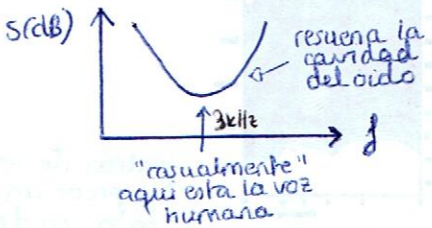
Sesión 1. Psicoacústica de la percepción espacial del sonido

• Introducción

- El modelo cerebral del oído es mas complejo que el visual
  ↳ localiza con rapidez y precisión el origen del sonido

- Sensibilidad del oído



• Localización espacial en el plano horizontal

- Diferencia de intensidad

IID

Difracción : sombra de la cabeza

No funciona bien si cabeza << λ

- Diferencia de tiempo interaural (ITD)

ITD

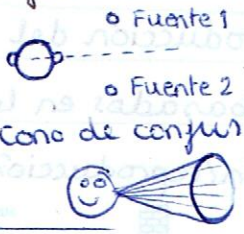
$ITD = \frac{r}{c}(\theta + \text{sen}\theta)$

La ITD se interpreta como una diferencia de fases (el oído detecta fase, no retardo)

⇓

Ambigüedad de fase a altas frecuencias → $f > 1'5$ kHz ⇒ Ya no sirve ITD

Confusión delante-detrás



Mecanismo combinado:
↳ transición gradual

- $f < 1'5$ kHz → ITD (detecta desfase)
- $f > 1'5$ kHz → IID (difracción cabeza)

f

• Localización espacial en elevación

Ambos oídos están en el mismo plano (no se puede detectar desfase en elevación)

- Filtrado de las orejas y hombro según la dirección
  ↳ ecos con retardos

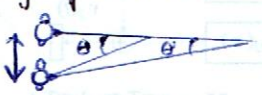
• Percepción de la distancia

→ Intensidad : $A_t = \frac{P_t}{P_r}$

- Física : propagación atenúa : el cerebro deduce la distancia
- Inteligencia : según el tipo de sonido (voz humana, pájaro, ...) el cerebro estima P_t

→ Atenuación de altas frecuencias :
- el aire tiene mas 'rozamiento' a frecuencias altas
- según la relación entre faltas y bajas el cerebro deduce la distancia

→ Paralaje por movimiento



se mueve ligeramente la cabeza y se compara el ángulo (típico para encontrar a un grillo)
(así es como se mide la distancia de las estrellas)

→ Exceso de diferencia de IID : para objetos muy cercanos → ej : mosquito en el oído

→ Relación entre sonido directo y reverberación : (en salas)

la energía del sonido directo depende de la distancia
la energía TOTAL de los ecos es mas o menos igual en todos los puntos de la sala

Función de transferencia relacionada con la cabeza (HRTF)

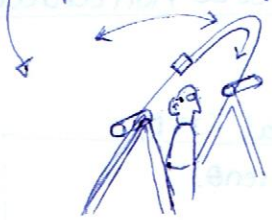
$HRIR(\phi,\theta,t) \underset{TF^{-1}}{\overset{TF}{\rightleftarrows}} HRTF(\phi,\theta,f)$

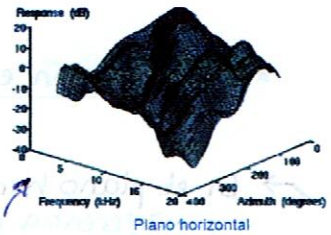
impulse response — transfer function

Una distinta para cada oído!! (¡nadie es simétrico!)

Medida:

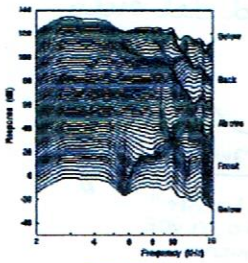
Fuente en ϕ,θ que emita un pulso $\delta(t)$





Plano horizontal

plano horizontal bastante parecido entre personas



Plano medio

zona suave (solo hombres)

patrón de valles y picos único para cada persona

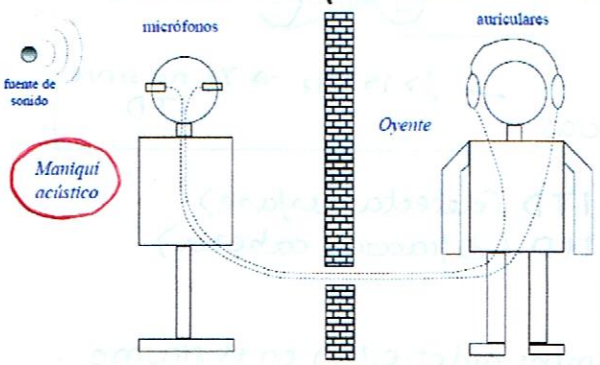
· Reproducción del sonido espacial

Obtener la señal { - grabación binaural / - síntesis binaural

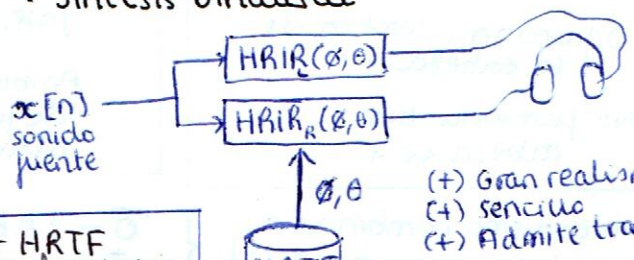
Reproducción { - cascos / - altavoces

→ Sistemas basados en la HRTF

· Grabación/reproducción binaural



· Síntesis binaural



- Medir HRTF para cada usuario
  - inviable
  - cámara anecoica
- Nearest-neighbour
  - el usuario elije
- Scale standard HRTF
  - dividir en partes (pinna, head, ...) para ir ajustando
- Adaptive model

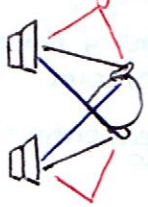
(+) Gran realismo
(+) Sencillo
(+) Admite tracking

(-) particularidad en la forma de la oreja

(-) auriculares generan efecto-dentro-cabeza

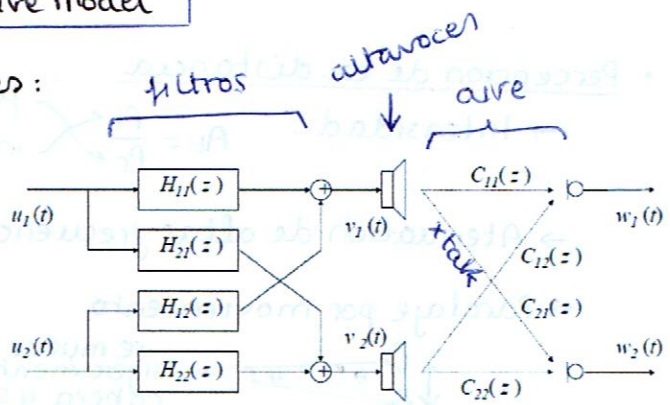
· Reproducción binaural con altavoces:

Problema
- crosstalk
- reflexiones

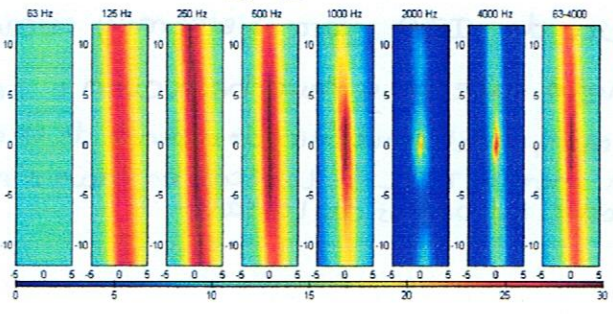


Solución: cancelar los caminos cruzados

↳ sólo sirve si el usuario no se mueve del sweet spot: estudios:
- teóricos en espacio libre
- medidos



$H = (C^T C)^{-1} C^T A$



moverse delante/detrás no afecta mucho

Degradación (relación directo/interferente) en función del desplazamiento

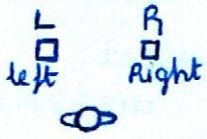
- En 2kHz hay singularidad debido a distancia entre oídos → único sweet spot

EVOLUCIÓN DEL SONIDO ENVOLVENTE

potenciómetro PAN-POT (Panorama Potenciometer)

MIC1: L — R

· Estéreo



Efecto phantom:
- Por diferencia de volumen el cerebro interpreta una posición entre L y R
- Problema: no hay diferencia entre tiempos de llegada
- Solución: el cerebro se acostumbra
- Problema 2: sólo funciona si estás centrado

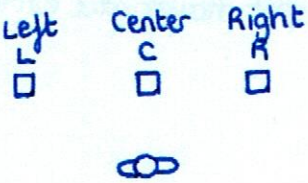
$x+y=1$

$L \leftarrow x \quad y \rightarrow R$

$V_L = y \cdot V$

$V_R = x \cdot V$

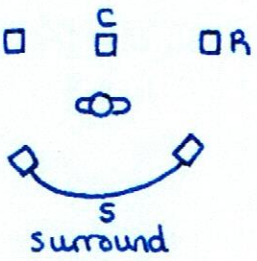
· Canal central para diálogo



Los espectadores no centrados se desorientaban al oír el diálogo a los lados.

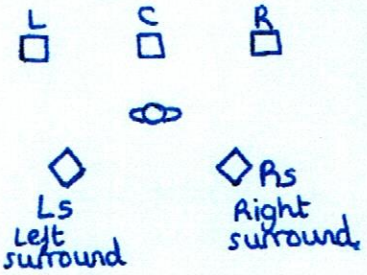
Canal central para el diálogo

· Canal surround



- Llena el ambiente
- Simula reverberación
- Efectos/sustos

· Desdoblar canal surround



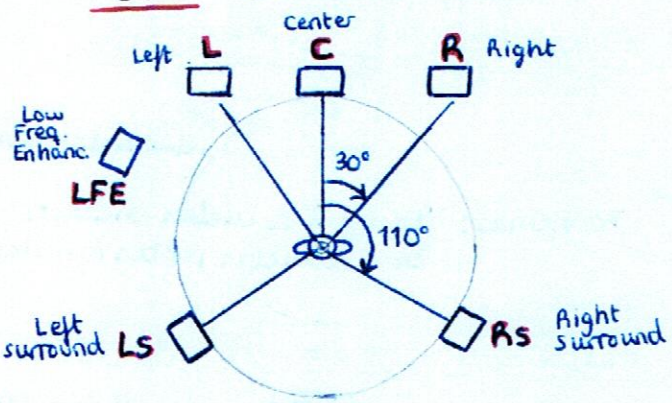
Permite el efecto phantom detrás

· LFE: Low frequency enhancement

usa gran altavoz (subwoofer) para frecuencias [20Hz, 120Hz]

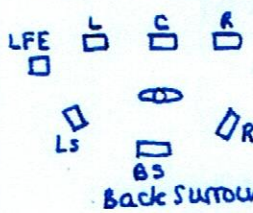
No requiere posicionamiento ya que apenas hay diferencia de fase y volumen entre los oídos a bajas frecuencias ($\lambda \uparrow\uparrow$ → difracción de cabeza $\downarrow\downarrow$)

5.1 (estándar ITU)



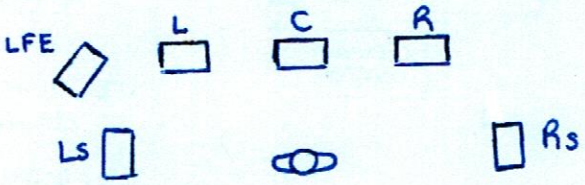
· Tercer canal de surround

6.1

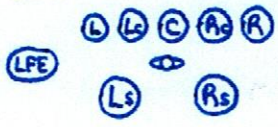


· Desdoblar el Back surround → 7.1

7.1



Otro 7.1 con 5 delante



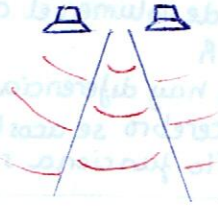
- más tolerante a la posición del usuario
- Poco exito
- compatible con 5.1 (matriz de compatibilidad)

Stereo: - phantom image created L-C-R

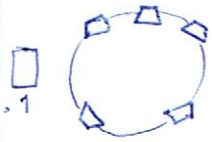
X-Y technique: variando intensidad ← Pan pot: el más usado

A-B technique: variando time-of-arrival

Stereo Sweet Spot:



Surround 5.1



→ pan-pot delante con 3 altavoces → sweet spot mayor que estéreo

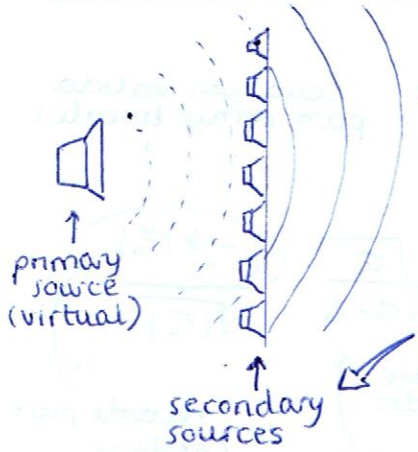
→ localización lateral y detrás es pobre

7.1 lo mejora, pero no lo suficiente

Solución → Wave Field Synthesis

SESIÓN 2. WAVE-FIELD SYNTHESIS → 2'5D (no tiene elevación)

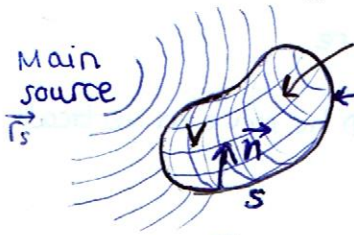
otra técnica aún mejor:
Ambisonics: descomposición del campo acústico en armónicos esféricos (con array de 32 micrófonos se capta hasta 4º orden)



Huygen's principle:

WFS: Theory Background

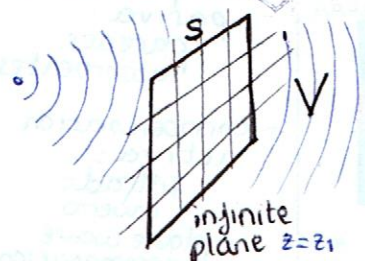
Kirchoff-Helmholtz integral equation



$$P(\vec{r},\omega) = \oint_S \frac{1}{4\pi}\left[P(\vec{r}_s,\omega)\cdot\frac{\partial}{\partial n}\left(\frac{e^{-jk|\vec{r}-\vec{r}_s|}}{|\vec{r}-\vec{r}_s|}\right) - \frac{\partial P(\vec{r}_s,\omega)}{\partial n}\left(\frac{e^{-jk|\vec{r}-\vec{r}_s|}}{|\vec{r}-\vec{r}_s|}\right)\right] ds$$

↑ presiones = monopolos = altavoces

↑ derivada de presiones = dipolos



separa el universo en dos

Rayleigh integral equation

$$P(\vec{r},\omega) = |z-z_1|\int_S \left[P(\vec{r}_s,\omega)\,\frac{1+jk|\vec{r}-\vec{r}_s|}{2\pi|\vec{r}-\vec{r}_s|^3}\,e^{-jk|\vec{r}-\vec{r}_s|}\right] dS_1$$

↖ ya no necesito dipolos!

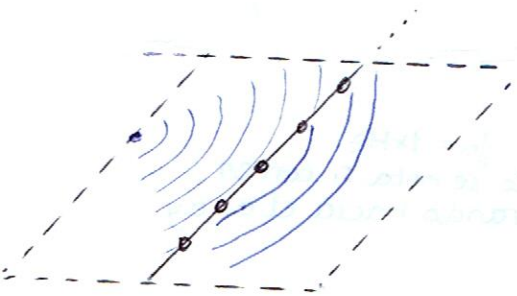
↓ Discretizar las fuentes

$$P(\vec{r},\omega) = |z-z_1|\sum_n \left[P(\vec{r}_n,\omega)\,\frac{1+jk|\vec{r}-\vec{r}_s|}{2\pi|\vec{r}-\vec{r}_s|^3}\,e^{-jk|\vec{r}-\vec{r}_s|}\right]\Delta x\,\Delta y$$

↖ plano infinito de altavoces

↓ Simplification to a line array

↳ suponer que cada altavoz emite onda cilíndrica en lugar de esférica (no es cierto, pero el error es pequeño si estamos en el plano de los altavoces)



$$P(\vec{r},\omega) = \frac{jk\rho c}{2\pi}\sum_n u_n(\vec{r}_n,\omega)\,\frac{e^{-jk|\vec{r}-\vec{r}_n|}}{|\vec{r}-\vec{r}_n|}\,\Delta x$$

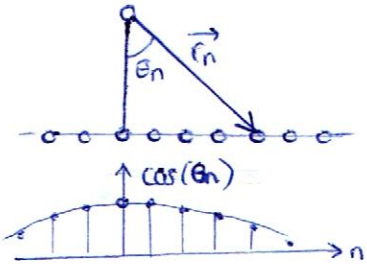
Para N altavoces

$$P(\vec{r},\omega) = \sum_{n=1}^{N} \left[Q(\vec{r}_n,\omega) \cdot G(\theta_n,\omega) \cdot \frac{e^{-jk|\vec{r}-\vec{r}_n|}}{|\vec{r}-\vec{r}_n|} \right] \Delta x$$

directividad altavoz

Ecuación válida para arrays lineales

Excitación de cada altavoz



$$Q(\vec{r}_n,\omega) = \underbrace{S(\omega)}_{\text{señal mono}} \cdot \underbrace{\frac{\cos(\theta_n)}{G(\theta_n,\omega)}}_{\text{controla amplitud}} \cdot \underbrace{\sqrt{\frac{jk}{2\pi}}}_{\text{preacentuación de los agudos}} \cdot \sqrt{\frac{\sigma_0}{\sigma_0+\rho_0}} \cdot \frac{e^{-jk|\vec{r}_n|}}{\sqrt{|\vec{r}_n|}}$$

error por no usar ondas esféricas

retardo puro (desfase variable con la frecuencia)

Limitaciones:

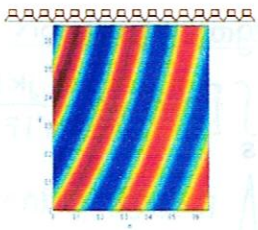
→ Muchos altavoces: cada uno con su señal y su hardware

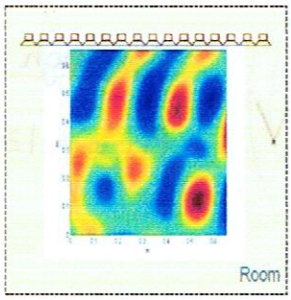
→ Limitación: no se puede jugar con elevación?
  ↳ ¿Y si hacemos phantom effect en elevación? – No, IID no existe en vertical
  ↳ Mejor solución: combinar con HRTF

→ Tamaño de array limitado (truncation)

→ Efectos de la sala

Ecos estropean el efecto

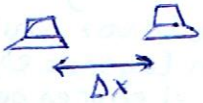

Free field


Room

Soluciones:
- compensación pasiva: paredes absorbentes
- compensación activa: filtrado inverso
  - plane wave decomposition
  - multichannel inversion

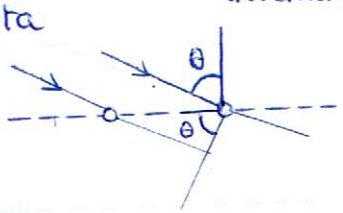
→ Spatial aliasing

· separación entre altavoces



· Sólo se puede sintetizar bien hasta

$$\Delta x \cos\theta_{max} < \frac{\lambda_a}{2}$$



$$f_a = \frac{c}{2\Delta x \cos\theta_{max}}$$

Soluciones:

· Ignorar el problema → en el ITEAM $\Delta x = 18$cm, $f_a = 1$kHz
  la fase errónea solo se nota si corres lateralmente y mirando hacia el array

· usar mas altavoces agudos

· OPSI (Optimized Phantom Source Image)
  ↳ Usar WFS para $f < f_a$
  ↳ Usar efecto phantom para $f > f_a$

· Usar DML (Distributed Mode Loudspeaker)
  ↳ membrana que se mueve entera con MAP (multi actuator panels)

WFS in GTAC
- array de 32 (4x8)
- array de 96 (12x8) → de los mas grandes de Europa
- MAP (DML) → 4 paneles x 6 excitadores

Applications of WFS

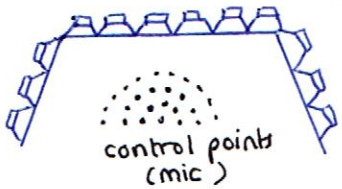
- Telepresencia
- Realidad virtual
- Reproducción de música con fidelidad espacial
- Planetarium, IMAX, ... (short term)
  movie theaters (medium term)
  home cinema (long term)

Research lines for WFS

- room compensation
- Discrete Time Modelling (FDTD)
- Authoring & Real Time Tools
- High-power applications
- Improvements in DMLs

Compensación de la sala

- Plane-wave decomposition →  emitir los ecos "al revés"
- MIMO inverse filters



- measure impulse response between each loudspeaker and each control point.
- obtain bank of inverse filters
- Simulate sound field

corrección multipunto → se han obtenido buenas correcciones en un gran área

- Simulación FDTD
  - Micrófonos virtuales para obtener RIR
  - Superordenadores

- High-power applications
  - Diameter of loudspeakers
  - Directivity

Experiments & simulations:
- Field analysis
- Precise aliasing Frequencies

Authoring tools for WFS

- standalone application
- VST plugin that works with 'Cubase'

Software Block Diagram

SESIÓN 3. SOUND SOURCE SEPARATION (SSS)

· INTRODUCTION

- Recovering each source signal from a given mixture
- Difícil → active research

Aplications:
- remixing
- speech enhancement (e.g. hearing devices)
- speech recognition
  ↳ singer recognition and melody extraction
- Wave-Field-Synthesis

Evolution:

Beamforming (mic. array)
- problem in audio
  ↓↓ ⇒ N·Δx ↑↑ } N ↑↑
  ↓↑ ⇒ Δx ↓↓
  ↙ separación baja
  ↓ gran apertura

ICA: coste computacional elevadísimo

CASA: Perceptual sound processing principles based on listening experiments
problem: very specific e.g. speech singing voice

Sparse methods:
Idea: baja probabilidad de que fuentes distintas sean simultáneas en tiempo o frecuencia o ambos
↳ time/frequency masking

· MIXTURES

→ Audio Sources

- Voz: secuencia de fonemas
  ↳ cuerdas vocales (pitch fundamental + armónicos) ← parciales
  ↳ ruidos de susurro (tambien tiene formantes)
  ↳ transitorios (labios)

  Male around 140 Hz
  Female 200 Hz

- Music: musical instruments + singers
  ↳ tonos/notas = harmonic partials → acordes = mezcla notas
  ↳ transient signals (ej: tambor)

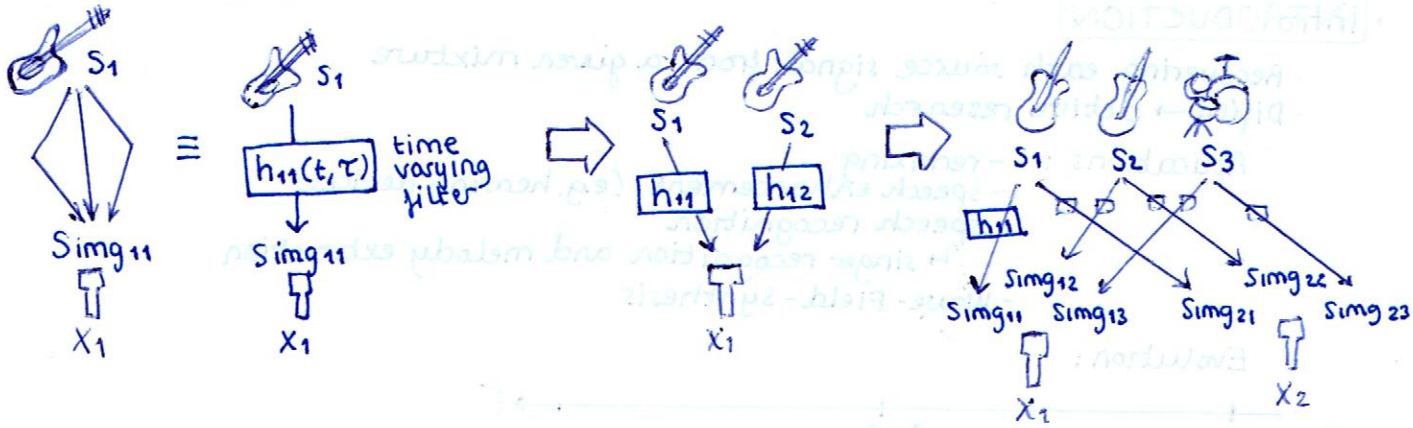
- Sonidos naturales: muy distintos ej: pito de coche
  ej: lluvia

- Studio recordings: multitrack + downmix recording

- Live recordings: Puede intentarse la separación con micros
  ej: 2 micros X-Y → apuntando ↑ ↳ cardioide ↳ en 8
  ej: 2 micros spaced → sencillo y rápido ↑ ↑

- Synthetic mixtures: tabla de mezclas

The mixing process

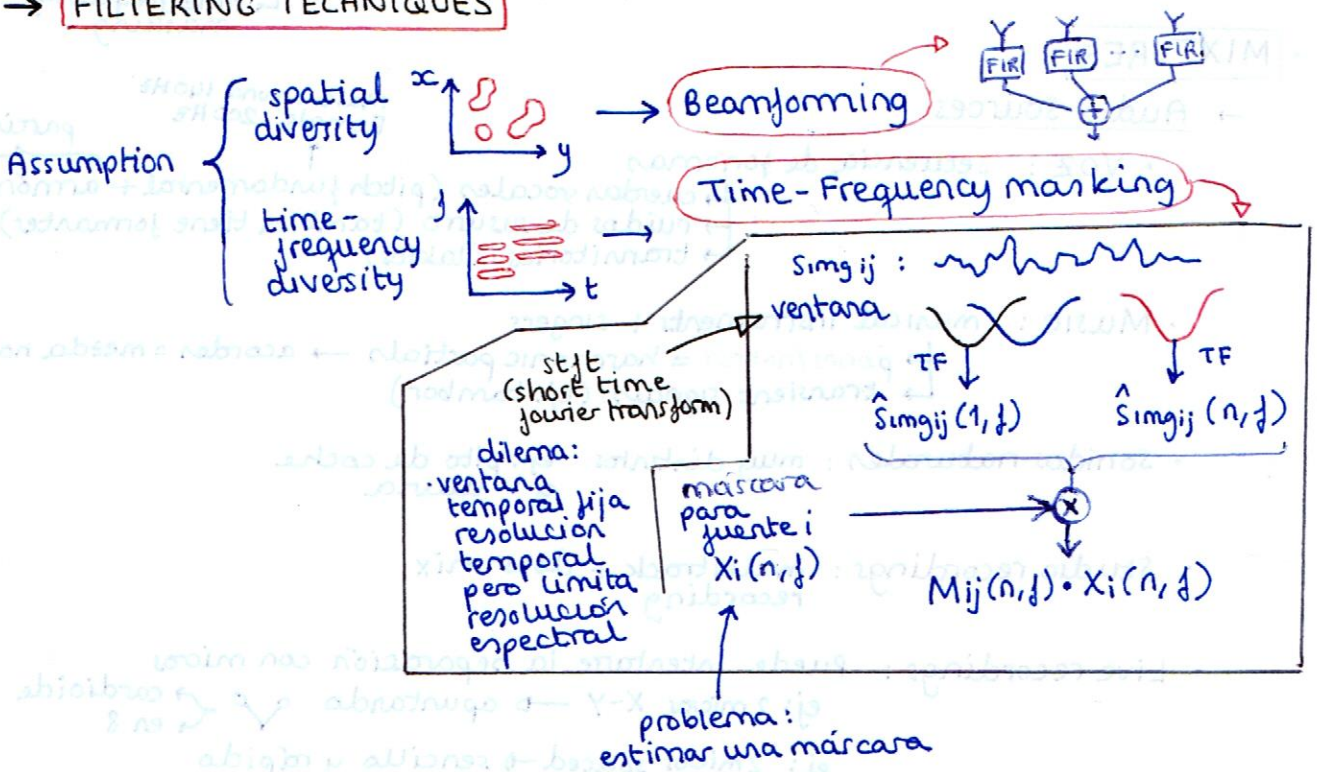


- En sss típicamente se intenta estimar $Simg_{ij}(t)$ ← efecto sala
- si trato de estimar directamente $s_j(t)$ se llama derreverberación

$$x_i(t) = \sum_{j=1}^{J} \underbrace{Simg_{ij}(t)}_{s_j(t) * h_{ij}(t,\tau)}$$

- **BLIND SOURCE SEPARATION** → separar haciendo el menor número de suposiciones posible

→ **FILTERING TECHNIQUES**



→ **MULTI-CHANNEL SEPARATION** → You can exploit spatial clues

- **ICA (Independent Component Analysis)**

  Assumption: the sources are statistically independent
  - ↳ minimize mutual information between estimated source signals
  - ↳ find optimal demixing matrix (MAP)

  Miras estadísticos de alto orden (>2) (más allá de la varianza)

  ejemplo:
  - una gaussiana da cero
  - suma de 2 gaussianas no da cero, pero si las separo cada una da cero

- **ADRess (Azimuth Discrimination and Resynthesis)**

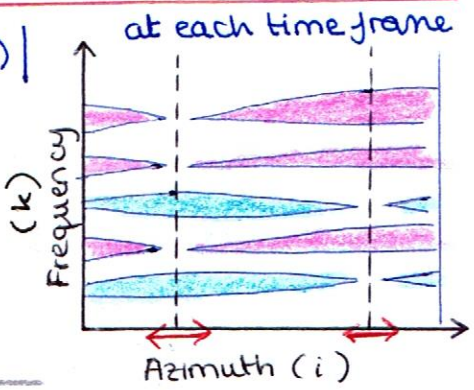
  Idea stereo:
  - Left $l(t) =$  (guitarra + voz)
  - Right $r(t) =$  (guitarra + voz)

  - $l(t) - g_j r(t)$ ⇨  causes the cancellation of the j-th source in the left channel
  - $r(t) - g_j l(t)$ causes the cancellation of the j-th source in the right channel

We know how to cancel a source, but how can we recover it?

① $Azr(k,i) = \left| L(k) - \underbrace{\frac{i}{\beta}}_{g(i)} R(k) \right|$ $\quad i \in [0, \beta]$

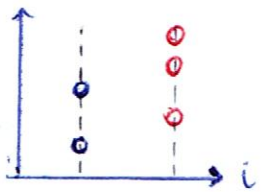
y lo mismo con canal left $Azl(k,i)$



② Convertir nulos en picos

Buscar picos en un azimuth subspace width
↳ ya sabemos a qué frecuencias está cada fuente en un determinado time frame

$$Az'(k,i) = \begin{cases} Az(k)_{\max_{\forall i}} - Az(k)_{\min_{\forall i}} & \text{if } Az(k,i) = Az(k)_{min} \\ 0 & \text{resto} \end{cases}$$



③ señal separada:

$$Y_R(k) = \sum_i Azr'(k,i)$$

$$Y_L(k) = \sum_i Azl'(k,i)$$

↑ módulo (la fase se cogerá del original $R(k)$, $L(k)$)

(+) Good quality of separation
(+) Works with stereo → casi todo está en estéreo
(+) Very simple

Problems:
- Sources with same pan position are extracted together
- Musical noise in extracted sources

• DUET (Degenerate Unmixing Estimation Technique)

Assumption: only one source is present at any time-frequency point (W-disjoint orthogonality)

↓

Problema:
- Rarely holds
- musical noise artifacts
- it cannot deal with convolutive mixes where mixing delay is larger than one sample

It is possible to determine the source in each point from the spatial location information

Interchannel Intensity Difference

$$IID(n,f) = 20 \log \frac{X_2(n,f)}{X_1(n,f)}$$

Interchannel Phase Difference

$$IPD(n,f) = \angle \left[\frac{X_2(n,f)}{X_1(n,f)} \right]$$

→ SINGLE CHANNEL SEPARATION

- Spatial clues cannot be included
- More challenging problem

• Hidden Markov Models (HMM)

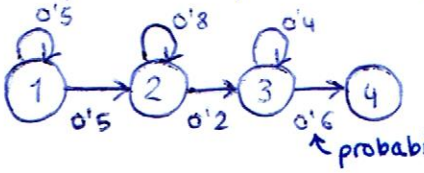
Markov model: states correspond to random processes whose outcomes are observable

HMM: the states are not observable directly but rather have a certain probability distribution of the observation

In Speech: – observations are significant speech parameters
– Each state models a particular phoneme or particular chord

e.g. HMM modelling a vowel

e.g. HMM modelling a vowel in a singing voice has 10 states



↑ probability of changing state

Three probabilities:
- A : state transitions
- B : distribution of observations for a given state
- π : probability of initial state

steps:
① : Train individual HMM beforehand to get A, B, π
② : Calculate state path (most likely succession of states) (e.g. Viterbi)
③ : Each state corresponds to a short-term spectrum of the source

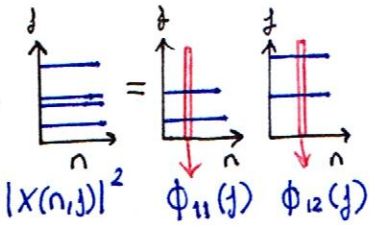
(+) Good separation of male and female voice
(+) Acceptable separation of singing voice
(-) Difficult to determine optimum number of states for each case
(-) Train beforehand with the solo voice

• Spectral decomposition

spectrum as weighted sum of normalised bases plus error

$$|X(n,f)|^2 = \left(\sum_j \sum_k \underbrace{e_{jk}(n)}_{\text{time varying coefficients}} \cdot \Phi_{jk}(f) \right) + \varepsilon(n,f)$$

e.g.:



the coefficients vary with n

↳ Non-negative matrix factorization

$$\underline{\underline{X}} = \underline{\underline{B}} \cdot \underline{\underline{G}}$$

$\underline{\underline{B}}$: basis functions
$\underline{\underline{G}}$: time varying coefficients

- Estimate B and G (non-negative)
- once estimated, each source will correspond to sets of basis spectra

- **Computational Auditory Stream Analysis (CASA)**

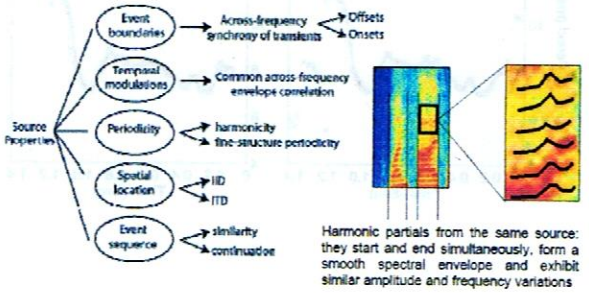
→ Inspirado en la capacidad de los humanos para distinguir sonidos
↳ experimentos de audición → progresos durante 10 años

① Descomponer la muestra en componentes tiempo-frecuencia
② Agrupar los componentes en sus correspondientes fuentes

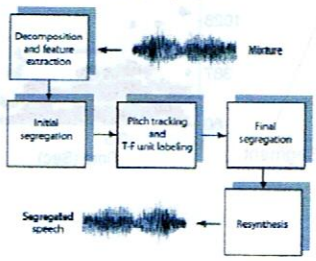
CASA : Resumen del Proceso

10 years of progress

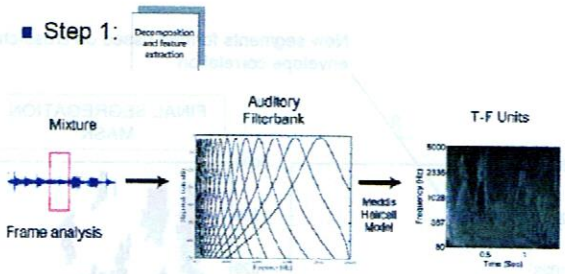
- 100 mixture set used in Cooke (1991)
- Speech + { white noise, tone, telephone, impulses, siren, music, speech, ... }
- Largely data-driven systems below
- Unrealistic corpus: mainly voiced speech + instrusion

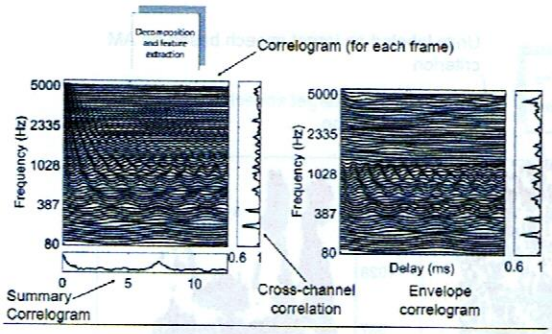


- Monaural Speech Segregation:

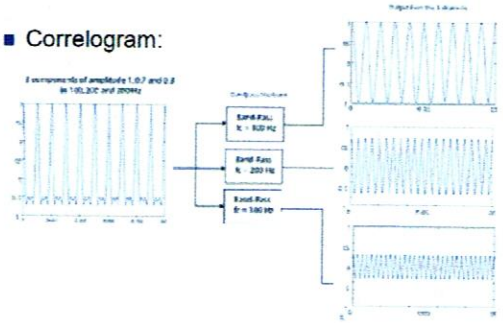


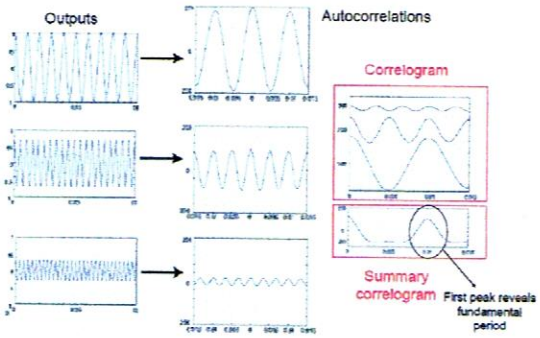
- Step 1:





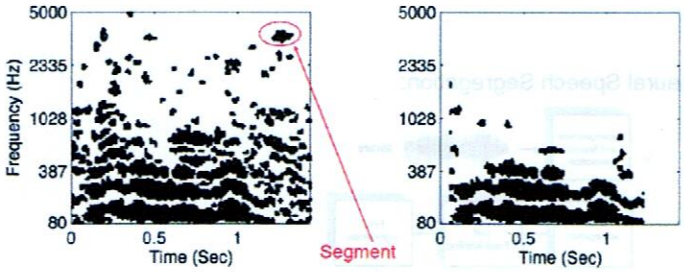
- Correlogram:



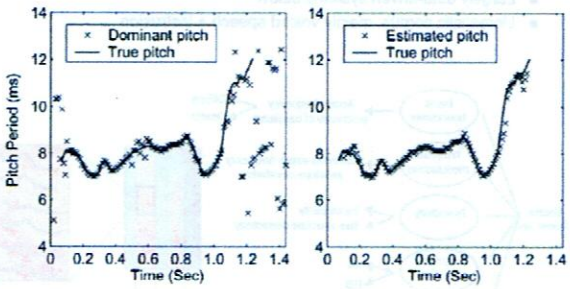


- **Step 2:** Initial segregation

If a T-F unit has correlagram at zero-lag and cross-channel correlation greater than a certain threshold, the unit is selected → Adjacent units are merged into segments



Pitch tracking → pitch is estimated in the plausible pitch range according to 2 constraints: agreement with the dominant pitch and slow pitch change.

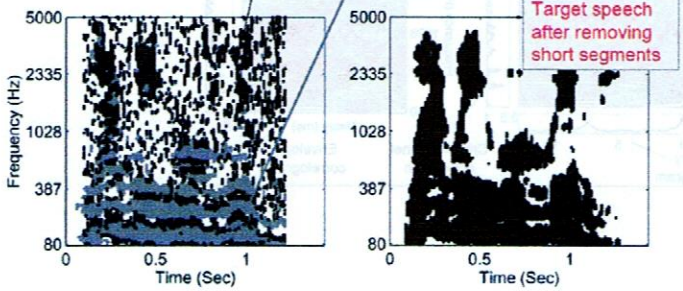


- **Step 3:** Pitch tracking and T-F unit labeling

Units labeled as target speech based on AM criterion

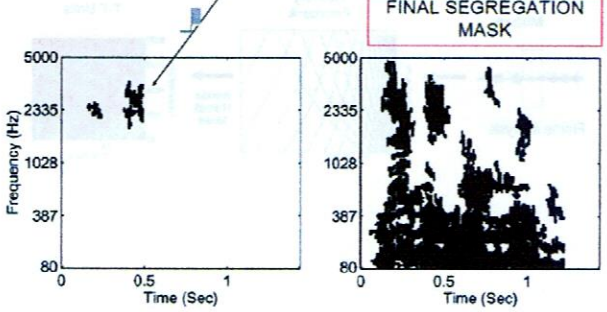
Units labeled as target speech based on periodicity criterion

Target speech after removing short segments



New segments formed based on cross-channel envelope correlation

FINAL SEGREGATION MASK



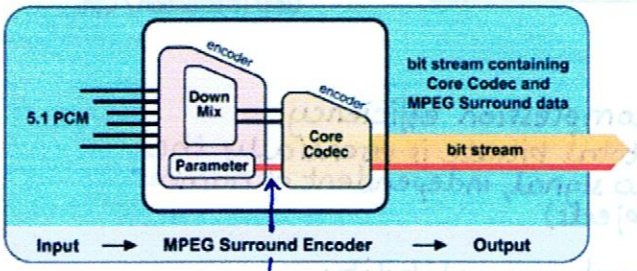
SESIÓN 4 : SPATIAL AUDIO CODING (SAO) SPATIAL AUDIO OBJECT CODING (SAOC)

Spatial Audio Coding (mp3 Surround)

(MPEG Surround)
(MPS)

CODEC:

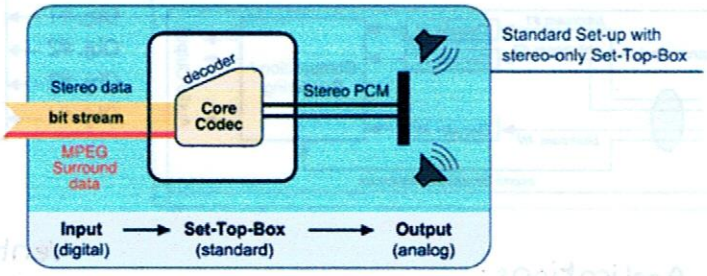
- Codifica multicanal en un stream stereo + info adicional ↑ spatial parameters
  El downmix puede ser automático o externo (a gusto del artista)



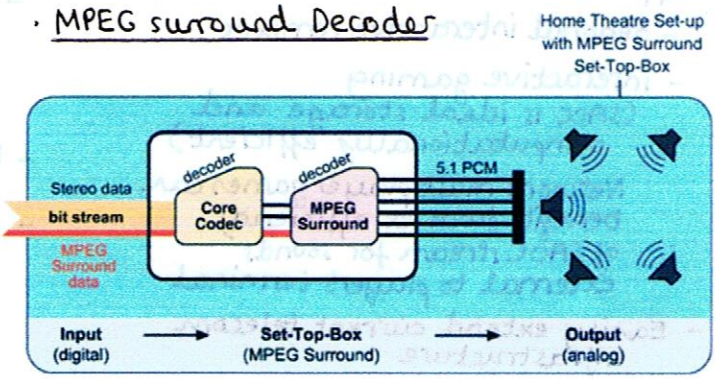
info adicional (reduced set of spatial parameters)

DECODEC

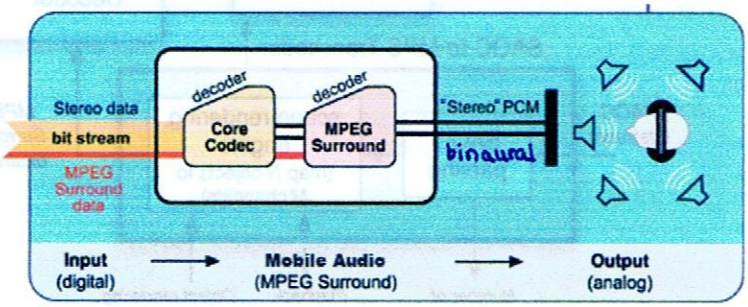
- Legacy decoder :
  - ignora info adicional
  - backward compatibility



- MPEG surround Decoder



- Capacidad adicional En Sonido
  - A partir del stream estéreo + parametros genera 2 canales con HRTF para usar cascos (perfecto para dispositivo móvil!)



lo hace "directamente" sin pasar por 5.1

Ventajas y aplicaciones

- Tasa de bits muy cercana a estéreo → bueno para broadcast IPTV
  - → Fácil de integrar por los broadcasters
  - → Está en estandares de broadcast ej: DAB
- Dispositivos móviles:
  - → surround binaural
- Music and video downloads
  - → same music file plays on conventional stereo player

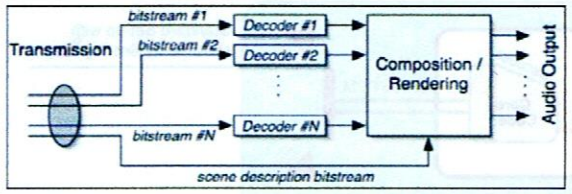
Spatial Audio Object Coding (SAOC)

Object based representation:

- Individual dry sources + scene description or user input

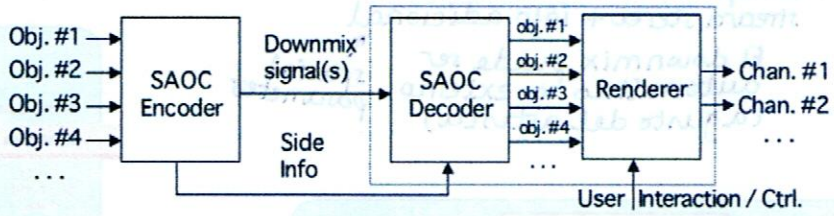
↓

Context based interactivity



SAOC busca lograr la misma funcionalidad que el Object Based Representation, de forma eficiente

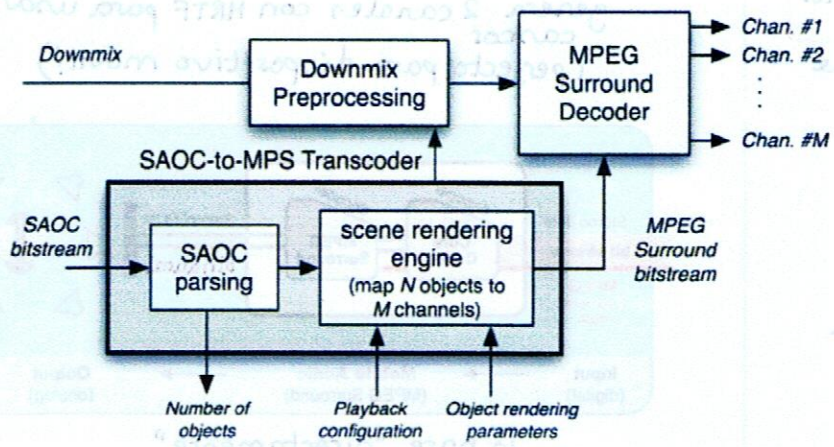
tx:
- stereo stream
- side info → ayuda a separar los distintos dry sources



Applications:
- personal interactive remixes
- interactive gaming (SAOC is ideal storage and computationally efficient)
  Network multiplayer games can benefit from tx efficiency of SAOC stream for sounds external to player's terminal
- Easily extend current telecom infrastructure

Ventajas:
- high compression efficiency (the total bitrate is essentially the stereo signal, independent of number of objects)
- backward compatibility
- el SAOC decoder + renderer pueden integrarse para evitar el paso intermedio de separación en objetos → reduce complejidad computacional

MPEG SAOC Architecture



Objective:
- Integrated Decoder to avoid complexity

N objects to M channels
N > M
by simply using an M-channel MPS decoder driven by parameters