Directional Audio Coding (DirAC)

Marc Franco Meca

1 Maig 2020

1 STFT

Com bé s'ha dit anteriorment, mitjançant la transformada de Fourier FT es pot obtenir informació sobre la distribució d'energia d'un senyal $x(t)$ a través de les seves components frequencials. Per això, aquesta eina matemàtica és molt útil per l'anàlisi de senyals estacionaries ja que proporciona una bona resolució freqüencial. Ara bé, si l'objectiu és determinar de forma precisa quan o on estan presents les diferents components en freqüència, aquesta proporciona una mala resolució temporal, fent-la inadequada per a senyals no estacionàries on el contingut espectral varia amb el temps.

Per a resoldre aquest problema de resolució temporal s'introdueix l'eina de la transformada de Fourier amb finestra (Short-Time Fourier Transform, STFT) que aporta una variació a la transformada de Fourier fent ús de l'enfinestrament. El procediment per a calcular la STFT consisteix en dividir el senyal $x(t)$ de temps en segments més curts d'igual longitud. D'aquesta manera es pot asumir que en cada un d'aquests segments el senyal és estacionari i es pot procedir a calcular la transformada de Fourier.

En temps continu, el senyal $x(t)$ o funció a ser transformada es multiplicada per una funció de finestra la qual és diferent de zero per un instant de temps molt petit. Aleshores, la transformada de Fourier del senyal resultant es desplaça per l'eix temporal fent una representació de dos dimensions del senyal. Aquest procés es pot expressar matemàticament com [1],

$$STFT\left[x(t)\right](\tau, w) \equiv X(\tau, w) = \int_{-\infty}^{\infty} x(t)w(t-\tau)e^{-iwt}dt \tag{1}$$

on $w(t)$ és la funció de finestra, comunament una finestra de Hann o una Gaussiana centrada en zero, $x(t)$ és la senyal a la qual es vol aplicar la transformada, $X(\tau, w)$ és la transformada de Fourier de $x(t)w(t-\tau)$, una funció complexe que representa la fase i la magnitud del senyal sobre el temps i la freqüència.

En temps discret, el senyal x(t) o funció a ser transformada es pot dividir en

fragments que normalment es solapen entre ells per a reduir les possibles irregularitats en els límits. A cada una d'aquestes divisions s'aplica la transformada de Fourier i el resultat s'emmagatzema en una matriu de valors complexes que conté la magnitud i la fase per a cada punt en temps i en freqüència. Matemàticament aquest procediment s'expressa mitjançant la següent expressió,

$$STFT\left[x[n]\right](m,w) \equiv X(m,w) = \sum_{n=-\infty}^{\infty} x[n]w[n-m]e^{-jwn} \tag{2}$$

on $x[n]$ és el senyal i $w[n]$ la finestra, m és de temps discret i w és continu, tot i que normalment ambdues es quantitzen i són de temps discret,

De la mateixa manera que la transformada de Fourier, la STFT és invertible. Per tant, el senyal original es pot recuperar de la transformada mitjançant la transformada inversa STFT, la qual ve donada per la següent expressió [2],

$$x(t)w(t-\tau) = \frac{1}{2\pi}\int_{-\infty}^{\infty} X(\tau,w)e^{jwt}dt \tag{3}$$

2 Directional Audio Coding (DirAC)

El Directional Audio Coding (DirAC) és un mètode paramètric de representació del so en l'espai que sosté que no es necessari reproduir perfectament el camp sonor, sinó que és suficient en reproduir-lo perfectament en el sentit en què els humans el percebem. El model de DirAC assumeix que amb una representació temporal i frequencial similar a la del sistema auditori humà, és adequat codificar i decodificar el camp sonor local amb un conjunt d'àudios i amb dos paràmetres que són el la direcció d'arribada (direction of arrival, DOA) de l'energia sonora incident i la difusió (diffuseness). El mètode de DirAC consta de diverses fases però en el nostre cas la fase que ens interessa és l'anàlisi. En aquesta fase es fa una estimació de la DOA i la difusió en una única localització depenent del temps i de la freqüència. La primera es relaciona amb la localització i la direcció, mentre que la difusió es relaciona amb la reverberació o l'extensió de la font sonora representada per la coherència interaural.

2.1 B-Format

La matriu de micròfons utilitzada en el mètode de DirAC ha de permetre l'anàlisi de la direcció i la difusió d'una àmplia regió de freqüència. Es per això que es fa ús de senyals B-format com a senyals d'entrada estàndard per a aquest mètode. Aquest format representa una matriu de micròfons situats en la mateixa localització espacial i produeix quatre canals de micròfons amb diferents característiques direccionals. Un primer micròfon omnidireccional el qual és proporcional a la pressió p, i els altres tres micròfons amb directivitat figura de vuit situats ortogonalment als eixos de les coordenades cartesianes x, y i z respectivament, que son proporcionals a les components de velocitat acústica v_x, v_y i

v_z. Els senyals corresponents obtinguts per a cadascun d'aquest micròfon són $w(t)$, $x(t)$, $y(t)$ i $z(t)$ respectivament, on t és l'índex temporal.

Aquest senyal B-format pot ser creat sintèticament a partir d'una gravació mono o gravada utilitzant micròfons especials que disposen d'aquesta característica.

En cas de sintetitzar el camp sonor a partir d'un senyal mono s(t), el senyal corresponent al micròfon omnidireccional W(t) s'escala per un factor 1/ 2 mentre que els altres tres senyals que contenen la informació de la direcció s'expressen en funció de l'angle d'elevació i azimut, representant-se així els senyals amb les següents expressions matemàtiques [3],

$$\begin{aligned} W(t) &= \frac{s(t)}{\sqrt{2}} \\ X(t) &= s(t) \cdot cos\phi \cdot cos\delta \\ Y(t) &= s(t) \cdot sin\phi \cdot cos\delta \\ Z(t) &= s(t) \cdot sin\delta \end{aligned} \tag{4}$$

on $s(t)$ és un senyal mono, ϕ és l'azimut i δ l'angle d'elevació de la font sonora.

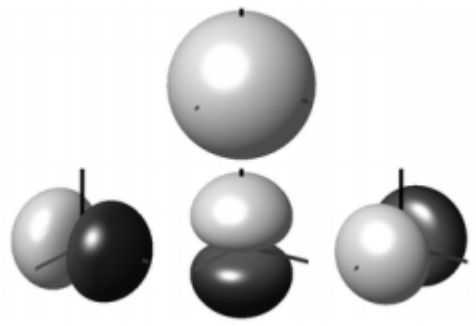

Figure 1: Representació 3D dels patrons de directivitat polar d'un micròfon omnidireccional a la part superior i tres figures de vuit a la part inferior direccionats cap a cada un dels eixos. [4]

2.2 Anàlisi DirAC

Els paràmetres de DirAC s'extreuen a partir de l'anàlisi de l'energia basada en la pressió sonora $p(t)$ i la velocitat acústica $u(t)$ en la posició de gravació.

El processat es dur a terme de manera separada per a bandes de freqüència les quals s'han d'obtenir del senyal original. Per tant, el primer pas per al càlcul

dels paràmetres consisteix en passar a domini freqüencial el senyal. Per a fer-ho normalment s'utilitzen dos plantejaments que es corresponen amb fer ús de filtres de bandes o de la transformada de Fourier amb finestra STFT. El sistema auditori humà separa les freqüències audibles en bandes de diferent tamany. La STFT conté una resolució fixe, és a dir, les de freqüència que s'obtenen són del mateix tamany mentre que el filtres de bandes possibiliten crear una transformació similar a la proporcionada pel sistema auditori. Tot i aquesta possibilitat d'aproximar-nos a la recreació del sistema auditori humà i donat que un dels objectius del DirAC és la no necessitat de reproduir perfectament la percepció humana, en el nostre cas s'ha decidit emprar la STFT per la seva rapidesa i eficàcia en els resultats.

En el domini de la STFT, es denota la pressió sonora p com P(k,n) i la velocitat acústica u com U(k,n), on k i n són els índex de la transformada en freqüència i temps. Com s'ha dit anteriorment, el micròfon omnidireccional capta un senyal proporcional a la pressió sonora i el figura de vuit obté un senyal proporcional a la velocitat. Aquestes magnituds físiques es relacionen amb el senyal B-format per el següent conjunt de relacions [5],

$$\begin{aligned} P(k,n) &= W(k,n) \\ \vec{U}(k,n) &= -\frac{1}{\sqrt{2}Z_0}\vec{X}'(k,n) \end{aligned} \tag{5}$$

on $\vec{X}'(k,n) = [X(k,n)Y(k,n)Z(k,n)]^T$ és el vector dels senyals provinents del gradient de pressió en B-format, $Z_0 = c\rho_0$ és la impedància de l'aire i el factor $\sqrt{2}$ és degut a la convenció dels senyals B-format.

Assumint que per a cada frame de freqüència el camp sonor és estacionari, i està format per una ona plana i un camp difús perfecte, es pot fer una estimació de la direcció de l'ona plana mitjançant el flux d'energia, expressat pel vector d'intensitat activa [6].

$$\vec{I}_a(k,n) = \frac{1}{2}\Re\left\{P(k,n)\cdot\vec{U}(k,n)^*\right\} \tag{6}$$

Utilitzant les equacions anteriors, es pot expressar el vector $\vec{I}_a$ en termes de senyal B-format com,

$$\vec{I}_a(k,n) = -\frac{1}{2\sqrt{2}Z_0}\Re\left\{W(k,n)\cdot\vec{X}'(k,n)^*\right\} \tag{7}$$

La direcció d'arribada DOA de l'ona sonora s'estima com el vector oposat al vector d'intensitat activa, ja que volem la direcció en què l'objecte es troba respecte el micròfon, i s'expressa com,

$$\vec{u}_{DOA}(k,n) = -\frac{\vec{I}_a(k,n)}{\left\|\vec{I}_a(k,n)\right\|} \tag{8}$$

o en termes de senyal B-format com,

$$\vec{u}_{DOA}(k,n) = \frac{\Re\left\{W(k,n)\cdot\vec{X}'(k,n)^*\right\}}{\left\|\Re\left\{W(k,n)\cdot\vec{X}'(k,n)^*\right\}\right\|} = \begin{bmatrix} cos\phi\cdot cos\delta \\ sin\phi\cdot cos\delta \\ sin\delta \end{bmatrix} \tag{9}$$

on $\phi(k,n)$,$\delta(k,n)$ són les estimacions dels angles incidents d'azimut i d'elevació respectivament i $\|\cdot\|$ és la norma euclidiana.

Un cop obtingut el vector unitari que es correspon amb la direcció d'arribada de la font sonora es converteix a coordenades esfèriques. Existeixen diverses variacions del sistema de coordenades esfèriques en l'àmbit matemàtic i físic, però en el nostre cas s'ha utilitzat el sistema de coordenades més utilitzat en audio espaial. En aquest sistema, r és la distància al punt des del centre de coordenades, $\phi\epsilon[-\pi,\pi]$ és l'angle azimut, que descriu la posició en l'eix horitzontal del pla, i $\delta\epsilon[-\pi/2,\pi/2]$ és l'elevació o altitud, i descriu la posició en l'eix vertical del pla.

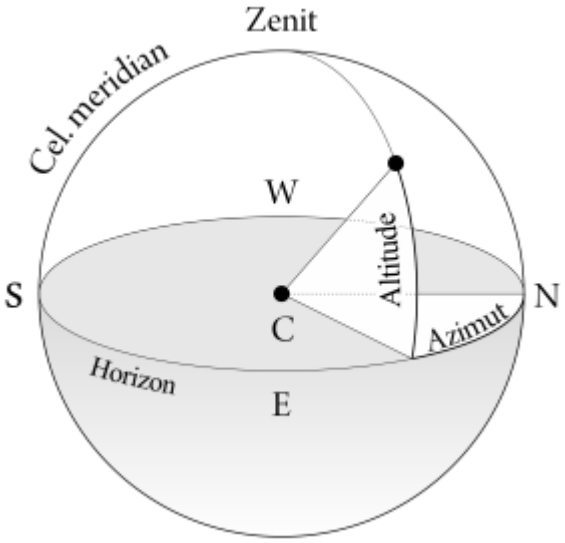

Figure 2: Sistema de coordenades esfèriques utilitzat en astronomia, equivalent al sistema utilitzat en el nostre cas. La coordenada N es correspon amb l'eix positiu X i l'altitud és equivalent a l'elevació[7]

Qualsevol punt de l'espai expressat en coordenades cartesianes (x,y,z) es pot referenciar a coordenades esfèriques (r,ϕ,δ) utilitzant les següents expressions matemàtiques [3],

$$
\begin{aligned}
r &= \sqrt{x^2 + y^2 + z^2} \\
\phi &= arctan\frac{y}{x} \\
\delta &= arcsin\frac{z}{r}
\end{aligned}
\tag{10}
$$

Un cop calculat el primer paràmetre DOA, l'altre paràmetre a estudiar es correspon amb la difusió, la qual es pot definir com una proporció d'energia sonora que oscil·la localment [8]. Aquest paràmetre de difusió mostra quant directiu és el camp sonor prenent valors entre 0 i 1. Una ona plana té un valor de difusió mínim quan el transport d'energia es correspon amb la densitat total d'energia, i màxim per a un camp sonor completament difús, quan el transport d'energia és nul. Es defineix la difusió del camp sonor com la relació entre la intensitat i la densitat d'energia, definint-se aquest vector d'energia del camp sonor com [6]

$$
E(k,n) = \frac{\rho_0}{4}\left\|\vec{U}(k,n)\right\|^2 + \frac{1}{4\rho_0 c^2}\left|P(k,n)\right|^2 \tag{11}
$$

i en termes de senyal B-format com,

$$
E(k,n) = \frac{1}{4\rho_0 c^2}\left[\frac{\left\|\vec{X}'(k,n)\right\|^2}{2} + \left|W(k,n)\right|^2\right] \tag{12}
$$

i obtenint així la següent expressió per a calcular el paràmetre de difusió

$$
\psi(k,n) = 1 - \frac{\left\|\left\langle \vec{I}_a(k,n)\right\rangle\right\|}{c\left\langle E(k,n)\right\rangle} \tag{13}
$$

i expressat amb senyal B-format com,

$$
\psi(k,n) = 1 - \frac{\sqrt{2}\left\|\left\langle \Re\left\{W(k,n)\cdot\vec{X}'(k,n)^*\right\}\right\rangle\right\|}{\left\langle\left|W(k,n)\right|^2 + \left\|\vec{X}'(k,n)^2\right\|/2\right\rangle} \tag{14}
$$

on $\langle\cdot\rangle$ denota la mitjana en temps i $(^*)$ el conjugat.

References

[1] Daniel Griffin and Jae Lim. Signal estimation from modified short-time fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 32(2):236–243, 1984.

[2] Wikipedia contributors. Short-time fourier transform. [Online; accessed 1-May-2020].

[3] Daniel Arteaga. Introduction to ambisonics, 2015.

[4] Wikipedia contributors. Spherical harmonics. [Online; accessed 1-May-2020].

[5] Archontis Politis, Tapani Pihlajamäki, and Ville Pulkki. Parametric spatial audio effects. *York, UK, September*, 2012.

[6] Taylor Francis Frank J. Fahy, Sound intensity. Sound intensity. *London, 2nd edition*, 1995.

[7] Wikipedia contributors. Horizontal coordinate system. [Online; accessed 1-May-2020].

[8] Juha Merimaa and Ville Pulkki. Spatial impulse response rendering i: Analysis and synthesis. *Journal of the Audio Engineering Society*, 53(12):1115–1127, 2005.

[9] Ville Pulkki, Mikko-Ville Laitinen, Juha Vilkamo, Jukka Ahonen, Tapio Lokki, and Tapani Pihlajamäki. Directional audio coding-perception-based reproduction of spatial sound. In *International Workshop on the Principles and Applications of Spatial Hearing*, pages 1–4, 2009.

[10] Ville Pulkki. *Implementing a modular architecture for virtual-world Directional Audio Coding*. PhD thesis, Aalto University, 2013.