Research practice II

Final report

# Building a better speaker separation algorithm for emotion recognition

José Andrés Carvajal [†] and Olga Lucia Quitero[‡*]
jacarvajab@eafit.edu.co, oquinte1@eafit.edu.co

[†]Mathematical Engineering, Universidad EAFIT
[‡]Mathematical Science Department, School of Sciences,
Mathematical Modeling Research Group, GRIMMAT, Universidad EAFIT

June 7, 2020

**Abstract**

In a conversation of two speakers there are many features, this ones contain important information that allows to identify each speaker voice signal. The Hilbert transform and dynamic threshold permit to build a better algorithm for prosodic silences identification and the Fourier analysis helps to display and synthesize different signals in terms of power spectral density distributions . Thus, through Fourier transform and the short time Fourier transform, which build a signal spectogram, we are going to identify the prosodic silences features obtained from the voice signal. Using these mathematical tools we made a Matlab implementation for two voice recorders, each one with two speakers, we are going to show the results and to conclude about this.

**Keywords:** speaker separation, Hilbert transform, prosodics silences, signal processing, Fourier transform.

## 1 Introduction

In the literature there are many articles about the emotion recognition, speakers separation, voice recognition in a conversation or something like that. In those papers the authors have used neural networks, convolutional neural networks and other artificial intelligence tools. Starting from this, we have found the relation between the emotion recognition and the speakers separation. If we use the algorithm describe by Puerta *et al.* (2015) to recognize the prosodic silences vector and after that, we use a similar algorithm for emotion recognition, then we can identify the speakers in a conversation through the silences signal spectogram and the Fourier transform.

[*]Tutor

This kind of procedure has many scopes, which will be mentioned in the state of the art, and this makes that the topic gets a huge interest by the author. However, here is also a personal motivation, because there is an interest on digital signal processing, on linear systems, on telecommunication and on the different mathematical techniques and theories that these ones embrace.

The expected results in this investigation are, understanding the applications of integral transform in the signals processing and the formulation about the filters. In addition to this trough the bibliographic review about algorithms that are used in the signal separation, we will create a simulation on Matlab.Some resources that will be used to carry out this work are, different computer programs that Universidad EAFIT has included, bibliographic bases and virtual and physical books. There may be some limitations with some references.

In the section 2 we describe the problem definition and use some references to justify this description. In the section 3 we show some authors that have done contributions to the investigation and we include the last advances in the speaker separation and emotion recognition, in section number 4 we describe some concepts,which will be useful in the algorithm formulation, we show the algorithm to identify prosodic silences, and the algorithm to voice recognition, which use the Fourier transform and the signal spectogram. In the section 5 we show some results that are obtained using the algorithms, described on the section 4, using voice recorders. Finally we show the conclusions about the project and future research.

## 2 Problem definition

A noise can be described as a perturbation of a signal, which interferes and changes its real values. A signal can have more than one noise, and that can be produced by electronic failures of the signal receptor or noises produced by close emitters, like the voice of others speakers. For this problem we can use a filter, which is a mechanism that processes input signal and modifies some properties of this. There are systems named adaptive filters, that are digital filters whose coefficients changes with an objective to make it converge to an optimal state, to carry put its function the filter use a structure of *Finite Impulse Response* (FIR). Puerta & Quintero (2014)

When we are in conversation our vocal cords vibrate and produce sound, this is called voiced and when our vocal chords are not vibrating is called unvoiced, moreover in the conversation exists silenced that is defined as the absence of audible sound or as a sound with a very low intensity. Using those terms like tools we can study the voice activity detection. The silences allow us to identify and separate the main components and they expose the rate at which the speaker delivers his speech, but when the sounds have low amplitude they are known as noises, that can be described as disturbances that interfere in the signal obtained by altering their real values. Puerta *et al.* (2015)

Using the idea about noise reduction and voiced activity detection we want to describe algorithms that remove the noise and separate the voice of a speaker on conversation. For this we will use concepts about linear algebra, integral transforms, adaptive filters and other mathematics tools that allow the development of different computational techniques for the audio signals study and its properties. In the development of the project is expected to start by studying Hilbert

transforms and their applications in the silenced recognition on an conversation. As a second part to the project, and since this is the way its exposed in the literature, we will do a research about others integral transforms and their applications in processing signals. Once this is completed we are going to compare the results and we will defined which to use in the implementation.

As mentioned previously, between the mathematical tools that will be used in the project are some concepts of linear systems , linear algebra, signal processing, calculus concepts, signal analysis and optimization. Finally, since it is also posed as the last objective of the project addressing the implementation in Matlab or Phyton of at least one of the existent algorithm for the speakers separation and the noise reduction.

# 3 State of the art

We can remove noise on a signal using adaptive filters, for this is necessary the use of two audio channels, where there is a signal that will be filtered and a reference of the noise. Puerta & Quintero (2014) propose a filtering of monophonic signals starting from Hilbert transform, which extracts the noise present on silences and this is used as the first step for the adaptation of the filter, in this step they use gradient descent. They make two tests using artificial noise and the natural noise of the signal. An important conclusion of this paper is that the method of Least squares is a good tool for the adaptation of the filter's coefficients, but its convergence velocity is very important for the structure of filtering.

Following the idea above, we can apply noise reduction to Hyperspectral Imagery. Hisham & Shen-En (2006) use the spectral derivative domain of the signal to carry out this. They use this technique because in this domain the the noise level is elevated and benefits from the dissimilarity of the signal regularity in the spatial and the spectral dimensions of hyperspectral images. For this development, they use Wavelet Shrinkage (WS) NR algorithms, because it benefits from the fact that wavelet transform provides a sparse representation for a wide class of signals, especially those that are piecewise smooth and of coherent regularity.

After applying the noise reduction algorithm we can use algorithms for speakers separation. Puerta *et al.* (2015) designs an algorithm based in Hilbert transform and dynamic threshold to pre-processing of audio signals. This allows the detection of prosodic and silence segments on a speech in presence of non-ideal conditions like a spectral overlapped noise. They propose as methodology the combination of three features from the signal: the zero-crossing rate, the signal energy and the signal coverage from the Hilbert transform. For the dynamic threshoold, they use crossings and energy as dynamic features of the signal, they say that using this dynamic the algorithm can adapt to the spectral characteristics over the signal.

Recently Wang *et al.* (2019) presents a system that separates the voice of a target speaker from multi-speaker signals, by making use of a reference signal from the target speaker. They achieve this by training two separate neural networks: A speaker recognition network and a spectrogram masking network that takes both noisy spectrogram and speaker embedding as input. In this paper is mentioned that the system is more applicable to real scenarios because it does not require prior

knowledge about the number of speakers and avoids the permutation problem. Luo & Mesgarani (2019) propose a fully-convolutional time-domain audio separation network (Conv-TasNet), a deep learning framework for end-to-end time-domain speech separation. They use a linear encoder to generate a representation of the speech waveform optimized for separating individual speakers, they say that the proposed Conv-TasNet system significantly outperforms previous time-frequency masking methods in separating two- and three-speaker mixtures.

The problem known as cocktail party, which is related to people's ability to focus their attention on a single conversation when at a noisy environment, in which there are a lot of conversations going on at the same time, was defined by Colin Cherry in 1993. Cory *et al.* (2017) proposes an algorithm to solve the cocktail party problem using a single microphone, they use neural networks regression to a vector space that is descriptive of independent speakers, such a vector space can embed empirically determined speaker characteristics and is optimized by distinguishing between speaker masks. Moreover, they use use weighted spectral features and masks to augment individual speaker frequencies while filtering out other speakers. Lemus & Ballesteros (2012) propose a solution to cocktail party problem using blind signal separation, they make a prepossessing of the signal (whitening and centering) and a decomposition phase based in the wavelet discrete transform, their model is called DWT-BSS, and it builds a virtual signal that is observed starting from a real signal and it uses the wavelet coefficients of the signals as input of the algorithm of separation in this case is JADE algorithm.

This kind of study represents a major step towards the realization of speech separation systems for real-world speech processing technologies, and the results could be used to establish relationships between the presence and frequency of these segments in a speech with the objective to detect deception, emotional states in social interaction, shortcomings of affective disorder or pathology associated with speech like the stuttering Puerta *et al.* (2015). Sierra (2016) through the Fourier transform he does emotion recognition from speech, They use that because "*The Fourier analysis allows for display and synthesizes different signals, in terms of power spectral density distributions*".

In this year, the scientific community has proposed some methods for to solve the speakers separation problem. Shi *et al.* (2020) proposes a novel architecture for speaker recognition is proposed by cascading speech enhancement and speaker processing. For this they use deep neural networks, with the aim of to improve speaker recognition performance when speech signals are corrupted by noise. Moreover, Delcroix *et al.* (2020) makes a algorithm using Speaker Beam Technology, which exploits an adaptation utterance of the target speaker to extract his/her voice characteristics that are then used to guide a neural network towards extracting speech of that speaker. Das *et al.* (2020) shows a state-of-the-art summary and present approaches for using the widely used machine learning and deep learning methods to detect the challenges along with future research directions of speech enhancement systems. This will be a important tools in the research about speaker separation.

# 4 Solution method / Methodology

## 4.1 Previous concepts

For this project, we must define some important concepts to understand the formulation and development of the speaker separation algorithm.

Suppose that we will recognize the silences in a conversation, which are processed by the signal $s(n)$. We are going to use overlapping windows with percentage of overlap $p$ on each window, and we calculate the Energy and zero crossing rate, using the following definitions

*Definition*: Let $s(n)$ be a discrete-time signal with $n = 1, 2..., N$. The signal's energy for each window is defined as following

$$E_j = \left[ \frac{1}{N} \sum_{n=(j-1)N+1}^{jN} |x(n)|^2 \right]^{\frac{1}{2}} \tag{1}$$

*Definition*: Let $s(n)$ be a discrete-time signal with $n = 1, 2..., N$. the zero crossing rate (ZRC) for each window is defined as

$$Z_j = \sum_{n=(j-1)N+1}^{jN} |\text{sgn}[x(i)] - \text{sgn}[x(i-1)]| \tag{2}$$

To recognize the silences we must use a signal's covering $s(n)$, for this we use the Hilbert transform and the analytic signal.

*Definition* If $f(t)$ is defined on the real line $-\infty < t < \infty$, its Hilbert transform, denoted by $\hat{f}_H$, is defined by

$$\mathcal{H}\{f(t)\} = \hat{f}_H = \frac{1}{\pi} \oint_{-\infty}^{\infty} \frac{f(t)}{t-x} dt \tag{3}$$

*Definition*: The analytic signal $f_a : \mathbb{R} \to \mathbb{C}$ is obtained by combining $f$ with $\hat{f}_H$ as a complex number

$$f_a = f + \hat{f}_H i \tag{4}$$

As $f_a$ is a complex number, then we can calculate the amplitude of it

$$|f_a| = \sqrt{\text{Re}(f_a)^2 + \text{Im}(f_a)^2}$$

This result will allow to build a signal covering and it is the first step to recognition of silences.

*Definition* If $f(x)$ is defined on the real line $-\infty < t < \infty$ and $|\int_{-\infty}^{\infty} f(x)| < \infty$, then the Fourier transform of $f(x)$ is defined by

$$\mathcal{F}\{f(x)\} = F(k) = \int_{-\infty}^{\infty} e^{-ikx} f(x) dx$$

To calculate the Hilbert transform we use the convolution, but this operation has a high computational component. To solve this we use the Fourier transform properties and it becomes in a multiplication as follows.

$$\mathcal{F}\{\mathcal{H}\{f(t)\}\} = \mathcal{F}\{\frac{1}{\pi t} * f(t)\} = -iF(w)\text{sgn}(w)$$

*Definition*: Let $f(t)$ be a signal. The short-time Fourier transform (STFT) is defined by

$$\textbf{STFT}\{f(T)\}(\tau, w) = F(\tau, w) = \int_{-\infty}^{\infty} f(t)w(t-\tau)e^{-iwt}dt$$

Where $w(t)$ is the window function, commonly a Han window or Gaussian window. This transform allows to build a signal spectogram.

## 4.2 Voice activity detection algorithm

Using the algorithm described in Puerta *et al.* (2015), we can build a simulation in Matlab. The steps to develop the algorithm are:

**input :** The voice signal **s** with frequency $F_s$
**output:** Silences vector **SL**

**if** $F_s \notin [100, 3200]$ **then**
  To Change frequency such that $F_s \in [100, 3200]$
**else**
  Building $n$ overlapping windows with $p$ overlap percentage
  For each window ($r$), calculating the windows energy $\mathbf{E}_s$ and the zeros crossing rate **ZCR**
  **if** $\boldsymbol{ZCR}_1 = 0$ **then**
    $ZCR_{\max} = \overline{\mathbf{ZCR}}$
    $ZCR_{\min} = \epsilon$
    **else**
    $ZCR_{\max} = \mathbf{ZCR}_1$
    $ZCR_{\min} = \min(\mathbf{ZCR})$;
  **end**
  $\mathrm{E}_{\max} = \overline{\mathbf{E}_s}$
  $\mathrm{E}_{\min} = \min(\mathbf{E}_s)$
  $\lambda_E = \frac{E_{\max} - E_{\min}}{E_{\max}}$
  $\lambda_Z = \frac{ZCR_{\max} - ZCR_{\min}}{ZCR_{\max}}$
  $\mathbf{E}_{th} = (1-\lambda_E)\mathbf{E}_{\max} + \lambda_E\mathbf{E}_{\min}$
  $\mathbf{Z}_{th} = (1-\lambda_Z)\mathbf{Z}_{\max} + \lambda_Z\mathbf{Z}_{\min}$
  **for** $i \leftarrow 1$ **to** $r$ **do**
    $\mathbf{E}_{\min} = E_{\min}(j-1)\Delta_E(j)$
    $\mathbf{Z}_{\min} = Z_{\min}(j-1)\Delta_Z(j)$
    $\Delta(j) = \Delta(j-1)\alpha$
    **TH** = TH(j) = (1-p)$\mathrm{E}_{th}(j) + pZ_{th}(j)$**end**
    ;
  **end**
  Apply **TH** for all the signal $\widetilde{\mathbf{TH}}$
  Calculate the Smooth covering **SC** using the analytic signal
  **for** $i \leftarrow 1$ **to** *len(s)* **do**
    **if** $\widetilde{\boldsymbol{TH}}(i) > \boldsymbol{SC}(i)$ **then**
      $\boldsymbol{SL}(i) = \boldsymbol{s}(i)$
    **end**
    **else**
    $\boldsymbol{SL}(i) = 0$ ;
  **end**

Figure 1: Algorithm to identify prosodic silences

### 4.3 Double Fourier analysis for silences vector

Using the silences vector and the algorithm described by Sosa *et al.* (2016), the results can be analyzed and we can recognize several features of the silences vector and it is providing the means to identify speakers.

**input :** The prosodic silences vector **f**
**output:** Fourier transform of signal spectogram **F**

1 Calculate the prosodic silences spectrogram, using the **STFT**
2 The spectogram is converted to gray-level image (**SG**).
3 Use the binary threshold
4 Calculate the fast Fourier transform for **SG**.
5 $F \leftarrow$ **SG**

Figure 2: Algorithm to voice recognition

## 5 Results

### 5.1 Example 1

For this example we use a voice recorder of two speakers , man and woman. The signals features are:

| Frequency (HZ) | Size (s) |
|---|---|
| 160000 | 236480 |

Table 1: Signal Features for example 1, voice recorder https://cutt.ly/iyXoGDb

Using the algorithm described in the section 4.2, with the following parameters

| New Freq (HZ) | Size windows | overlapping percentage | growth factor $\alpha$ | scaling factor $p$ | $\epsilon$ |
|---|---|---|---|---|---|
| 3000 | 256 | 90 % | 1.01 | 0.001 | 0.0001 |

Table 2: Signal Features for example 1

We obtain the covering signal, the dynamic threshold and the silences vector. We compare these results in the figure 3. Starting from these the Fourier analysis will be more interesting, because the silences frequency gives an important information about the speakers voice features.

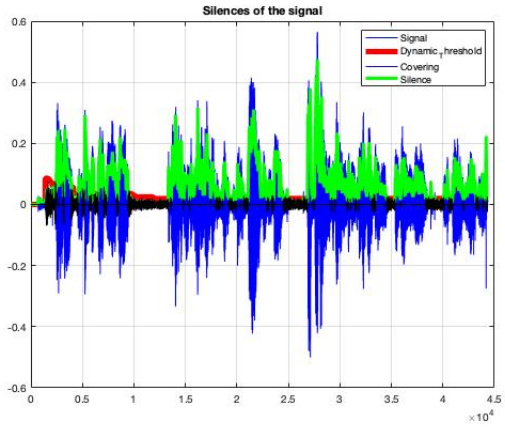
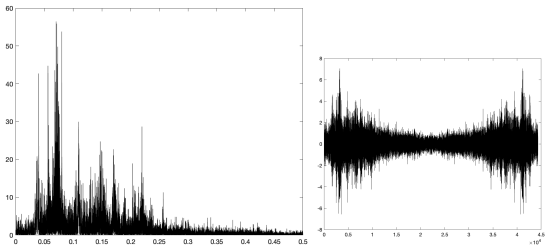



Figure 3: Signal covering, original signal, prosodic silences and dynamic threshold for example 1

Using the silences vector (black) and the Fourier transform, we can analyze the silence vector frequency, using the real part and amplitude of Fourier transform.



(a) Power (b) Real part

Figure 4: Fourier transform of the silences vector

Using the above figures, we can see the silences frequency, its amplitude and real part. Using this information we can recognize the highest frequencies in the silences vector. Applying the algorithm described in the section 4.3 we can analyze this information and we will understand the behavior of the silences frequency form each speaker.

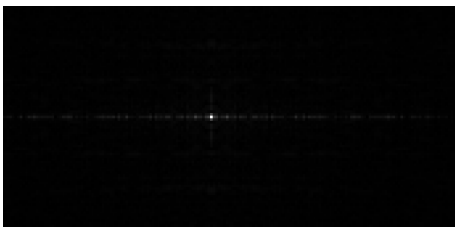


Figure 5: Fast Fourier transform of the spectogram

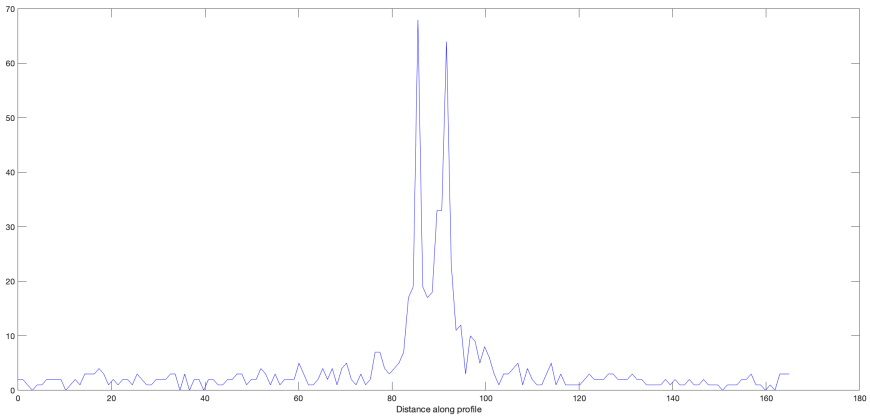



Figure 6: line profile

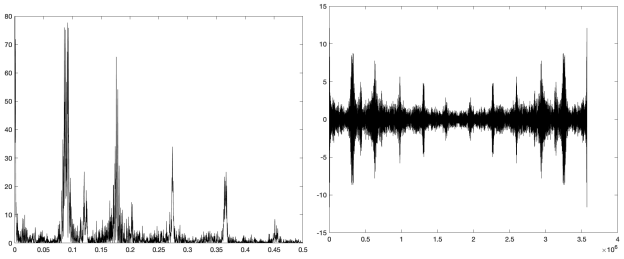
## 5.2 Example 2

For this example we use a voice recorder of two speakers , a child and her mom. The signals features are:

| Frequency (HZ) | Size (s) |
|---|---|
| 44100 | 52520896 |

Table 3: Signal Features for example 2, voice recorder https://cutt.ly/ayXoKA3

Using the same parameters from the example 1, and the Fourier transform for the silences vector the results are:



(a) Power (b) Real part

Figure 7: Fourier transform of the silences vector

In the figures above, the Fourier transform shows there are 5 spikes frequency, each one represents a features of the speakers. Those features are more intense than the example above. This behavior can be explained through the energy and power voice. It's clear that the child and mom voice are more different than the man and women, regarding to their intensity and sharp.

## 6 Conclusions and future research

The integral transforms have a lot of applications in the signal processing, in this project we used two of those. The Hilbert transform allows to identify silences through the analytic signal and the Fourier transform and STTF give information about the silences properties and the spectogram. Using this information we can find some features of the signal, using the power and energy.

This kind of algorithm than uses the Hilbert and Fourier transforms, proposes an efficient form to identify speakers in a conversation. Moreover, computational cost is low, the difficulty grade to implement this algorithm is not high, and its results are very important because they are the first step to recognized emotions in a conversation. However, the Fourier analysis gives tools to recognize emotions in an audio, it does all this through short time Fourier transform and the spectogram.

Using the above, as future works we want to recognize emotions in a conversation by using the speaker separation algorithm, the Fourier transform and STFF . Moreover, we want to implement this algorithms in real application and maybe this results will be compared one another and neural networks or artificial intelligence tools be applied, to make a more efficient algorithm.

## References


A., Revathi, R., Nagakrishnan, & Sasikaladevi. 2020. Twin identification from speech: linear and non-linear cepstral features and models. *N. Int J Speech Technol*.

Cory, Stephenson, Patrick, Callier, & Abhinav, Ganesh. 2017. Monaural Audio Speaker Separation Using Source-Contrastive Estimation.

Das, Nabanita, Chakraborty, Sayan, Chaki, Jyotismita, Padhy, Neelamadhab, & Dey, Nilanjan. 2020. Fundamentals, present and future perspectives of speech enhancement. *International Journal of Speech Technology*, 1–19.

Delcroix, Marc, Ochiai, Tsubasa, Zmolikova, Katerina, Kinoshita, Keisuke, Tawara, Naohiro, Nakatani, Tomohiro, & Araki, Shoko. 2020. Improving speaker discrimination of target speech extraction with time-domain SpeakerBeam. *arXiv Computer Science*.

Hisham, Othman, & Shen-En, Qian. 2006. Noise Reduction of Hyperspectral Imagery Using Hybrid Spatial-Spectral Derivative-Domain Wavelet Shrinkage. *Journal of Physics: Conference Series 705*.

Lemus, Camilo, & Ballesteros, Dora. 2012. Separación ciega de fuenetes no-determinada aplicada aa mezclas de voz con base en la transformada wavlet discreta. *Revista chilena de ingeniería*, 312–319.

Luo, Yi, & Mesgarani, Nima. 2019. Conv-TasNet: Surpassing Ideal Time- Frecuency Magnitude Masking for Speech Separation.

Puerta, David, & Quintero, O.L. 2014. Una aproximación al filtrado adaptativo para la cancelación de ruidos en señales de voz monofónicas. *Memorias del XV1 Congreso Latinoamericano de control Automático*, 510–515.

Puerta, David., Villa, Luisa F., Salazar, Carlos., & Quintero, O.L. 2015. A simple but efficient voice activity detection algorithm through Hilbert transform and dynamic threshold for speech pathologies. *Journal of Physics: Conference Series 705*.

Shi, Yanpei, Huang, Qiang, & Thomas, Hain. 2020. Robust Speaker Recognition Using Speech Enhancement And Attention Model. *Computer Science*.

Sierra, D. Bastidas, M. Quintero O.L. 2016. Double Fourier analysis for Emotion Identification in Voiced Speech. *Conference Series 705*.

Sosa, Daniel Sierra, Bastidas, Manuela, Ortiz, David, & Quintero, Olga Lucia. 2016. Double Fourier analysis for emotion identification in Voiced speech. *Journal of Physics Conferences Series*.

Wang, Quan, Muckenhirn, Hannah, Wilson, Kevin, & Wilson, Kevin. 2019. VoiceFilter: Targeted Voice Separation by Speaker-Conditioned Spectogram Masking e. *Google Inc.USA*.