Lección 3.1: Uso del Cepstrum para el procesamiento de señales de voz

Dr. Alessandro Presacco
Dr. Jesús Savage

4 de mayo de 2021

Índice

Sistema homomórfico

Un sistema homomórfico es una clase de sistemas non-lineales que se puede definir con el principio generalizado de sobreposición

$$y[n] = L\{x[n]\} = L\{x_1[n] + x_2[n]\} =$$
$$= L\{x_1[n]\} + L\{x_2[n]\} =$$
$$= y_1[n] + y_2[n]$$

y

$$= L\{ax[n]\} = aL\{x[n]\} = ay[n]$$

where $L\{*\}$ es el operarador linear que representa el sistema

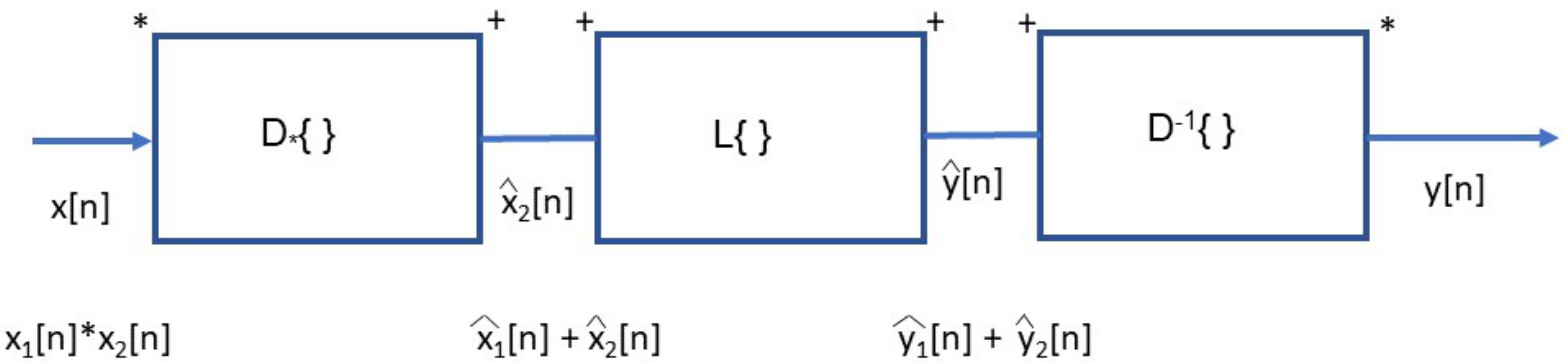
Sistema homomórfico de convolucion

Un sistema homomórfico para la convolucion es un sistema que obedece al principio de sobreposición donde la adición está remplazada por la convolucion

Para una señal $x[n] = x_1[n] * x_2[n]$, la señal de salida $y[n]$ es:

$$y[n] = H[x[n]] = H[x_1[n] * x_2[n]] =$$
$$= H[x_1[n]] * H[x_2[n]] = y_1[n] * y_2[n]$$

Un sistema homomórfico puede ser representado por una secuencia of 3 sistemas homomórphicos. El primero convierte la señal de ingreso obtenido por convolucion y lo transforma en una adición. El segundo es un sistema lineal que obedece al principio de sobreposición. El tercero es el inverso del primer sistema



Representación con la transformada discreta de Fourier DFT

Una de las propiedades más importantes de la convolucion es el hecho que el producto de convolucion se puede representar con una multiplicación en el dominio de la frecuencia:

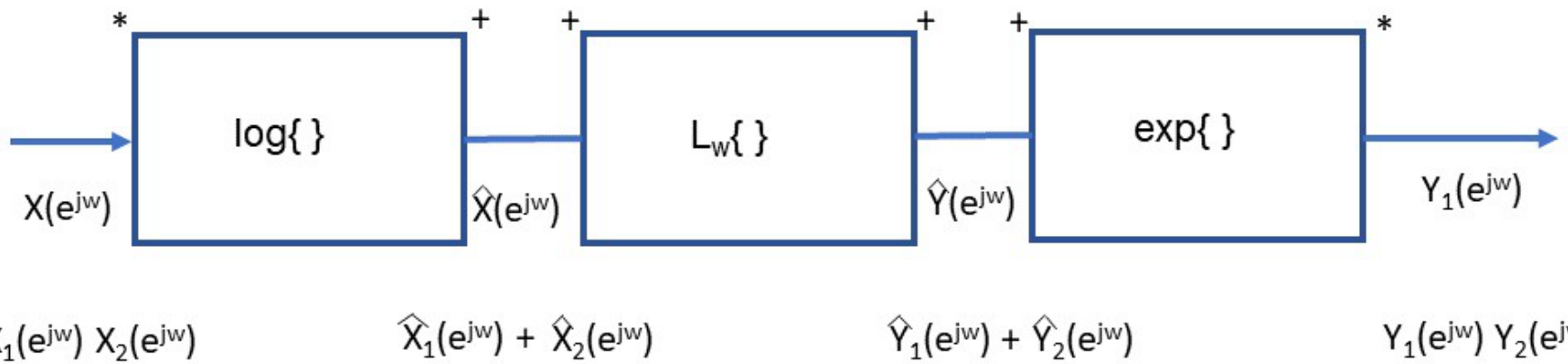
$$x[n] = x_1[n] * x_2[n] \rightarrow X(e^{j\omega}) = X_1(e^{j\omega}) \cdot X_2(e^{j\omega})$$

Entonces el sistema homomórfico se puede representar en el siguente modo:

$$\hat{X}(e^{j\omega}) = log\{X(e^{j\omega})\} =$$

$$= log\{X_1(e^{j\omega}) \cdot X_2(e^{j\omega})\} = log\{X_1(e^{j\omega})\} + log\{X_2(e^{j\omega})\} =$$

$$= \hat{X}_1(e^{j\omega}) + \hat{X}_2(e^{j\omega})$$


*
+
+
+
+
*
log{ }
$L_w\{ \}$
exp{ }
$X(e^{jw})$
$\hat{X}(e^{jw})$
$\hat{Y}(e^{jw})$
$Y_1(e^{jw})$
$X_1(e^{jw})\ X_2(e^{jw})$
$\hat{X}_1(e^{jw}) + \hat{X}_2(e^{jw})$
$\hat{Y}_1(e^{jw}) + \hat{Y}_2(e^{jw})$
$Y_1(e^{jw})\ Y_2(e^{jw})$

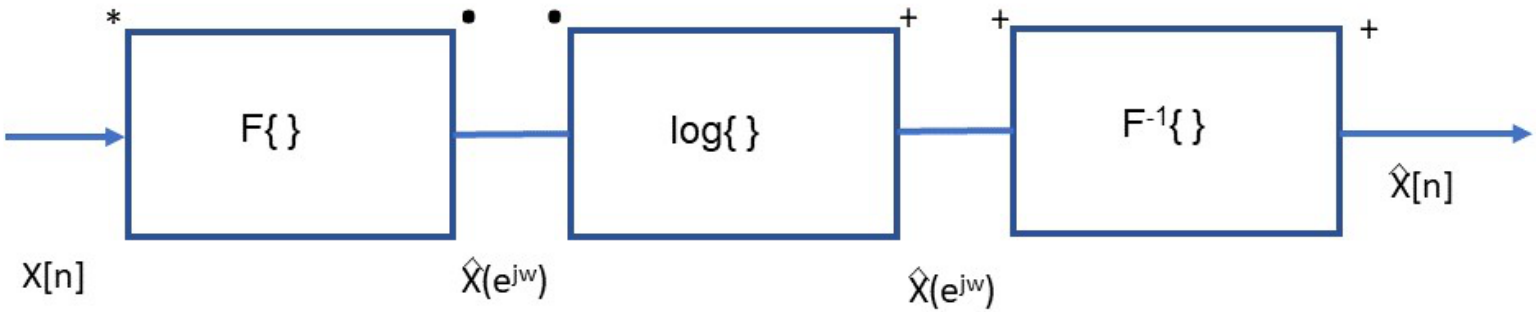
Representación con DFT para representar la señal como una secuencia

Si se quiere representar la señal como una secuencia, se puede utilizar la seguiente configuración de 3 ecuaciones:

$$X(e^{j\omega}) = \sum_{n=-\infty}^{+\infty} x[n]e^{-j\omega}$$

$$\hat{X}(e^{j\omega}) = log\{X(e^{j\omega})\}$$

$$\hat{x}[n] = \frac{1}{2\pi}\int_{-\pi}^{+\pi} \hat{X}(e^{j\omega})e^{jn\omega}$$



Donde F y F^{-1} son respectivamente la transformada de Fourier y la transformada inversa de Frourier. La señal de salida $\hat{x}[n]$ es el cepstrum complejo, asi llamdo no porque es complejo, sino porque se basa sobre un logaritmo complejo

Ambigüedad en la representación de un algoritmo complejo

Un ángulo en el plano complejo siempre es ambiguo por un múltiplo de 2π. Este problema no se aplica al exponencial complejo (inverso del logaritmo complejo), porque se representa de la siguiente manera:

$$X(e^{j\omega}) = e^{log(|X(e^{j\omega})|) + jarg\{X(e^{j\omega}))\}} =$$

$$= e^{log|X^{(}e^{j\omega})|}\, e^{jarg\{X^{(}e^{j\omega})\}} =$$

$$=\mid X(e^{j\omega}) \mid e^{jarg\{X^{(}e^{j\omega})\}}$$

Dada esta ecuación, se puede concluir que para cada entero múltiple de 2π, el valor del exponencial complejo no cambia.

Representación con la transformada Z

El sistema homomórfico describido con la DFT se puede obviamente representar tambien con la transformada Z:

$$\hat{X}(z) = log\{X(z)\} =$$

$$\hat{X}(z) = log\{X_1(z) \cdot X_2(z)\}$$

$$= log\{X_1(z)\} + log\{X_2(z)\}$$

Ejemplo: Con una secuencia exponencial

$x_1 = a^n u[n]$, con $|a| < 1$

Solución

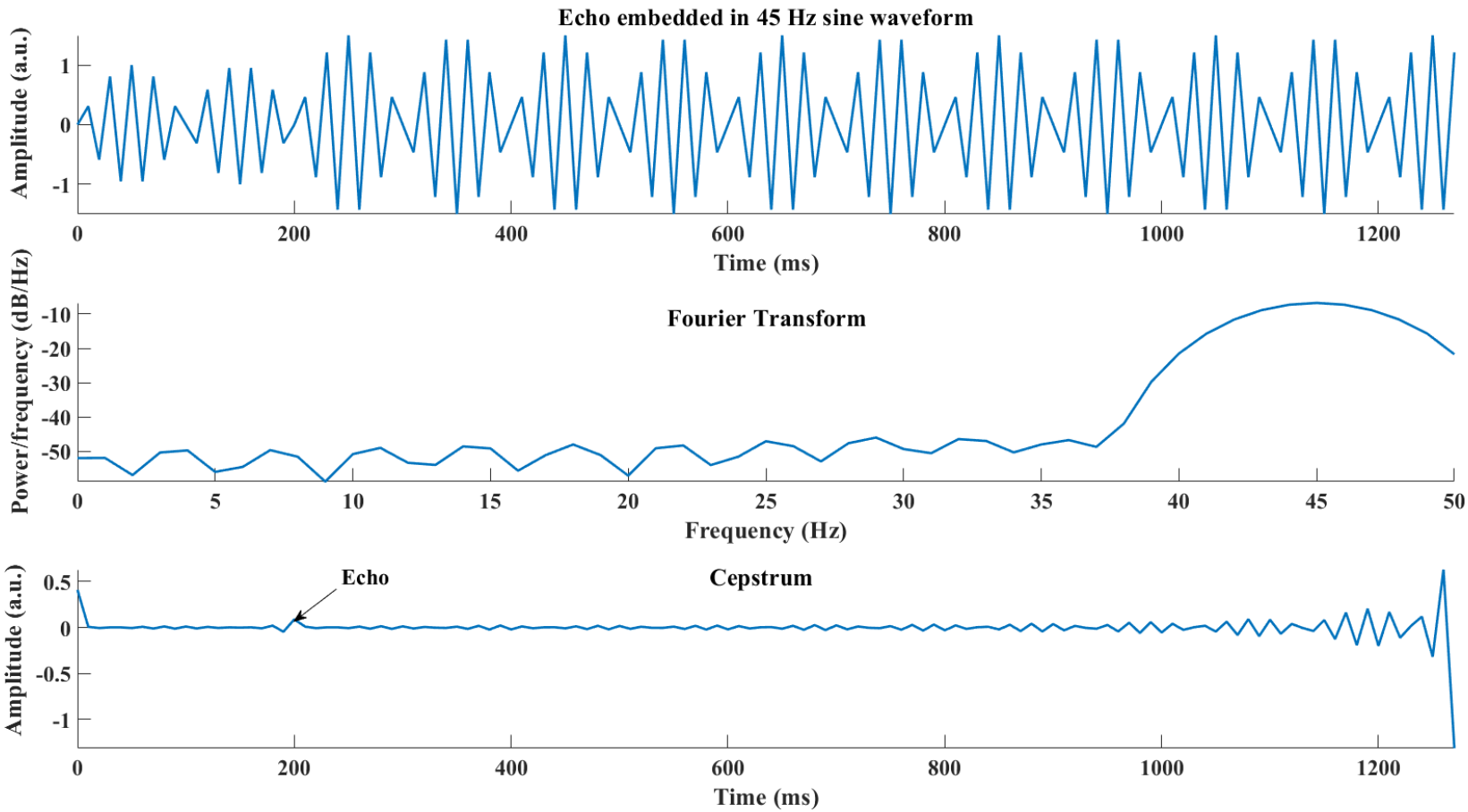
Transformada Z:

$$X_1(z) = \sum_{n=0}^{+\infty} a^n z^{-n} = \frac{1}{1 - az^{-1}}$$

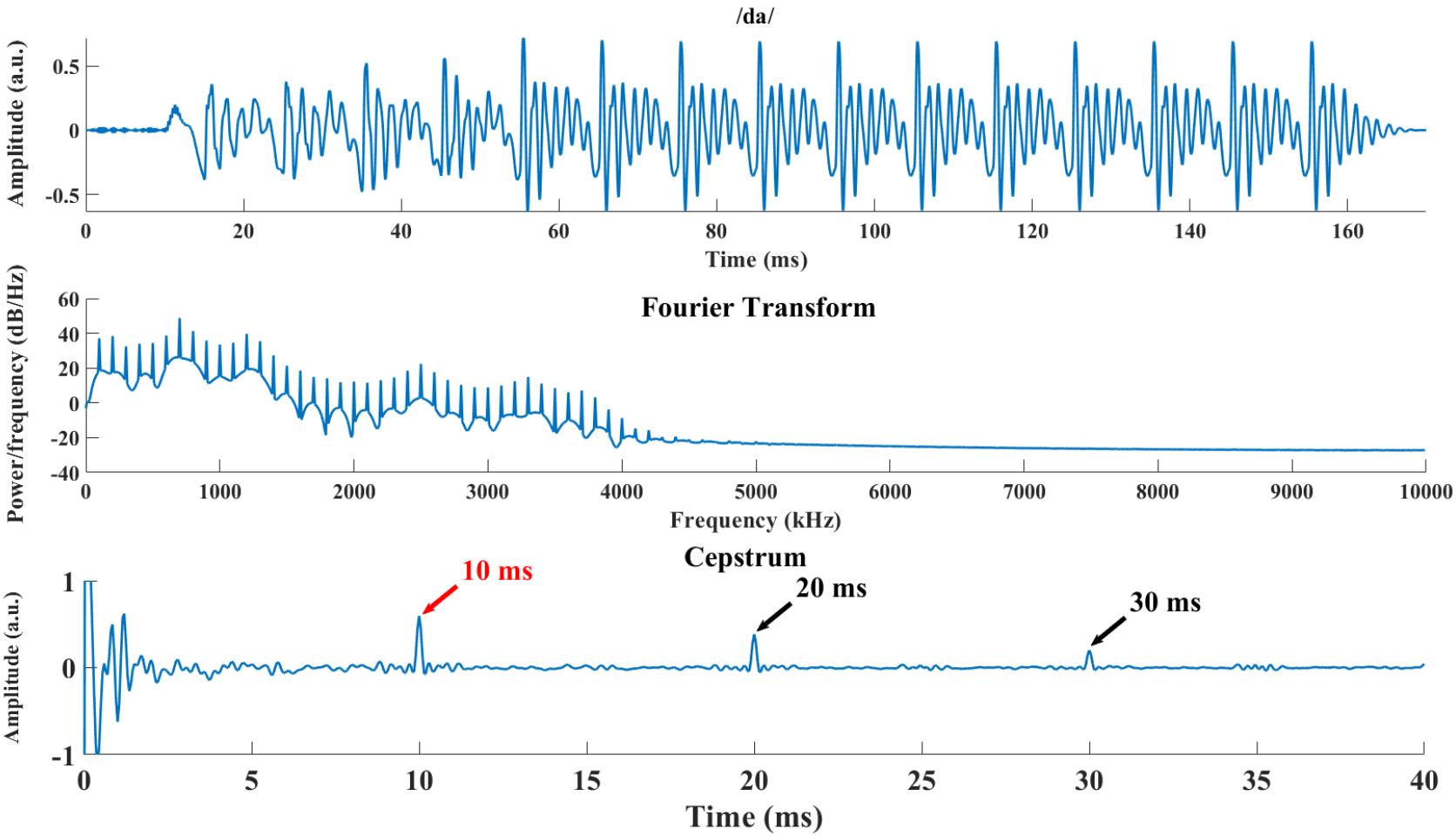
$$\hat{X}_1(z) = log[X_1(z)] = -log(1 - az^{-1}) = \sum_{n=1}^{+\infty} \frac{a^n}{n} z^{-n}$$

El cepstrum complejo es: $\hat{x}_1 = \frac{a^n}{n} u[n-1]$

Ejemplo: Detectar la presencia del echo en una onda sinusoidal con frecuencia de 45 Hz



Ejemplo: Estimar la frecuencia de timbre de la siguiente señal "da"



Cepstrum y coeficientes de Mel

Mel-Frequency Cepstrum (MFC)

MFC es una representación del espectro de potencia a corto plazo de un sonido, basada en una transformada de coseno lineal de un espectro de potencia logarítmica en una escala de frecuencia mel no lineal.

Mel-frequency cepstral coefficients (MFCCs)

MFCCs son coeficients que forman el MFC. La diferencia entre el cepstrum y el cepstrum de frecuencia mel es que en el MFC, las bandas de frecuencia están igualmente espaciadas en la escala mel.

Ecuaciones para el Mel-Spectrum

El Mel-spectrum de la m^{th} ventana de la DFT se define de la siguiente manera:

$MF_m[r] = \frac{1}{A_r} \sum_{k=L_r}^{U_r} \mid V_r[k] X_m[k] \mid^2, r = 1, 2, ..., R$

con R = numero de filtros (40 by default en Matlab).

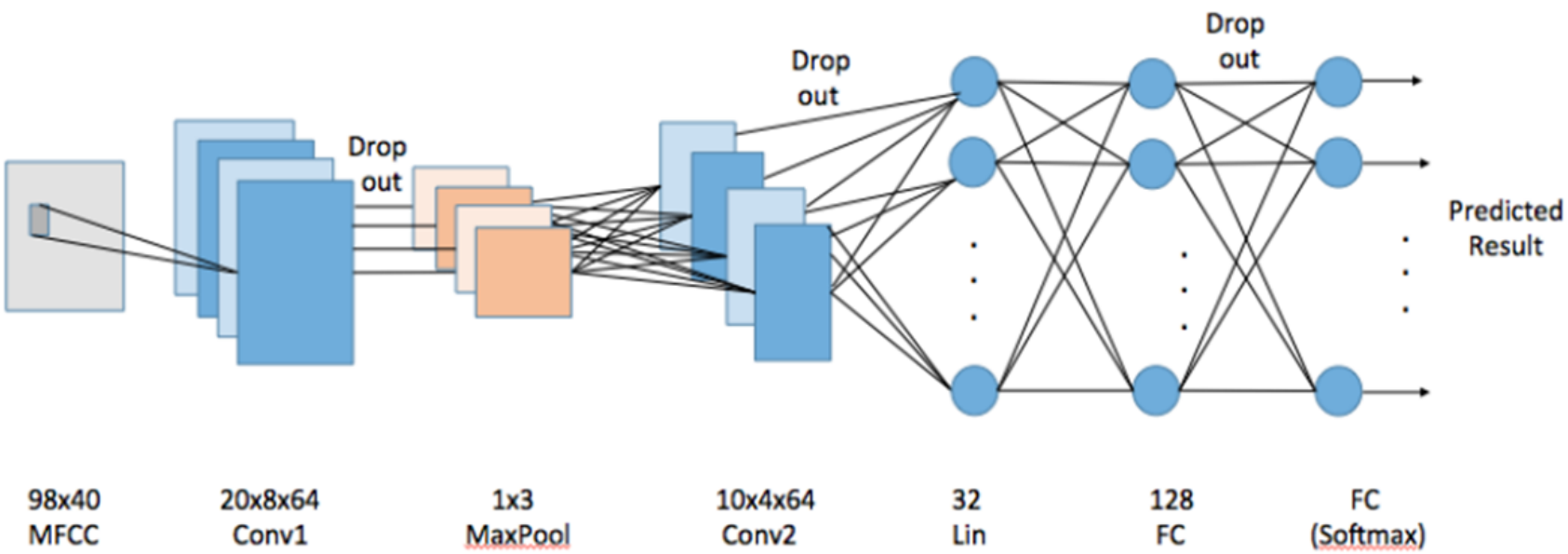
$V_r[k]$ = funcion de peso de la DFT.

$A_r = \sum_{k=L_r}^{U_r} \mid V_r[k] \mid^2$ = factor de normalización para el r^{th} filtro de mel, que se utiliza para obtener uno espectro de mel aplanado

Para cada ventana, m, se calcula una transformada de coseno discreta del logaritmo de la magnitud de las salidas del filtro para formar la función $mfcc_m[n]$:

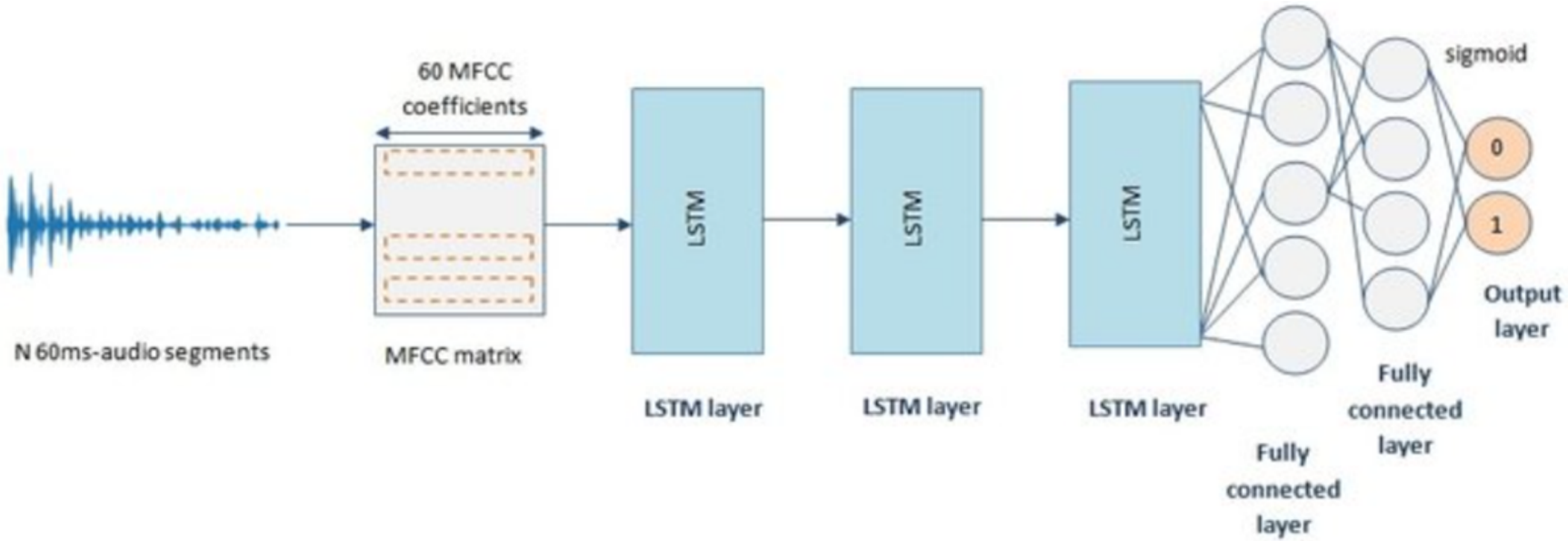
$$mfcc_m[n] = \frac{1}{R} \sum_{r=1}^{R} log(MF_m[r]) cos[\frac{2\pi}{R}(r + \frac{1}{2})n]$$

Ejemplo: MFCC utilizado como señal de ingreso para Convolution Neural Network (CNN)



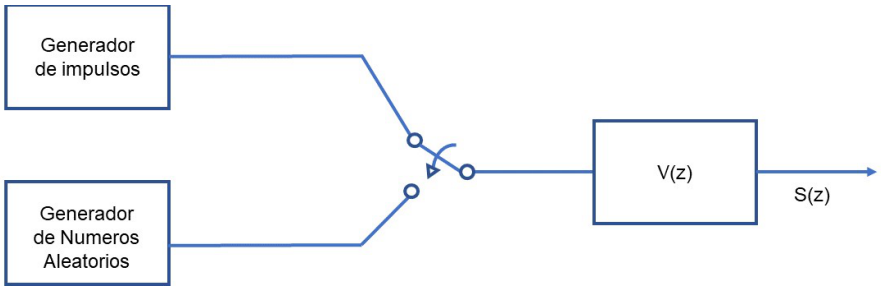
Li et al (2017) "*Speech Command Recognition with Convolutional Neural Network*" (Computer Science)

Ejemplo: MFCC utilizado como señal de ingreso para a Long short-term memory (LSTM) artificial recurrent neural network (RNN)



Rejaibi et al (2019) "*MFCC-based Recurrent Neural Network for Automatic Clinical Depression Recognition and Assessment from Speech*" (arxiv.org/abs/1909.07208v1)

Uso del Cepstrum para el procesamiento de Voz.



$$S(n) = v(n) * p(n)$$

El objectivo es separar p(n) de s(n) para así hacer el reconocimiento usando solamente v(n) como patrón

$$S(z) = V(z) \cdot P(z)$$

$$\hat{S}(z) = log[V(z) \cdot P(z)] = log[V(z)] + log[P(z)]$$

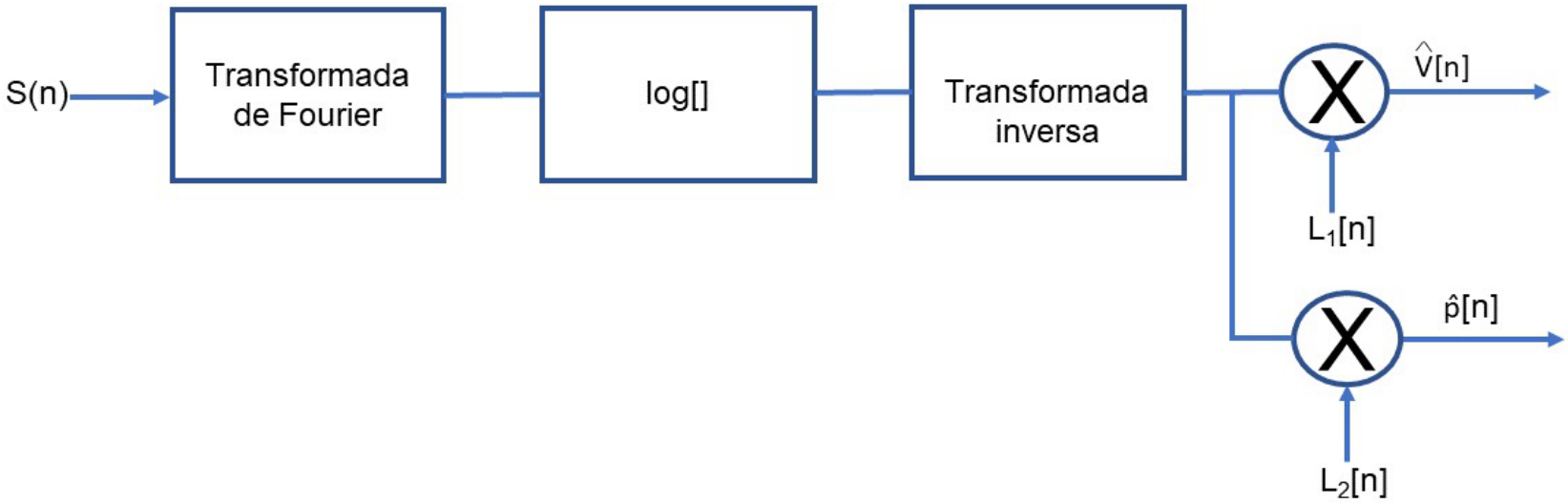
$$\hat{s}(n) = \hat{v}(n) + \hat{p}(n)$$

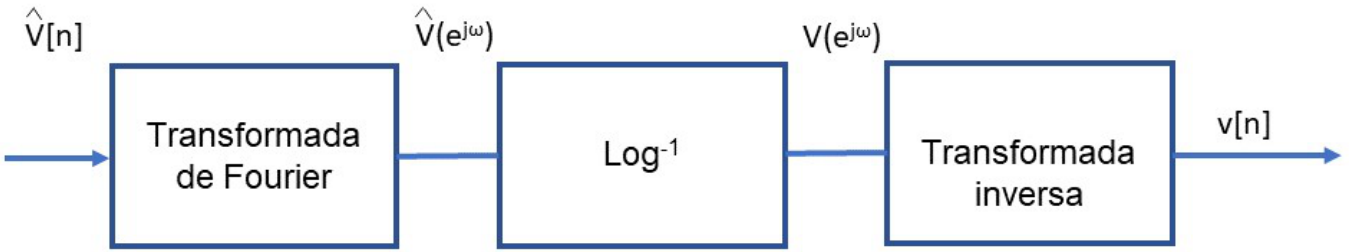
Si $\hat{s}(n)$ y $\hat{p}(n)$ tienen una region de suporte que no se traslapa, entonces se podrian separar $\hat{s}(n)$ y $\hat{p}(n)$. Por ejemplo, si $\hat{v}[n] = 0$ para $n \geq n_0$ y $\hat{p}(n) = 0$ para $n < n_0$, entonces

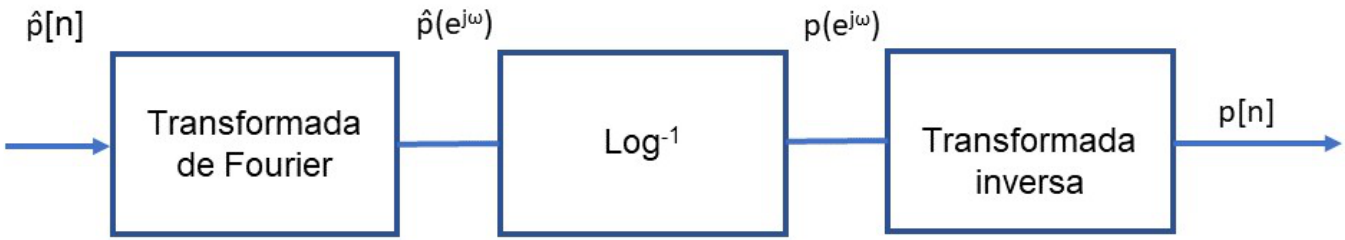
$$\hat{v}[n] = \hat{s}[n]l[n]$$

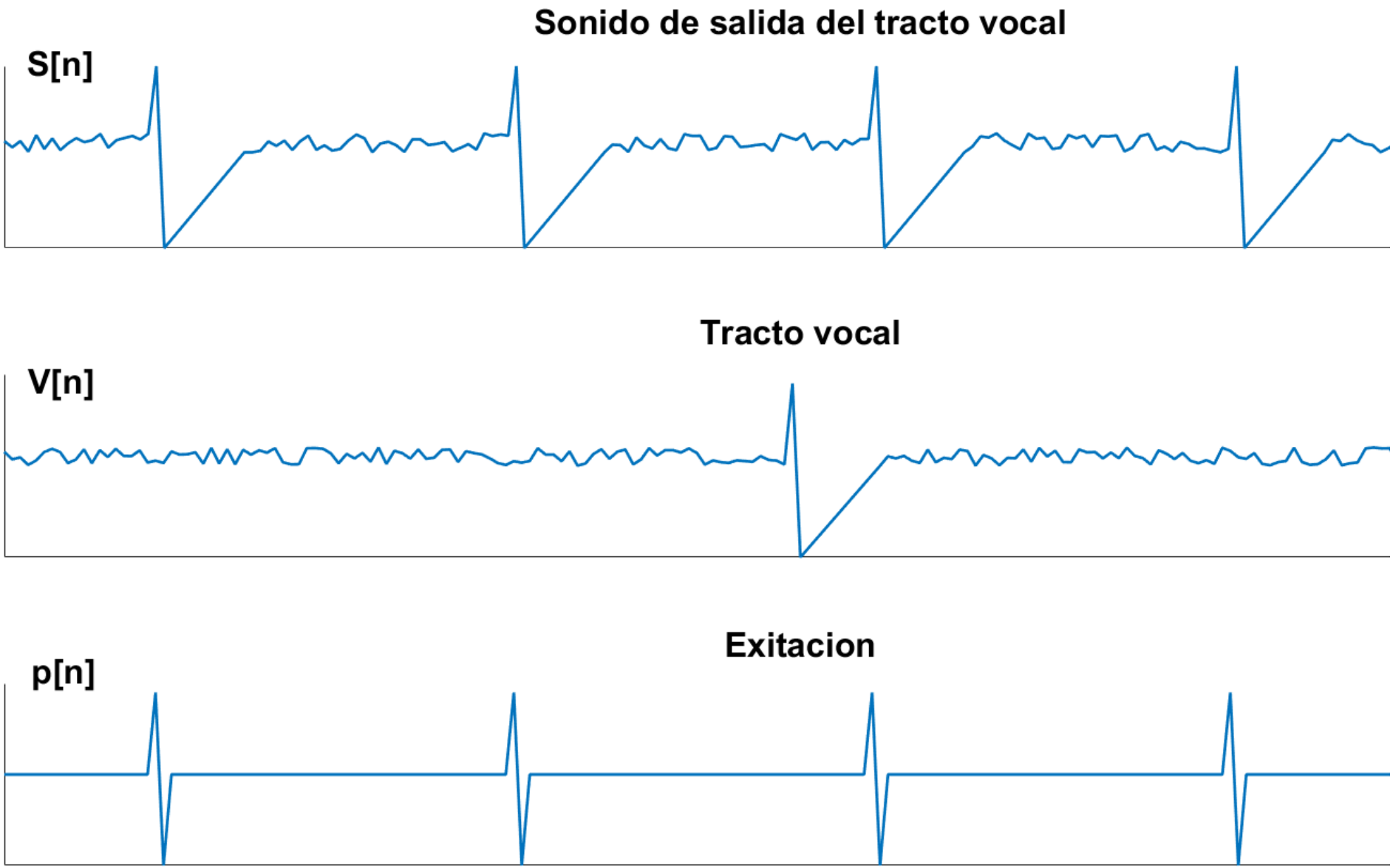
con

$$l[n] = \begin{cases} 1, & n < n_0 \\ 0, & n \geq n_0 \end{cases}$$


S(n)
Transformada de Fourier
log[]
Transformada inversa
$\hat{V}[n]$
$L_1[n]$
$\hat{p}[n]$
$L_2[n]$


$\hat{V}[n]$
$\hat{V}(e^{j\omega})$
$V(e^{j\omega})$
Transformada de Fourier
Log^{-1}
Transformada inversa
v[n]


$\hat{p}[n]$
$\hat{p}(e^{j\omega})$
$p(e^{j\omega})$
Transformada de Fourier
Log^{-1}
Transformada inversa
p[n]


Sonido de salida del tracto vocal
S[n]
Tracto vocal
V[n]
Exitacion
p[n]

Para $V(z) = \frac{G}{A(z)}$ (usando la expansion de Laurent) obtenemos el seguiente resultado

$log[\frac{G}{A(z)}] = logG + \sum_{n=1}^{\infty} a_k z^{-k}$

Diferenciando ambos los lados con respecto a z^{-1} e igualando los coeficientes para z^{-1}, se puede derivar la seguiente recursion:

$$C_n = -a_n - \frac{1}{n} \sum_{k=1}^{n-1} k c_k a_{n-k}$$

para $a_0 = 1$ y $a_k = 0$ para $k > p$

Usando la serie de Taylor para:

$log[\frac{G^2}{|A(e^{j\omega})|^2}] = \sum_{n=-\infty}^{+\infty} c_n e^{-jn\omega}$

$c_0 = logG^2$ y $C_{-n} = C_n$

Medidas de distancia usando Cesptrum

$$d_2^2(s, s') = \int_{-\pi}^{+\pi} |logS(\omega) - logS'(\omega)|^2 \frac{d\omega}{2\pi} =$$

$$= \int_{-\pi}^{+\pi} | \sum_{n=-\infty}^{+\infty} C_n e^{-jn\omega} - \sum_{n=-\infty}^{+\infty} C'_n e^{-jn\omega}|^2 \frac{d\omega}{2\pi} =$$

$$= \int_{-\pi}^{+\pi} | \sum_{n=-\infty}^{+\infty} (C_n - C'_n) e^{-jn\omega}|^2 \frac{d\omega}{2\pi} =$$

$$= \int_{-\pi}^{+\pi} \sum_{n=-\infty}^{+\infty} \sum_{m=-\infty}^{+\infty} (C_n - C'_n)(C^*_m - C'^*_m) e^{-j(n-m)\omega} \frac{d\omega}{2\pi} =$$

$$= \sum_{n=-\infty}^{+\infty} \sum_{m=-\infty}^{+\infty} (C_n - C'_n)(C^*_m - C'^*_m) \int_{-\pi}^{+\pi} e^{-j(n-m)\omega} \frac{d\omega}{2\pi} =$$

$$= \sum_{n=-\infty}^{+\infty} \sum_{m=-\infty}^{+\infty} (C_n - C'_n)(C^*_m - C'^*_m)\delta(n-m) =$$

$$= \sum_{n=-\infty}^{+\infty} \mid C_n - C'_n \mid^2 = \sum_{n=-\infty}^{+\infty} (C_n - C'_n)^2$$

S y S' son espectros de potencia, los cuales son funciones pares, por lo tanto sus coeficientes cepstrales son reales