Introducción al Reconocimiento de Patrones – Instituto de Ingeniería Eléctrica
Facultad de Ingeniería – Universidad de la República

INFORME PROYECTO FINAL

SEPARACIÓN DE FUENTES INDEPENDIENTES DE AUDIO USANDO ANÁLISIS DE COMPONENTES INDEPENDIENTES (ICA).

Haldo Spontón
C.I. 4.107.891-6

1.- INTRODUCCIÓN

1.1.- *Cocktail Party Problem*

Imaginemos que estamos en un cuarto en donde dos personas hablan simultáneamente. Se tienen dos micrófonos dispuestos en diferentes lugares. Grabaríamos dos señales temporales que podemos llamar $x_1(t)$ y $x_2(t)$. Cada una de estas señales es una suma ponderada de cada señal emitida por las personas que hablan, que llamaremos $s_1(t)$ y $s_2(t)$. Entonces, podemos expresar:

$$\begin{cases} x_1(t) = a_{11}s_1(t) + a_{12}s_2(t) \\ x_2(t) = a_{21}s_1(t) + a_{22}s_2(t) \end{cases} \qquad (1)$$

donde los a_{ij} dependen de la ubicación de los micrófonos con respecto a las personas[1].

Sería muy útil poder estimar las señales $s_1(t)$ y $s_2(t)$, usando solo las señales grabadas $x_1(t)$ y $x_2(t)$. Esto es llamado *cocktail-party problem*.

Nuestro acercamiento a resolver este problema será usar información sobre las propiedades estadísticas de las señales $s_i(t)$ para aproximar los a_{ii}. En realidad se deduce que es suficiente asumir que $s_1(t)$ y $s_2(t)$ son estadísticamente independientes en todo instante de tiempo. Esto no es una asunción irreal en la mayoría de los casos, y en la práctica esto puede no ser exactamente cierto.

1.2.- *Blind Signal Separation*

Consiste en la recuperación de señales independientes desconocidas desde la observación de una combinación lineal de las mismas.

Se tienen m señales desconocidas, una matriz de mezcla A_{mxn}, y por tanto n combinaciones lineales de las señales desconocidas:

$$\begin{aligned} s(t) &= \begin{bmatrix} s_1(t) & s_2(t) & \cdots & s_m(t) \end{bmatrix}^T \\ x(t) &= \begin{bmatrix} x_1(t) & x_2(t) & \cdots & x_n(t) \end{bmatrix}^T \end{aligned} \quad \Rightarrow \quad x = As\,. \qquad (2)$$

---

[1] Se supone que no hay retraso (delay) entre las señales x_1 y x_2, ni ningún otro efecto que escape al modelo de mezcla simplificado.

La columna a_k de A es denominada vector direccional asociado a la k-ésmia fuente $s_k(t)$.

ICA requiere que el número de señales recibidas n sea al menos igual al número de fuentes independientes m.

En la mayoría de las aplicaciones, o bien A no está disponible, o los errores en A hacen que las estimaciones no sirvan. En estas situaciones lo mejor es no asumir nada acerca de A. *ICA* es una clase de técnicas para aplicar *Blind Signal Separation*. En nuestro caso, consideraremos sinónimos, tanto a *ICA* como a *BSS*.

2.- ANÁLISIS DE COMPONENTES INDEPENDIENTES

2.1.- Definición de *ICA*

Supongamos, como antes, que observamos n mezclas lineales $x_1,\ldots,x_n$ de n componentes independientes:

$$\forall j: \quad x_j = a_{j1}s_1 + a_{j2}s_2 + \cdots + a_{jn}s_n \qquad (3)$$

En el modelo de *ICA*, asumimos que cada mezcla x_j y cada componente independiente s_k es una variable aleatoria. Entonces los valores observados $x_j(t)$, son muestras de dicha variable aleatoria. Sin pérdidas de generalidad, podemos asumir que tanto las mezclas como las fuentes son de media nula. Sino fuera cierto, las variables x_i pueden ser centradas, restando la media de las muestras.

Es conveniente utilizar una notación vectorial/matricial para el problema. Llamemos x al vector aleatorio cuyos elementos son las mezclas $x_1,\ldots,x_n$, y de la misma manera el vector s aleatorio, con elementos $s_1,\ldots,s_n$. Llamemos A a la matriz con elementos a_{ij}. Siempre consideraremos los vectores como columnas, entonces, x^T es un vector fila.

$$\Rightarrow \quad x = As. \qquad (4)$$

La ecuación anterior es llamada modelo *ICA* y describe como son generados los datos observados mediante la mezcla de las fuentes s_i. Las fuentes no pueden ser directamente observadas. También se asume desconocida la matriz A. Todo lo que tenemos es

el vector aleatorio x, y con éste debemos estimar A y s. Esto debe hacerse bajo asunciones lo más generales posibles.

El punto de partida de *ICA* es una simple asunción de que los componentes s_i son estadísticamente independientes. Se verá más adelante que se debe asumir que los componentes independientes deben tener distribuciones no-gaussianas. Sin embargo, en el modelo básico, no asumimos conocidas esas distribuciones. Por simplicidad, también asumimos que la matriz de mezcla –desconocida- es cuadrada, pero esta condición puede ser relajada. Luego de estimar A, podemos calcular su inversa –llámese W- y obtener los componentes independientes:

$$s = Wx. \qquad (5)$$

2.2.- Ambigüedades de *ICA*

1. No podemos determinar las varianzas (energías) de las componentes independientes.

La razón es que, al ser A y s desconocidos, cualquier factor multiplicativo en una de las fuentes s_i puede siempre cancelarse, dividiendo la correspondiente columna a_i de A por el mismo factor. Como consecuencia, debemos adecuar las magnitudes de las componentes independientes. Como son variables aleatorias, la forma más natural de hacerlo es asumir varianza unitaria: $\mathrm{E}\{s_i^2\}=1$. La matriz A será adaptada en la solución de *ICA*, teniendo en cuenta esta restricción.

2. No podemos determinar el orden de las componentes independientes.

Al ser A y s desconocidos, podemos cambiar libremente el subíndice de los s_i. Formalmente, una matriz de permutación P y su inversa pueden agregarse al modelo para dar $x = AP^{-1}Ps$. Los elementos de Ps son las s_i originales, pero cambiadas de orden. La matriz AP^{-1} es una nueva matriz de mezcla desconocida, a determinarse por los algoritmos de *ICA*.

3.- ALGUNOS CONCEPTOS TEÓRICOS

Veremos a continuación algunos conceptos teóricos que nos serán de utilidad a la hora de estudiar los algoritmos de *ICA*.

3.1.- Independencia, definición y propiedades

Consideremos dos variables aleatorias escalares y_1 e y_2. Básicamente, dichas variables son independientes si la información del valor de una no aporta nada de información sobre el valor de la otra. Antes, habíamos señalado que este es el caso de las variables s_1 y s_2, pero no de x_1 y x_2.

Técnicamente, la independencia puede definirse por las densidades de probabilidad. Definimos que y_1 e y_2 son independientes si y solo si la pdf conjunta es factorizable de la siguiente manera:

$$p(y_1, y_2) = p_1(y_1)p_2(y_2) \qquad (6)$$

siendo $p_1(y_1)$ y $p_2(y_2)$ las pdfs marginales de y_1 e y_2 respectivamente[2].

La definición puede ser usada para derivar una de las propiedades más importantes de las variables aleatorias independientes. Dadas dos funciones h_1 y h_2, se tiene:

$$\mathrm{E}\{h_1(y_1)h_2(y_2)\} = \mathrm{E}\{h_1(y_1)\}\mathrm{E}\{h_2(y_2)\} \qquad (7)$$

Una forma más débil de independencia es la no correlación. Dos variables aleatorias y_1 e y_2 son no correlacionadas, si su covarianza es nula:

$$\mathrm{E}\{y_1 y_2\} - \mathrm{E}\{y_1\}\mathrm{E}\{y_2\} = 0 \qquad (8)$$

Independencia implica no correlación, pero no al revés. Algunos métodos de *ICA* proceden estimando los componentes independientes con señales no correlacionadas. Esto simplifica el problema.

3.2.- Problema con variables gaussianas

Para ver por que variables gaussianas hacen imposible a *ICA*, asumamos que la matriz de mezcla es ortogonal, y que las s_i son gaussianas. Entonces x_1 y x_2 son gaussianas, no correlacionadas, y de varianza unitaria. Su densidad conjunta está dada por:

$$p(x_1, x_2) = \frac{1}{2\pi} e^{-\left(\frac{x_1^2 + x_2^2}{2}\right)} \qquad (9)$$

---

[2] La definición se extiende naturalmente para n variables aleatorias.

Se ve que la densidad es completamente simétrica, por lo tanto no contiene información de la dirección de las columnas de la matriz de mezcla A, por lo que la misma no puede estimarse.

4.- PRINCIPIOS DE LA ESTIMACIÓN *ICA*

4.1.- No-Gaussianidad

Intuitivamente hablando, la clave para la estimación del modelo *ICA* es la no-gaussianidad. Sin ella, la estimación se hace imposible.

El Teorema Central del Límite, un resultado clásico en probabilidad, dice que la distribución de una suma de variables aleatorias independientes tiende a una gaussiana, bajo ciertas condiciones. Por lo tanto, la suma de dos variables aleatorias independientes tiene, en general, una distribución más gaussiana que las originales variables aleatorias.

Ahora asumamos que el vector x está distribuido acorde al modelo *ICA*, o sea, es una mezcla de componentes independientes. Por simplicidad, en esta sección asumamos que todos los componentes independientes tienen idénticas distribuciones.

Para estimar uno de los componentes independientes, consideramos una combinación lineal de los $x_i \rightarrow y = w^T x = \sum_i w_i x_i$, donde w es un vector a determinar. Si w fuese una de las filas de la inversa de A, la combinación lineal sería igual a uno de los componentes independientes. La cuestión está en como usar el Teorema Central del Límite para determinar w y que sea igual –o lo más parecido- a una de las filas de la inversa de A. En la práctica no se puede determinar exactamente w, porque no conocemos la matriz A, pero se puede llegar a una buena aproximación.

Para ver como esto lleva al principio básico de la estimación con *ICA*, hagamos un cambio de variable $z = A^T w \Rightarrow y = w^T x = w^T A s = z^T s$, entonces, y es combinación lineal de los s_i, ponderados por z_i.

Dado que la suma de incluso dos variables aleatorias independientes es más gaussiana que las originales, $z^T s$ es más gaussiana que cualquiera de las s_i, y se vuelve menos gaussiana cuando es, de hecho, igual a alguna s_i. En este caso, obviamente solo uno de los elementos de z_i es distinto de cero.

Por lo tanto, podemos tomar como w un vector que maximice la no-gaussianidad de $w^T x$. Dicho vector necesariamente se corresponderá con un z que tenga solo un componente no nulo. Esto significa que $w^T x = z^T s$, o sea, una de las componentes independientes.

Entonces, maximizar la no-gaussianidad de $w^T x$ nos da una de las componentes independientes. De hecho, el tope de la optimización de la no-gaussianidad en dimensión n tiene $2n$ máximos locales, dos para cada componente independiente, correspondientes a s_i y $-s_i$[3].

Para encontrar todas las componentes independientes, debemos hallar todos esos máximos locales. Esto no es difícil, porque las diferentes componentes independientes son no-correlacionadas: siempre podremos restringir la búsqueda al espacio que brinda estimaciones no correlacionadas con las previas. Esto corresponde a la ortogonalidad de un espacio adecuadamente transformado (por ejemplo, blanqueado).

Nuestro enfoque fue puramente heurístico, pero veremos en las siguientes secciones que tiene una rigurosa justificación.

4.2.- Medida de la No-Gaussianidad

Para usar la no-gaussianidad en la estimación con *ICA*, debemos tener una medida cuantitativa de la no-gaussianidad de una variable aleatoria, llámese y.

Para simplificar, asumamos que y está centrada (media nula) y de varianza unitaria[4].

4.2.1.- *Kurtosis*

La medida clásica de la no-gaussianidad es el *Kurtosis* ó "*fourth-order cumulant*".

$$kurt(y) = \mathrm{E}\{y^4\} - 3\left(\mathrm{E}\{y^2\}\right)^2 \qquad (10)$$

---

[3] Recordemos que las componentes independientes pueden ser estimadas a menos de un signo multiplicativo.

[4] Más adelante se verá un procedimiento de pre-procesamiento para siempre lograr esta simplificación.

En realidad, si asumimos varianza unitaria para y, tenemos $kurt(y) = \mathrm{E}\{y^4\} - 3$. Esto muestra que el *kurtosis* es una versión normalizada del cuarto momento $\mathrm{E}\{y^4\}$. Para una y gaussiana, $kurt(y) = 0$. Para la mayoría de las variables aleatorias no gaussianas, el *kurtosis* no se anula.

El *kurtosis* puede ser positivo o negativo. Variables aleatorias con *kurtosis* negativo son llamadas sub-gaussianas, mientras que con *kurtosis* positivo son super-gaussianas.

Típicamente la no-gaussianidad se mide con el valor absoluto del *kurtosis*. El cuadrado del mismo también puede usarse. Ambos son nulos para variables gaussianas, y mayores que cero cuanto más no-gaussiana sea la variable.

El *kurtosis* (o su valor absoluto) ha sido ampliamente usado como medida de no-gaussianidad en *ICA* y campos relacionados. La principal razón es su simplicidad, tanto computacional como teórica. Computacionalmente, el *kurtosis* puede ser estimado usando el cuarto momento de las muestras de los datos. El análisis teórico se simplifica por la siguiente propiedad de linealidad (se deriva de la definición); Si x_1 y x_2 son dos variables aleatorias independientes:

$$\begin{aligned} kurt(x_1 + x_2) &= kurt(x_1) + kurt(x_2) \\ kurt(\alpha x_1) &= \alpha^4 kurt(x_1) \end{aligned} \qquad (11)$$

siendo α un escalar.

En la práctica comenzamos con cierto vector w, calculamos la dirección en la cual el *kurtosis* de $y = w^T x$ crece más (si $kurt > 0$) o decrece más (si $kurt < 0$), basados en las muestras disponibles $x(1), \cdots, x(T)$ del vector mezclado x, y usamos algún método de gradiente o alguna de sus extensiones para encontrar un nuevo vector w.

Sin embargo el *kurtosis* tiene algunos puntos bajos en la práctica, cuando su valor tiene que ser estimado con pocas muestras. El principal problema es que el *kurtosis* puede ser muy sensible a *outliers*. En otras palabras, el *kurtosis* no es una medida robusta de la no-gaussianidad.

4.2.2.- Negentropía

Basada en cantidad de información teórica de la entropía.

La entropía es el concepto básico de la teoría de la información. La entropía de una variable aleatoria puede ser interpretada como el grado de información que aporta la observación de la variable. Cuanto más aleatoria, impredecible e inestructurada es la variable, mayor es su entropía.

La entropía H está definida para una variable aleatoria:

Y Discreta $\Rightarrow$ $H(Y) = -\sum P(Y = a_i)\log P(Y = a_i)$, donde a_i son los posibles valores de Y.

y Continua $\Rightarrow$ $H(y) = -\int f(y)\log f(y)dy$, donde $f(y)$ es la densidad de y.

Un resultado fundamental en teoría de la información es que una variable gaussiana posee la mayor entropía entre todas las variables aleatorias de igual varianza. Esto significa que la entropía puede ser usada como medida de no-gaussianidad. De hecho, esto muestra que la distribución gaussiana es la "más aleatoria", o la menos estructurada de todas las distribuciones. La entropía es chica para distribuciones que están claramente concentradas en ciertos valores.

Para obtener una medida de no-gaussianidad, que se anule para una variable gaussiana, y sea siempre no negativa, se usa una versión modificada de la definición de entropía diferencial, llamada negentropía:

$$J(y) = H(y_{gauss}) - H(y) \qquad (12)$$

donde y_{gauss} es una variable aleatoria gaussiana con la misma matriz de covarianza que y.

Dado lo mencionado arriba, la negentropía es siempre no negativa, y solo se anula si y tiene distribución gaussiana. La negentropía además es invariante ante transformaciones lineales invertibles.

La ventaja de usar negentropía o entropía diferencial como medida de no-gaussianidad es que tiene una fuerte justificación teórica. De hecho en algunos sentidos es el mejor estimador de no-gaussianidad. El problema es que es computacionalmente muy costoso de calcular (requiere estimar la pdf, probablemente no paramétrica). Por lo tanto aproximaciones simples de la negentropía serán muy útiles.

4.2.3.- Aproximaciones de la Negentropía

El clásico método para aproximar la negentropía es usando momentos de alto orden:

$$J(y) \approx \frac{1}{12}\mathrm{E}\{y^3\}^2 + \frac{1}{48}kurt(y)^2 \qquad (13)$$

Se asume que la variable aleatoria y tiene media nula y varianza unitaria. Sin embargo, la validez de esta aproximación puede ser bastante limitada. En particular, esta aproximación sufre los mismos problemas de robustez que el *kurtosis*.

Para evitar estos problemas, se desarrollaron nuevas aproximaciones basadas en el principio de máxima entropía[5]:

$$J(y) \approx \sum_{i=1}^{p} k_i \left[\mathrm{E}\{G_i(y)\} - \mathrm{E}\{G_i(v)\}\right]^2 \qquad (14)$$

donde k_i son constantes positivas, y v es una variable aleatoria gaussiana de media nula y varianza unitaria. Se asume que y es de media nula y varianza unitaria, y las funciones G_i son no-cuadráticas.

En el caso que usemos solo una función no-cuadrática G, la aproximación es:

$$J(y) \propto \left[\mathrm{E}\{G(y)\} - \mathrm{E}\{G(v)\}\right]^2 \qquad (15)$$

para prácticamente cualquier función G no-cuadrática.

Esta es claramente una generalización de (13), si y es simétrica. Más aún, tomando $G(y) = y^4$ se obtiene exactamente (13).

El punto esta en elegir adecuadamente G, para obtener mejores aproximaciones de la negentropía. En particular, eligiendo G que no crezca muy rápido, uno obtiene estimadores más robustos. Las siguientes funciones son útiles:

$$G_1(u) = \frac{1}{a_1}\log\cosh a_1 u \qquad G_2(u) = -e^{-\frac{u^2}{2}} \qquad (16)$$

donde $1 \le a_1 \le 2$ es una constante adecuada.

---

[5] Nótese que incluso cuando la aproximación no es buena, (14) igual puede ser usada para construir una medida de no-gaussianidad, coherente con las propiedades de la negentropía.

4.2.4.- Otras

Existen otras maneras de medir la no-gaussianidad de una variable aleatoria, que escapan a lo que utilizaremos en nuestros algoritmos mas adelante.

Uno consiste en la minimización de la información mutua, lo que se puede probar que se basa en el mismo principio de encontrar la dirección más no-gaussiana.

Otro manera de estimar *ICA*, o sea, medir la no-gaussianidad, es mediante un EMV, o Estimador de Máxima Verosimilitud. Esto está directamente conectado con el principio *infomax*, lo que es esencialmente equivalente a la minimización de la información mutua antes mencionada.

4.3.- *Projection Pursuit*

Es interesante notar como nuestro acercamiento hacia *ICA* hace explícita la conexión entre el mismo y *"projection pursuit"*. Esta última es una técnica desarrollada en estadística para encontrar proyecciones interesantes de datos multidimensionales. Dichas proyecciones pueden ser usadas para la óptima visualización de los datos, y para objetivos como estimación de densidades y regresiones. En la más básica (1D), se trata de encontrar direcciones tales que las proyecciones de los datos en esas direcciones tengan distribuciones de interés. Se argumenta que la distribución gaussiana es la de menor interés, siendo las direcciones de mayor interés aquellas que muestran una distribución lo menos gaussiana. Eso es exactamente lo que hacemos para estimar el modelo *ICA*.

Por lo tanto, en su formulación general, *ICA* puede considerarse una variante de *projection pursuit*. En particular, *projection pursuit* nos permite enfrentar la situación en donde hay menos componentes independientes que variables originales.

5.- PRE-PROCESAMIENTO PARA *ICA*

Antes de aplicar *ICA* sobre los datos, es muy útil realizar cierto pre-procesado. En esta sección veremos ciertas técnicas de procesamiento para hacer el problema de estimación *ICA* más simple y mejor condicionado.

5.1.- Centrado

El más básico y necesario pre-procesamiento es centrar x, restándole el vector medio $\mathrm{E}\{x\}$, haciendo al vector x de media nula. Esto implica que también s es de media nula, como puede verse tomando esperanzas a ambos lados de la ecuación (4).

Este pre-procesamiento es únicamente para simplificar los algoritmos *ICA*, no significa que las medias no puedan estimarse. Luego de estimar la matriz de mezcla A con los datos centrados, podemos completar la estimación agregando el vector de medias de s al vector estimado centrado. Dicho vector de medias esta dado por $A^{-1}m$, donde m es el vector antes restado.

5.2.- Blanqueado (*Whitening*)

Otra estrategia útil de pre-procesamiento en *ICA* es primero blanquear las variables observadas. Esto significa que antes de la aplicación de los algoritmos de *ICA* (pero después del centrado), transformamos linealmente el vector observado x para obtener un nuevo vector $\tilde{x}$, el cual es blanco (sus componentes son no correlacionadas y sus varianzas son unitarias). En otras palabras, la matriz de covarianza de $\tilde{x}$ es igual a la identidad.

$$\mathrm{E}\{\tilde{x}\tilde{x}^T\} = \mathrm{I} \qquad (17)$$

La transformación de blanqueado es siempre posible. Un método popular para hacerlo es usar la descomposición en valores propios (EVD) de la matriz de covarianza $\mathrm{E}\{xx^T\} = EDE^T$, donde E es la matriz ortogonal de vectores propios de $\mathrm{E}\{xx^T\}$ y D es la matriz diagonal con sus vectores propios, $D = diag(d_1, \cdots, d_n)$.

Notar que $\mathrm{E}\{xx^T\}$ puede ser estimado de una manera estándar usando las muestras disponibles $x(1), \cdots, x(T)$. El blanqueado puede ahora realizarse:

$$\tilde{x} = ED^{-1/2}E^T x \qquad (18)$$

donde la matriz $D^{-1/2}$ es calculada elemento a elemento: $D^{-1/2} = diag\left(d_1^{-1/2}, \cdots, d_n^{-1/2}\right)$.

Es fácil ver ahora que $\mathrm{E}\{\tilde{x}\tilde{x}^T\} = \mathrm{I}$.

El blanqueado transforma la matriz de mezcla en una nueva, $\tilde{A}$.

$$\tilde{x} = ED^{-1/2}E^T As = \tilde{A}s \qquad (19)$$

La utilidad del blanqueado reside en que la nueva matriz de mezcla $\tilde{A}$ es ortogonal:

$$\mathrm{E}\{\tilde{x}\tilde{x}^T\} = \tilde{A}\mathrm{E}\{ss^T\}\tilde{A}^T = \tilde{A}\tilde{A}^T = \mathrm{I} \qquad (20)$$

Aquí vemos que el blanqueado reduce el número de parámetros a ser estimados. En vez de tener que estimar n^2 parámetros (elementos de A), solo necesitamos estimar la nueva matriz de mezcla ortogonal $\tilde{A}$. Una matriz ortogonal posee $n(n-1)/2$ grados de libertad. Por ejemplo, en dos dimensiones, una transformación ortogonal es determinada por un único parámetro (ángulo). En mayores dimensiones, una matriz ortogonal contiene aproximadamente la mitad del número de parámetros que una matriz cualquiera, por lo que se puede decir que el blanqueado resuelve la mitad del problema en *ICA*.

De ahora en adelante, asumimos que los datos han sido pre-procesados (centrados y blanqueados). Para simplificar la notación, denotaremos a los datos pre-procesados simplemente por x, y la matriz de mezcla transformada como A.

5.3.- Efecto del filtrado

El éxito de *ICA* para un conjunto de datos depende crucialmente de la realización de pasos de pre-procesamiento específicos para la aplicación. Por ejemplo, si los datos son señales temporales, algún filtro pasa-banda puede ser muy útil.

Se puede ver que si filtramos linealmente las señales $x_i(t)$ para obtener nuevas señales $x_i^*(t)$, el modelo *ICA* mantiene su matriz de mezcla.

6.- EL ALGORITMO *FASTICA*

Hasta ahora introdujimos diferentes medidas de no-gaussianidad, funciones objetivo para la estimación *ICA*. En la práctica también se necesita un algoritmo para maximizar la función de contraste adecuada para esta tarea. Está asumido aquí que los datos están centrados y blanqueados.

6.1.- *FastICA* para una unidad

Para empezar veremos la versión de una unidad de *FastICA*. Por una "unidad" nos referimos a una unidad computacional, eventualmente una neurona artificial con un vector de pesos w, el cual dicha neurona es capaz de entrenar. La regla de aprendizaje de *FastICA* encuentra una dirección, es decir un vector w tal que la proyección $w^T x$ maximiza la no-gaussianidad. Aquí la no-gaussianidad es medida por la aproximación de la negentropía $J(w^T x)$ dada en (15). Recordar que la varianza de $w^T x$ está obligada a valer 1; para los datos blanqueados, esto equivale a obligar que la norma de w sea 1.

FastICA está basado en una estrategia de iteración de punto fijo para encontrar el máximo de no-gaussianidad de $w^T x$, medida como en (15). Puede también derivarse como una iteración aproximativa de Newton. Llamamos g a la derivada de la función no-cuadrática G usada en (15); por ejemplo, las derivadas de las funciones en (16) son:

$$g_1(u) = \tanh(a_1 u) \qquad g_2(u) = u.e^{-\frac{u^2}{2}} \qquad (21)$$

donde $1 \le a_1 \le 2$ es alguna constante adecuada, usualmente igual a 1. La forma básica del algoritmo *FastICA* es la siguiente:

1) Elegir un vector de pesos inicial w (por ejemplo aleatorio).

2) Se calcula: $w^+ = \mathrm{E}\{x g(w^T x)\} - \mathrm{E}\{g'(w^T x)\} w$

3) Se actualiza w: $w = \dfrac{w^+}{\|w^+\|}$

4) Si no converge, volver a 2).

La convergencia significa que el viejo y nuevo valor de w apuntan en la misma dirección, es decir, su producto escalar es casi igual a 1. No es necesario que el vector converja a un único punto, ya que w y $-w$ definen la misma dirección. Esto, nuevamente, se debe a que las componentes independientes pueden ser definidas a menos de un signo.

En la práctica, las esperanzas en *FastICA* deben reemplazarse por sus estimaciones, que naturalmente son las medias de las muestras. Idealmente deberían usarse todos los datos disponibles, pero no es una buena idea dado el costo computacional. Entonces los

promedios pueden estimarse usando un subconjunto de muestras, cuyo tamaño puede influir demasiado en la precisión de la estimación final.

6.2.- *FastICA* para varias unidades

El algoritmo para una unidad antes descrito estima solo uno de los componentes independientes, o una dirección de *projection pursuit*. Para estimar varias componentes independientes, necesitamos correr el algoritmo *FastICA* de una unidad pero usando varias unidades (neuronas) con vectores de pesos $w_1,\ldots,w_n$.

Para evitar que diferentes vectores converjan al mismo máximo debemos decorrelacionar las salidas $w_1^T x,\ldots,w_n^T x$ luego de cada iteración. Veremos tres métodos para hacerlo:

Una simple manera de lograr no correlación es con una estrategia de deflación basada en una decorrelación similar a Gram-Schmidt. Estimamos los componentes independientes uno a uno. Cuando tenemos estimados p componentes independientes, ó p vectores $w_1,\ldots,w_p$, corremos el algoritmo de punto fijo de una unidad para w_{p+1}, y luego de cada paso iterativo le restamos a w_{p+1} las "proyecciones" $w_{p+1}^T w_j w_j$, $j=1\ldots p$ de los p vectores estimados previamente, y luego normalizamos w_{p+1}.

$$1)\ w_{p+1} = w_{p+1} - \sum_{j=1}^{p} w_{p+1}^T w_j w_j$$

$$2)\ w_{p+1} = \frac{w_{p+1}}{\sqrt{w_{p+1}^T w_{p+1}}} \qquad (22)$$

6.3.- Propiedades del algoritmo *FastICA*

El algoritmo *FastICA* tiene algunas propiedades muy buenas, en comparación con otros métodos existentes para *ICA*.

i.- Bajo la asunción de validez del modelo *ICA*, la convergencia es cúbica (o al menos cuadrática). En otros algoritmos de *ICA* basados en descenso por el gradiente, la convergencia es solo lineal. Esto significa por tanto una muy rápida convergencia.

ii.- Al contrario con los algoritmos basados en descenso por el gradiente, no hay ningún parámetro a seleccionar, lo que hace al algoritmo sencillo de usar.

iii.- El algoritmo *FastICA* encuentra directamente componentes independientes de casi cualquier distribución no-gaussiana usando cualquier no-linealidad g.

iv.- La performance del método puede optimizarse eligiendo adecuadamente la función g.

v.- Las componentes independientes pueden ser estimadas una a una, lo que es fuertemente equivalente a *projection pursuit*. Esto hace decrecer el costo computacional cuando no todas las componentes independientes necesitan ser estimadas.

vi.- El algoritmo *FastICA* tiene la mayoría de las ventajas de los algoritmos basados en redes neuronales; Se realiza en paralelo y distribuido, es computacionalmente simple, y requiere poco espacio en memoria.

6.4.- Algoritmo alternativo de *FastICA*

Se implementó además otro algoritmo de *FastICA* que describimos a continuación. Al igual que antes, se suponen los datos centrados y blanqueados. En este caso se busca directamente estimar la inversa de la matriz de mezcla, a la que llamaremos V. Los pasos a seguir son los siguientes:

1) Elegir una matriz ortogonal V al azar.

2) Calcular $z(t) = Vx(t)$.

3) Actualizar V como: $V \leftarrow V + \Gamma\left[diag(-\beta) + G\right]V$

$$\beta_i = \frac{1}{T}\sum_{t=1}^{T} z_i(t)\,g(z_i(t))$$

con:

$$\Gamma = diag\left(\frac{1}{\beta_i - \sum_{t=i}^{T} g'(z_i(t))}\right)$$

4) Ortogonalizar V: $V = V\left(V^T V\right)^{-0.5}$

5) Si no termino, volver a 2).

La no linealidad g puede ser elegida como:

$$g_1(x) = x^3 \qquad g_2(x) = \tanh(x) \qquad g_3(x) = x.e^{-\frac{x^2}{2}}$$

Tanto el algoritmo presentado en la sección 6.1 (6.2), como el presentado ahora, fueron implementados en los archivos `fastica.m` y `fastica2.m` respectivamente.

7.- PRUEBAS Y RESULTADOS

Se realizaron variadas pruebas de los algoritmos, con variadas condiciones de simulación y tipos de archivos de audio.

En principio, se optó por generar mezclas de componentes independientes conocidos, usando una matriz de mezcla aleatoria. Luego, aplicar los algoritmos implementados a las mezclas obtenidas, y comparar los resultados de componentes independientes determinados con los originales antes de mezclar. Estos resultados se pueden apreciar en los archivos adjuntos a este informe.

Se notaron excelentes resultados a la hora de separar 2 componentes independientes (usando dos mezclas). Se pueden apreciar buenos resultados al separar voces, instrumentos tanto sonoros como de percusión, e incluso mezcla de ambas. Se apreció que el algoritmo de la sección 6.1 funciona mejor a la hora de separar instrumentos, mientras que el algoritmo en 6.4 funciona mejor a la hora de separar voces. No se encontró una explicación convincente para dicho fenómeno, pero es bueno disponer de ambos algoritmos, para poder tener más de un resultado de separación.

También, como agregado a las pruebas, y con el objetivo de comparar el funcionamiento de los algoritmos implementados en este proyecto, con los de otros compañeros, se usaron los mismos archivos de mezclas utilizados en el otro proyecto. También se podrán apreciar los resultados en la muestra de este proyecto, así como en los archivos adjuntos a este informe.

7.1.- Simulación de habitación

Como extra al proyecto, y con le objetivo de generar mezclas en situaciones más reales, se uso para generar las mismas un simulador de habitación. Este, ubica las 2 personas hablando y los dos micrófonos necesarios en una habitación virtual, de la cual se

pueden configurar sus dimensiones. En nuestro caso usamos una habitación de 5m x 5m, y 3m de alto. La ubicación de las personas que hablan y de los micrófonos se aprecia en la siguiente imagen:

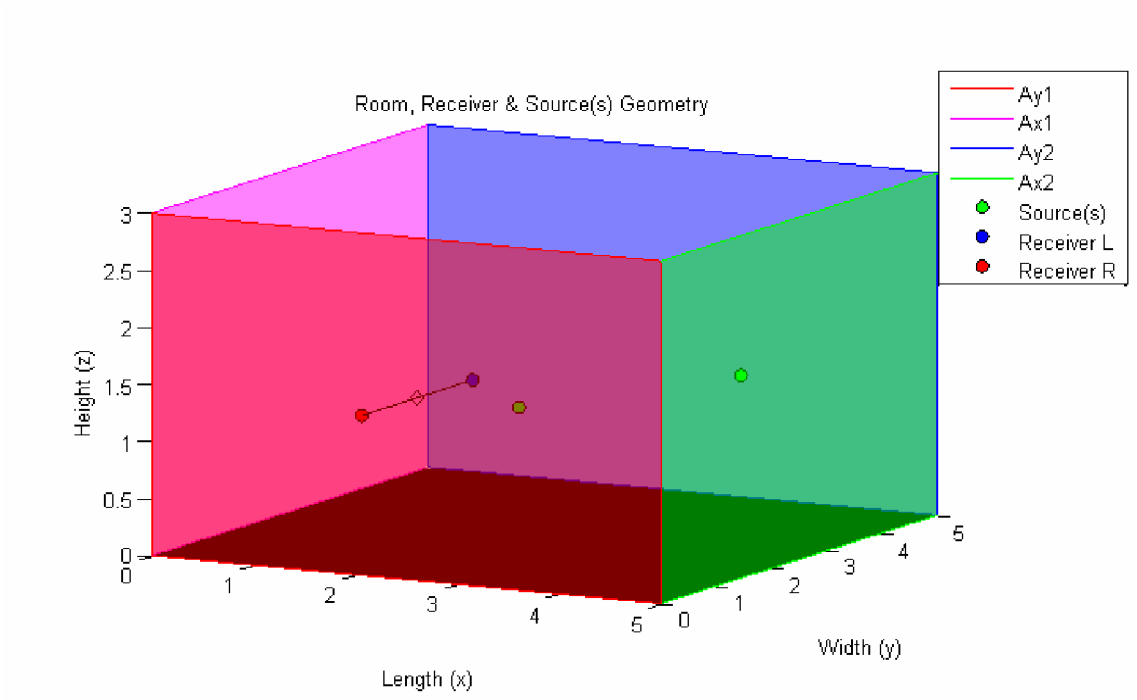

Figura 1.- Geometría de la habitación virtual, y ubicación de los emisores y receptores en la misma.

Una vez configurada la habitación, el programa genera automáticamente un archivo estéreo, en donde cada canal es la mezcla que entra en cada canal del micrófono. Esta mezcla estéreo fue convertida en dos mezclas mono, una por canal, sin modificarlas, para adecuarlas a la entrada de nuestros algoritmos. Se puede apreciar en los archivos adjuntos a este informe los resultados de salida de dichos algoritmos.

8.- CONCLUSIONES

Como primera conclusión de este trabajo, cabe destacar el buen funcionamiento de los algoritmos implementados. Desde un punto de vista perceptual se notan excelentes resultados en la separación de fuentes dadas las respectivas mezclas.

Otra cosa a destacar es en si el buen funcionamiento de los algoritmos de *FastICA*, en relación a su extrema sencillez y velocidad

de ejecución. Ambos algoritmos implementados funcionan rápido, y los resultados son suficientemente correctos. Además la implementación es sencilla, de fácil comprensión y depuración. No por nada este tipo de algoritmos tuvieron en su tiempo un auge muy importante para la separación de audio en fuentes independientes.

Como conclusión personal destaco el interés que en mi despiertan este tipo de trabajos y proyectos con audio, siendo muy gratificante llegar a los resultados obtenidos, que son, en algunos casos, hasta sorprendentes.

9.- REFERENCIAS

- Bodgan Matei – *"A Review of Independent Component Análisis Techniques"* – Electrical and Computer Engineering Department – Rutgers University, Piscataway, NJ, USA.

- Aapo Hyvärinen and Erkki Oja – *"Independent Component Analysis: Algorithms and Applications"* – Neural Networks Research Centre – Helsinki University of Technology, Finland.

- Zbynêk Koldovský[1] and Petr Tichavský[2] – *"Time-Domain Blind Audio Source Separation using Advanced ICA Methods"* – [1]Institute of Information Technology and Electronics, Technical University of Liberec, Liberec, Czech Republic. [2]Institute of Information Theory and Automation, Academy of Sciences of the Sciences of the Czech Republic, Czech Republic.

- Agradecimiento especial a los compañeros Ricardo Laureiro y Mariana Díaz, por la excelente disposición mostrada, y a los docentes involucrados en este proyecto.