

Procesamiento de la información paralingüística

- **Prosodia**
 - Registro del tono
 - Entonación, acento, y límites frasales
 - Emoción
- **Identificación del hablante**
- **Procesamiento multimodal**
 - Combinación del rostro y la ID (identificación) del hablante
 - Lectura de labios y reconocimiento de voz audiovisual
 - Gestos y comprensión multimodal

- **La prosodia es un término utilizado típicamente para describir aspectos extralingüísticos del discurso, como:**
 - Entonación
 - Límites frasales
 - Patrones de acentuación
 - Emoción
 - Afirmación/distinción de preguntas
- **La prosodia está controlada por la manipulación de:**
 - La frecuencia fundamental (F_0)
 - Las duraciones fonéticas y la velocidad de habla
 - La energía

Registro robusto del tono

- **Estimación de la frecuencia fundamental (F_0)**
 - Normalmente referido como *registro del tono*
 - Crucial para el análisis y el modelado de la prosodia del habla
 - Un problema muy estudiado con muchos algoritmos propuestos
- **Un algoritmo reciente de dos pasos (Wang, 2001)**
 - **Paso 1:** Estimar los tramos de F_0 y ΔF_0 , cada tramo de frecuencia basado en la combinación armónica
 - **Paso 2:** Realizar búsqueda dinámica con restricciones de continuidad para hallar el flujo óptimo de F_0

Transformada de Fourier logarítmica discreta

- Espectro de banda estrecha muestreado logarítmicamente:
 - Los picos armónicos presentan un espaciado fijo ($\log F_0 + \log N$)
 - Derivar estimaciones de F_0 y ΔF_0 mediante correlación

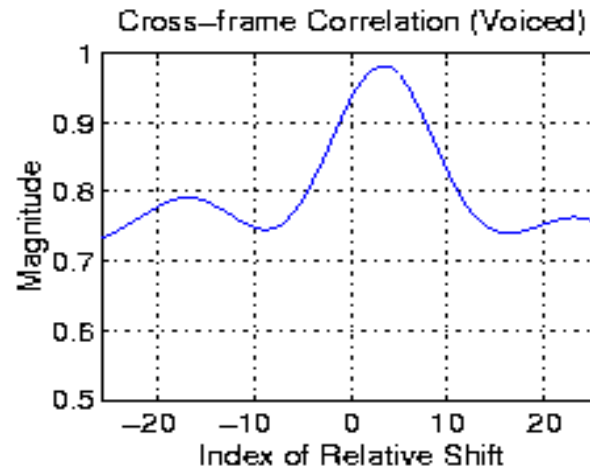
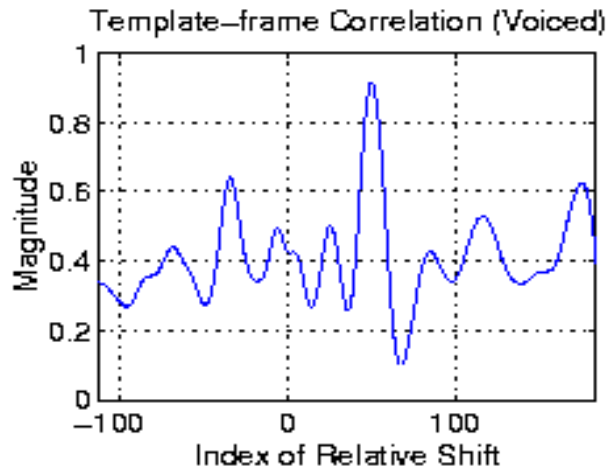


Dos funciones de correlación

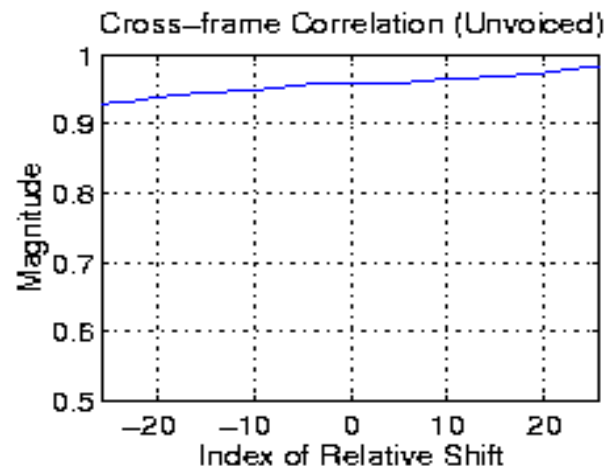
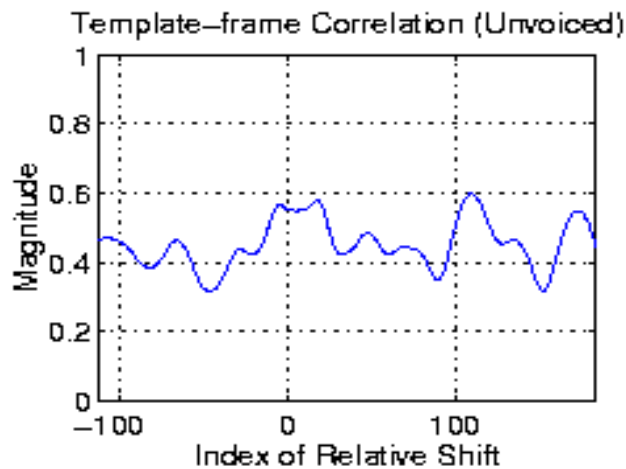
Plantilla - Tramo

Cruce - Tramo

Sonoro



Sordo

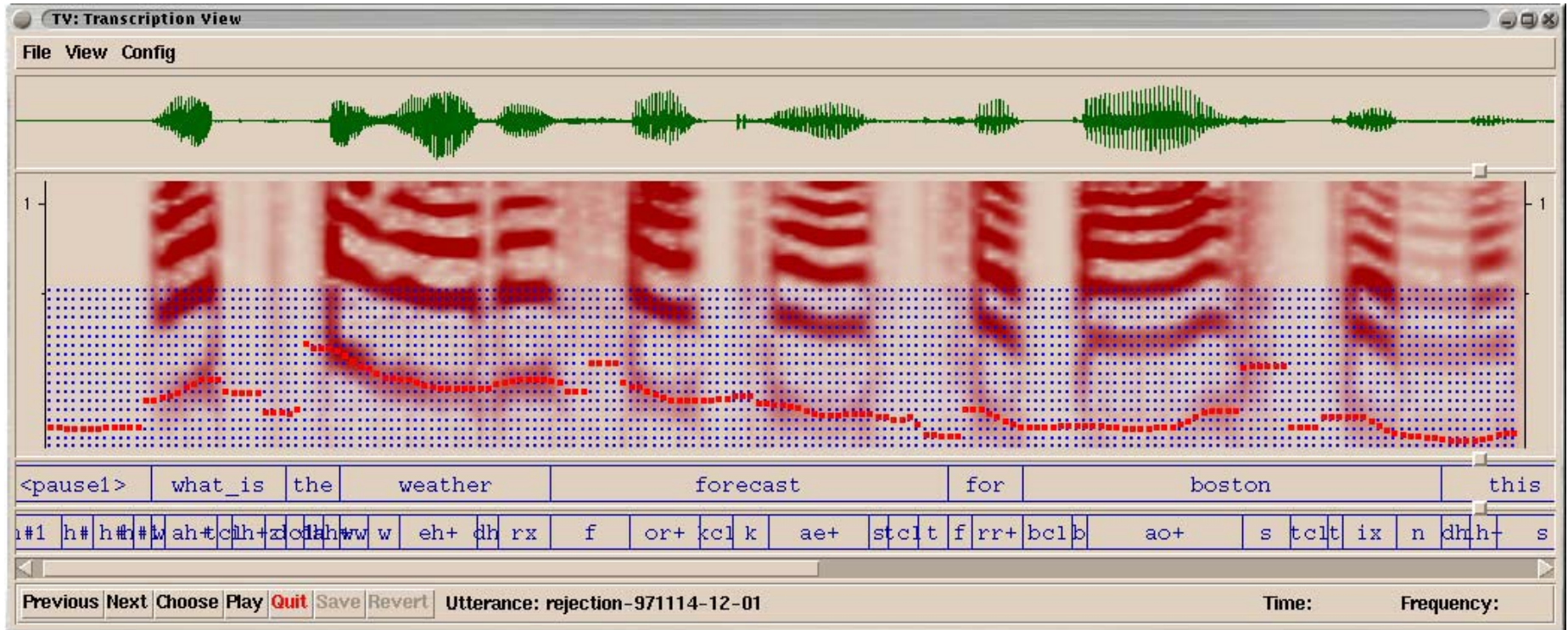


$$R_{Tx_t}(n) = \frac{\sum_i T(i)X_t(i-n)}{\sqrt{\sum_i X_t(i)^2}}$$

$$R_{X_t X_{t-1}}(n) = \frac{\sum_i X_t(i)X_{t-1}(i-n)}{\sqrt{\sum_i X_t(i)^2} \sqrt{\sum_i X_{t-1}(i)^2}}$$

Búsqueda de programación dinámica

- Solución óptima considerando restricciones de F_0 y ΔF_0
- Espacio de búsqueda cuantizado tal que $\Delta F/F$ es constante



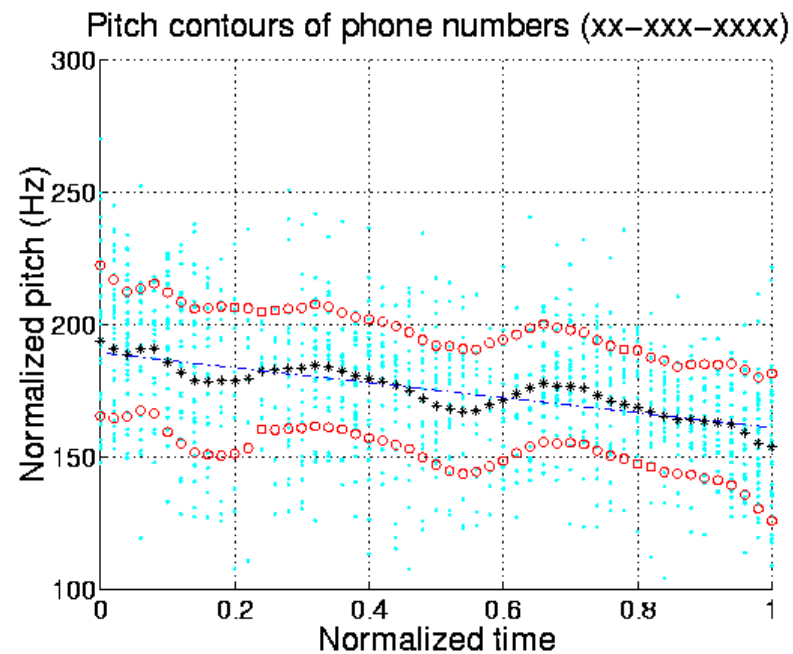
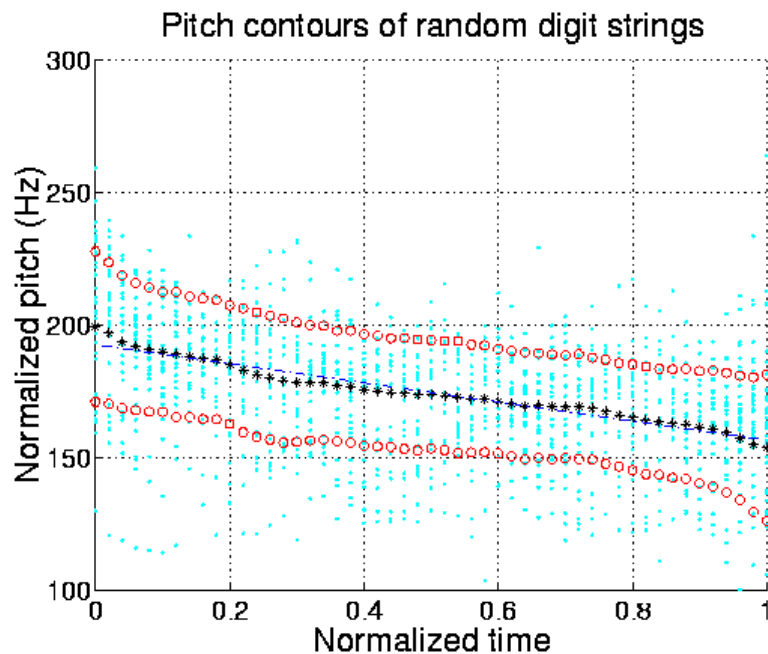
Frecuencia logarítmica

$$punto_t(i) = \begin{cases} \max_j \{score_{t-1}(j) \cdot R_{X_t X_{t-1}}(i-j)\} + R_{TX}(i-c) & (t > 0) \\ R_{TX_0}(i-c) & (t = 0) \end{cases}$$

R_{XX} : correlación cruce tramo R_{TX} : correlación plantilla tramo

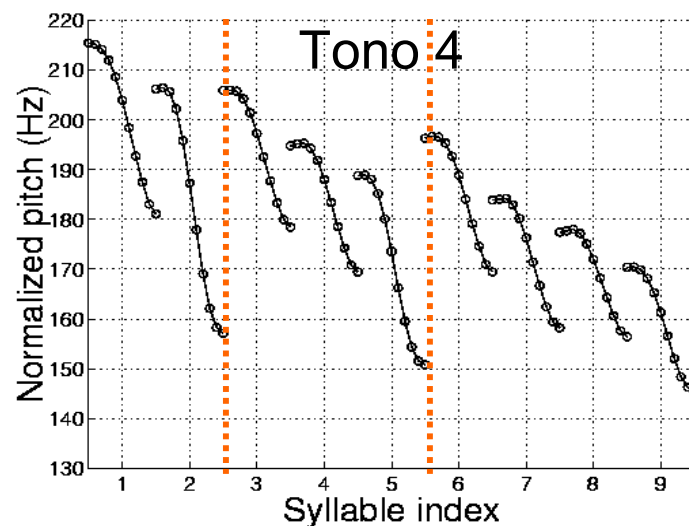
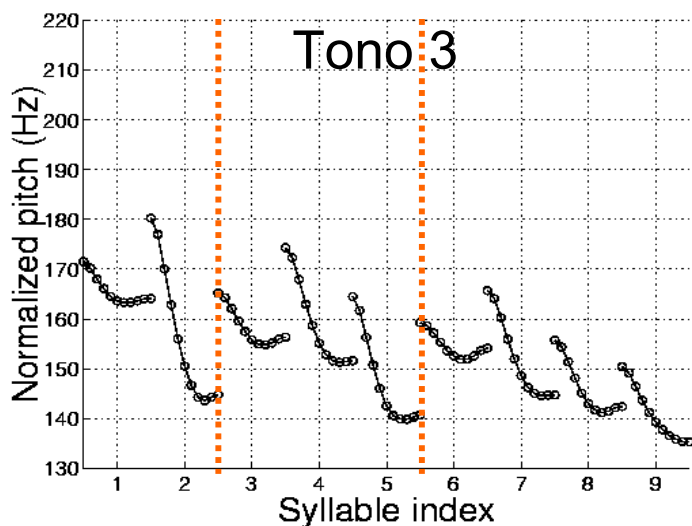
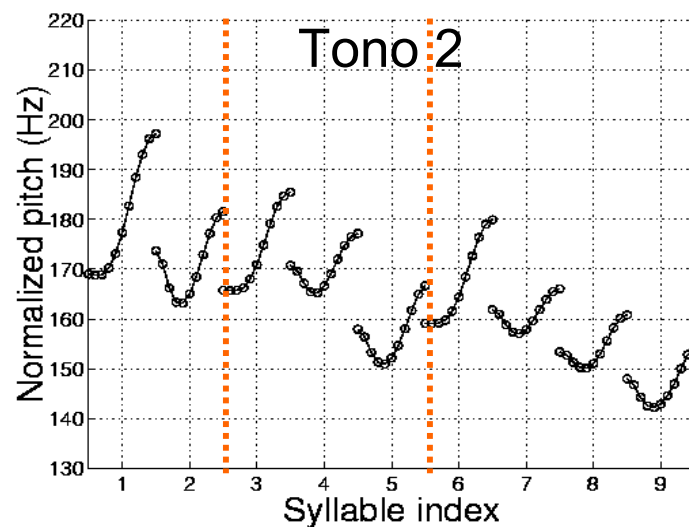
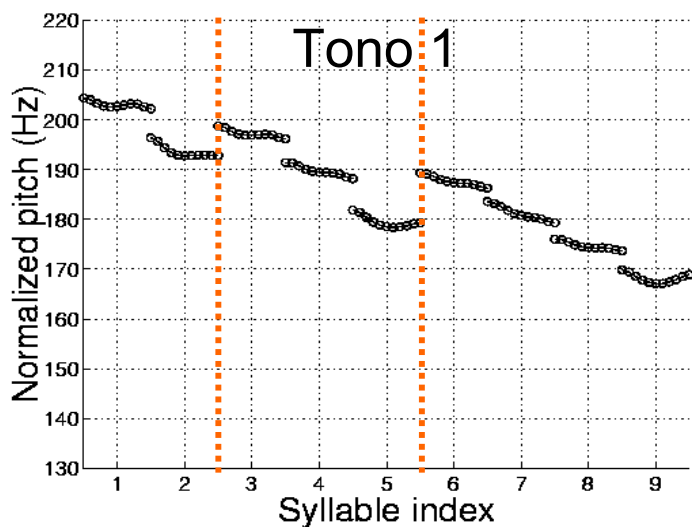
La naturaleza rítmica del habla

- Ejemplo en el que se utilizan dos tipos de cadenas de dígitos leídos en chino
 - Cadenas de dígitos aleatorias (5-10 dígitos por cadena)
 - Números de teléfono (9 dígitos, ej., 02 - 435 - 8264)
- Ambos tipos muestran una declinación tonal (ej. corriente descendente de la oración)
- Los números de teléfono muestran un patrón previsible de *ritmo*



Tonos locales frente a entonación global

- Contornos tonales dependientes de la posición en números de teléfono



MIT

Caracterización de contornos frasales

- Las frases llevan normalmente distinción de contornos F0
- Se pueden observar patrones canónicos para frases específicas
- Se ha realizado una investigación sobre la caracterización de los contornos prosódicos
 - Marcadores de límites frasales
 - Etiquetado TOBI (Tono e indicadores de límite)
- Diversas cuestiones sin respuesta
 - ¿Poseen las frases algún conjunto de patrones canónicos previsibles?
 - ¿Cómo son generalizadas las estructuras frasales prosódicas a nuevos enunciados?
 - ¿Existen interdependencias entre las frases del enunciado?
 - ¿De qué forma puede beneficiar el modelado prosódico al reconocimiento de voz y/o a la comprensión?

Estudio piloto de la prosodia frasal en JUPITER

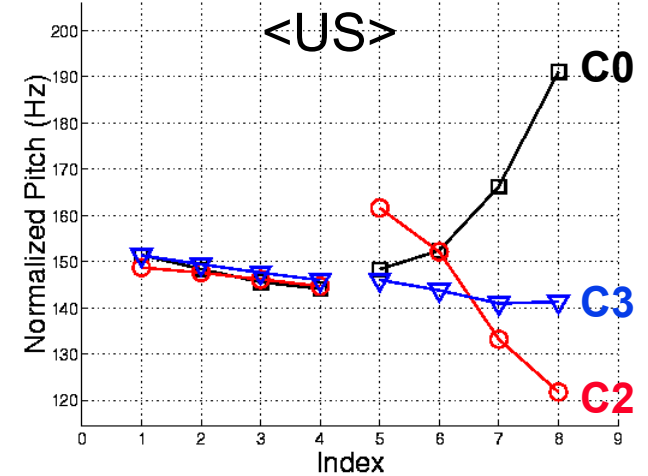
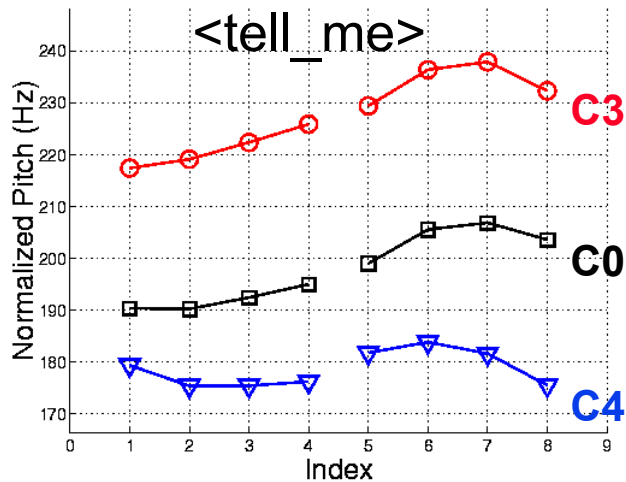
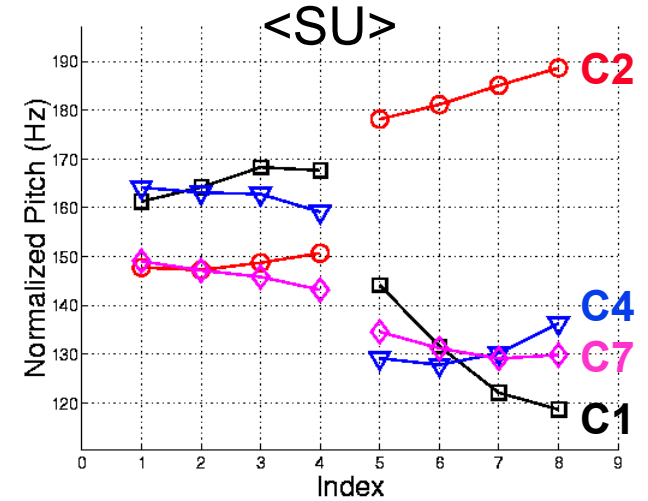
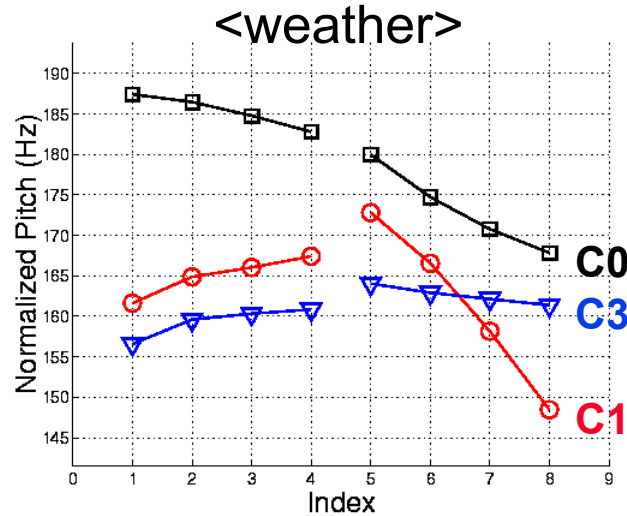
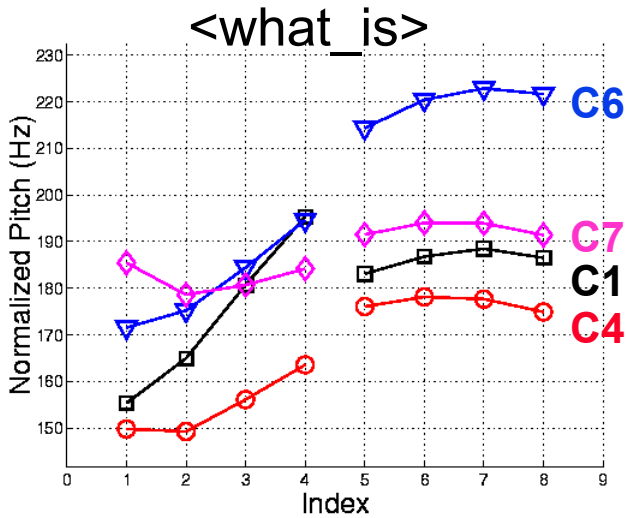
- Se estudiaron tipos de frase:
 - <what_is>: what is, how is, ...
 - <tell_me>: tell me, give me, ...
 - <weather>: weather, forecast, dew point, ...
 - <SU>: Boston, Monday, ...
 - <US>: Detroit, tonight, ...
- Frases estudiadas con una plantilla oracional fija:

<what_is> | <tell_me> the <weather> in | for | on <SU> | <US>

- Los contornos tonales para cada ejemplo fueron agrupados automáticamente en varias subclases
- La información común de las subclases puede predecir qué probabilidad aproximada existe de que éstas se den juntas en un enunciado

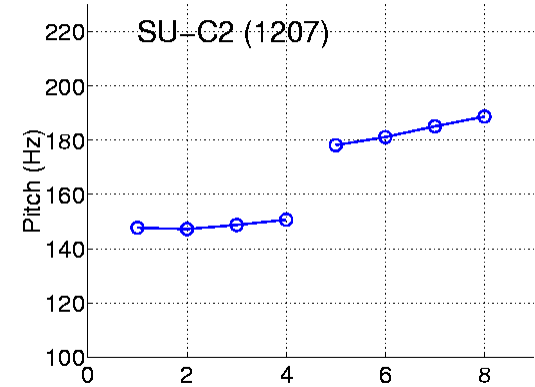
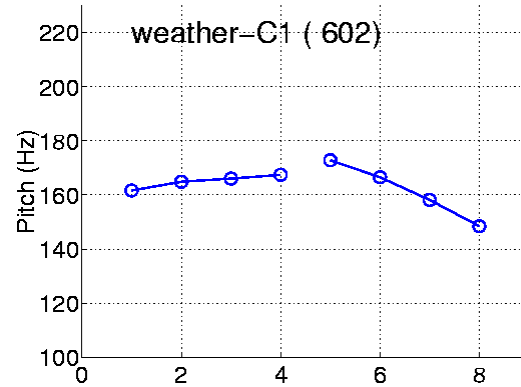
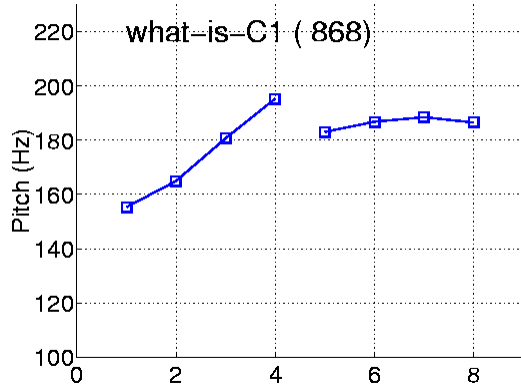
Subclases obtenidas por agrupamiento

- Agrupamiento K-medias en datos de entrenamiento seguidos por selección

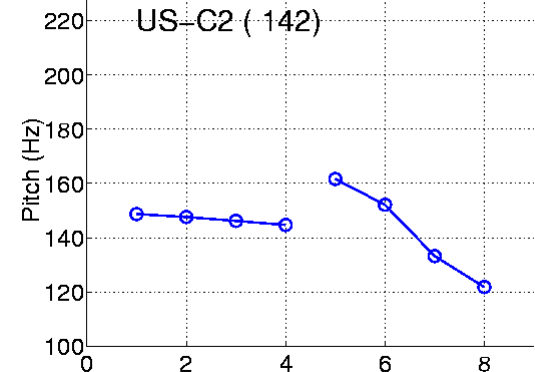
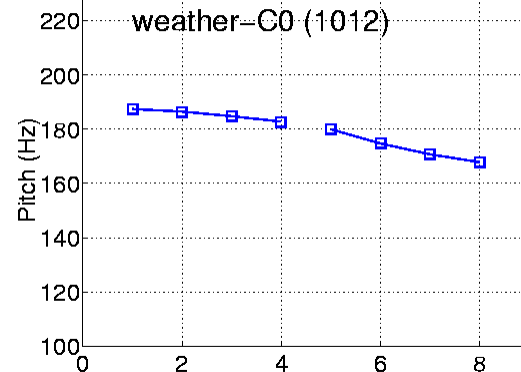
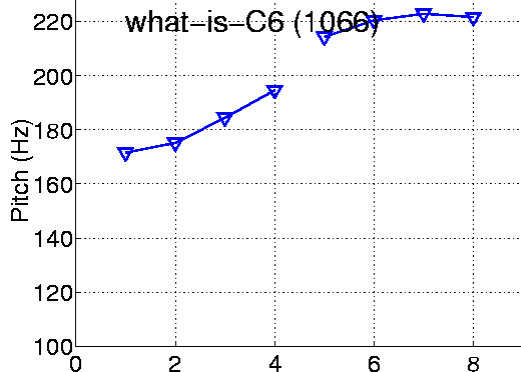


Ejemplo de enunciados

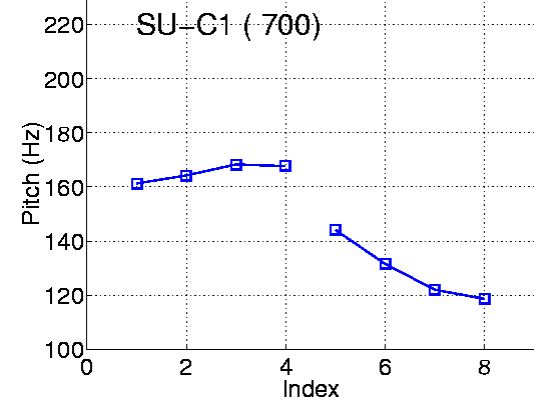
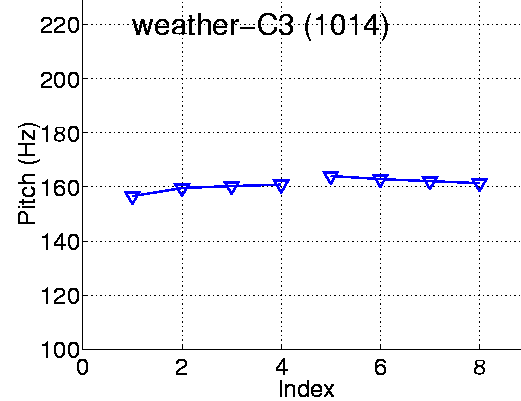
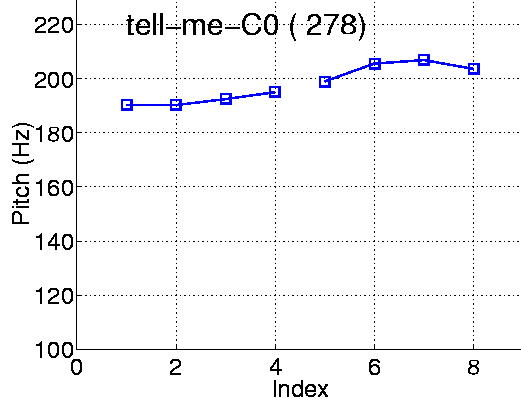
1. 



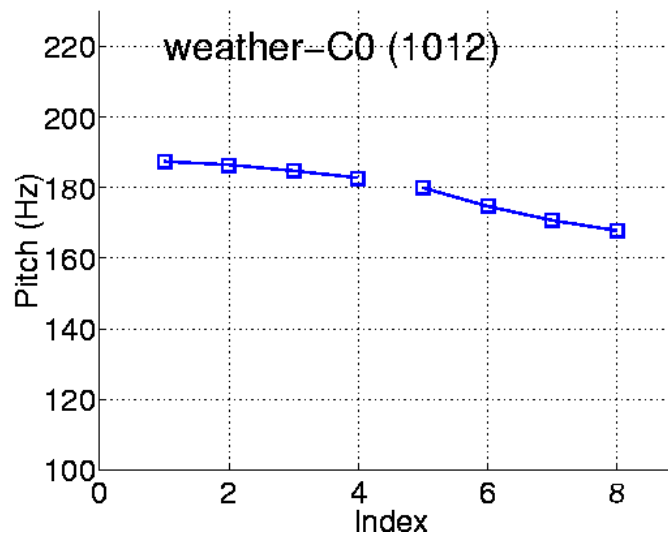
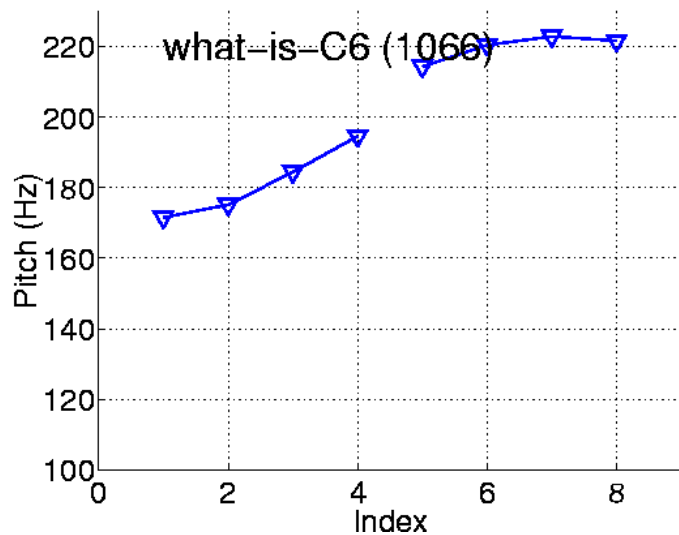
2. 



3. 

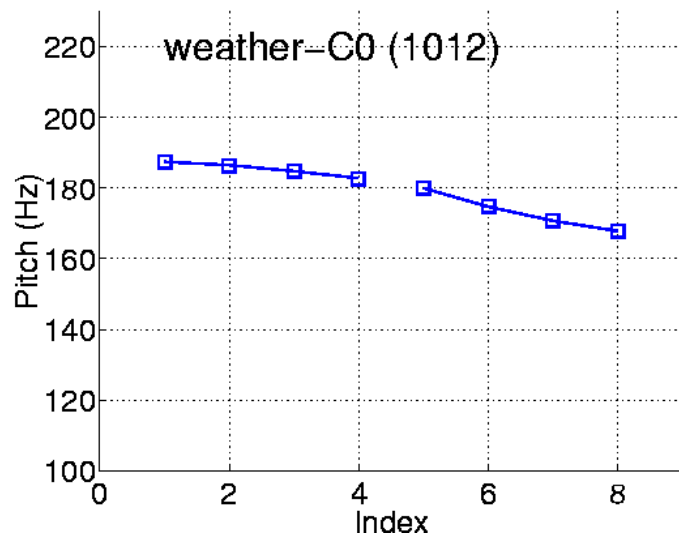
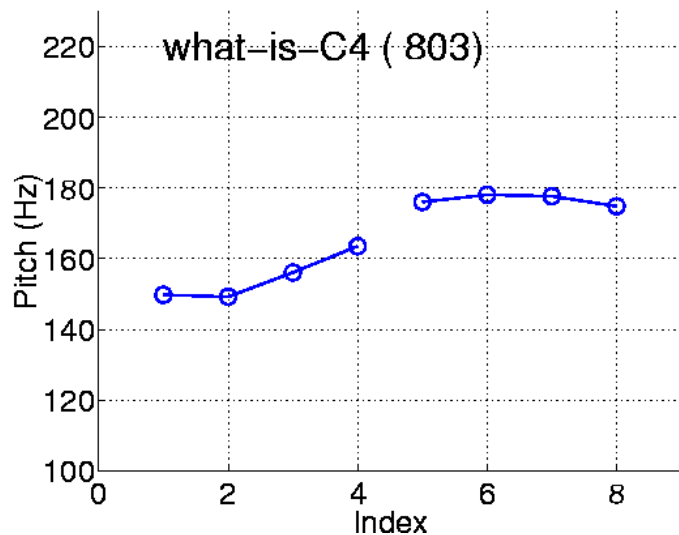


Información común de las subclases



MI = 0.67

Las subclases se utilizan normalmente juntas











MI = -0.58

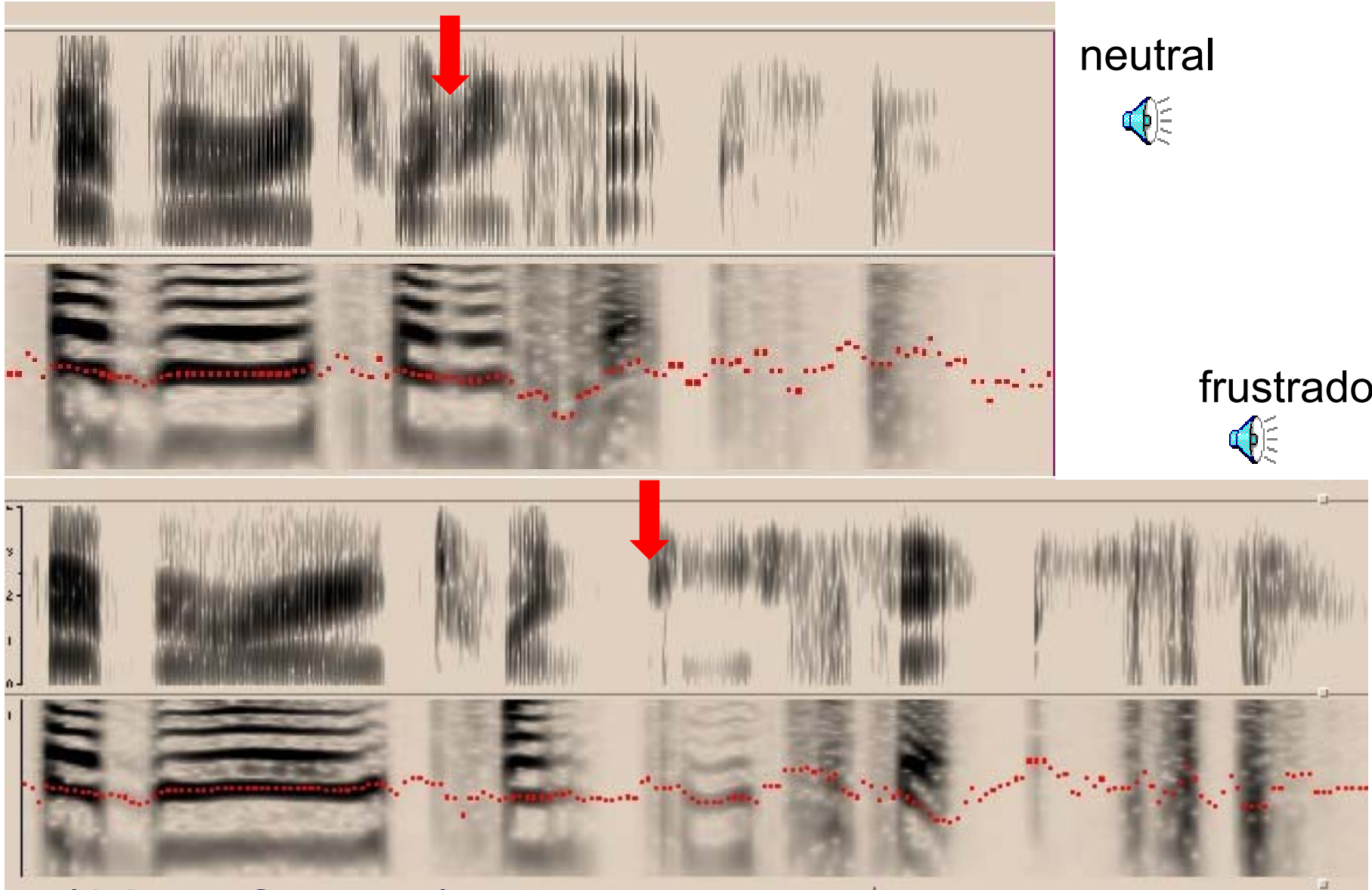
No es probable que las subclases se den juntas

MIT

Discurso emocional

- **El discurso emocional es difícil de reconocer:**
 - Tasa de error por palabra en el discurso neutral: 15%
 - WER en el discurso “alegre” en Mercury: 25%
 - WER en el discurso “frustrado”: 33%
- **Correlatos acústicos del discurso emocional/frustrado:**
 - Variación de la frecuencia fundamental 
 - Energía creciente 
 - Duración de la vocal y velocidad de habla 
 - Hiperarticulación 
 - Suspiros entre portados 
- **El contenido lingüístico puede también indicar frustración:**
 - Preguntas 
 - Constructores negativos 
 - Términos peyorativos 

Espectrogramas de un par emocional



(26 de febrero)

MIT Reconocimiento de la emoción

- Existen unos cuantos estudios sobre el reconocimiento emocional automático
- Rasgos comunes empleados en el reconocimiento de la emoción de enunciados:
 - Rasgos de F0 : medio, mediano, mínimo, máximo, desviación estándar
 - Rasgos de $\dot{F}0$: gradiente positivo medio, gradiente negativo medio, desviación estándar, ratio de gradientes ascendentes y descendentes
 - Rasgos de ritmo: velocidad de habla, duración entre regiones sonoras
- Algunos resultados:
 - 75% de exactitud sobre seis clases (feliz, triste, enfadado, indignado, sorprendido, miedo) utilizando sólo la desviación media y estándar de F0 (Huang *et al*, 1998)
 - 80% de exactitud sobre cuatro clases (feliz, triste, ira, miedo) usando 16 rasgos (Dellaert *et al*, 1998)

MIT

Identificación del hablante

- **Verificación del hablante: Aceptar o rechazar la identidad solicitada**
 - **Utilizado normalmente en aplicaciones que precisan transacciones seguras**
 - **No es 100% fiable**
 - * El discurso es muy variable y se distorsiona con facilidad
 - **Puede combinarse con otras técnicas**
 - * Posesión de una "clave" física
 - * Conocimiento de una contraseña
 - * Identificación del rostro u otras técnicas biométricas
- **Reconocimiento del hablante: Identificar al hablante desde un grupo de hablantes conocidos**
 - **Normalmente usado cuando los hablantes no ofrecen su identidad**
 - **Ejemplo de aplicaciones:**
 - * Transcripción de una entrevista e indexación
 - * Resumen de un correo de voz
 - * “Usuarios avanzados ” de un sistema de diálogo

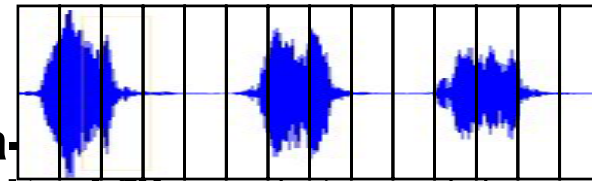
Enfoques sobre la identificación del hablante

- **Rasgos potenciales usados para la identificación del hablante**
 - Frecuencias del formante (correlacionadas con la longitud del tracto vocal)
 - Promedios y contornos de frecuencia fundamental
 - Duraciones fonéticas y velocidad de habla
 - Patrones de uso de palabras
 - Los rasgos espectrales (típicamente coeficientes MFCC) son los más utilizados normalmente
- **Algunos enfoques de modelado:**
 - **Independiente del texto**
 - * Modelos de mezclas de gaussianas globales (GMM) (Reynolds, 1995)
 - * Modelos GMM fonéticamente estructurados
 - **Dependiente del texto/ reconocimiento**
 - * Modelos GMM clasificados fonéticamente
 - * Puntuación de ASR (RAH) adaptable al hablante (Park and Hazen, 2002)

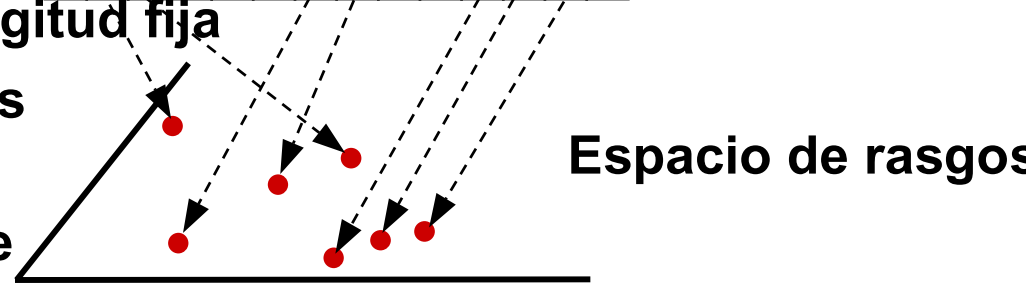
Modelos GMM globales

Entrenamiento

- Formas de onda de entrada para el hablante "i" divididas en tramos de longitud fija
- Vectores característicos computados desde cada tramo del discurso
- Modelos GMM entrenados a partir de un grupo de vectores característicos
- Un modelo GMM global por hablante

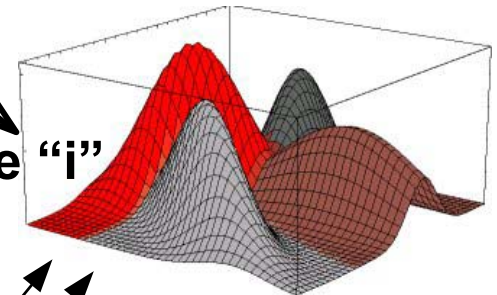


Enunciado de entrenamiento



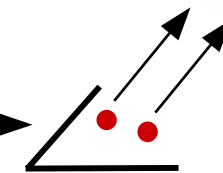
Espacio de rasgos

GMM para el hablante "i"
 $p(x_n | S_i)$



Pruebas

- Vectores característicos de entrada puntuados contra cada hablante GMM
- Puntuaciones de tramos para cada hablante sumadas sobre un enunciado completo
- La puntuación total más alta es el hablante hipotético

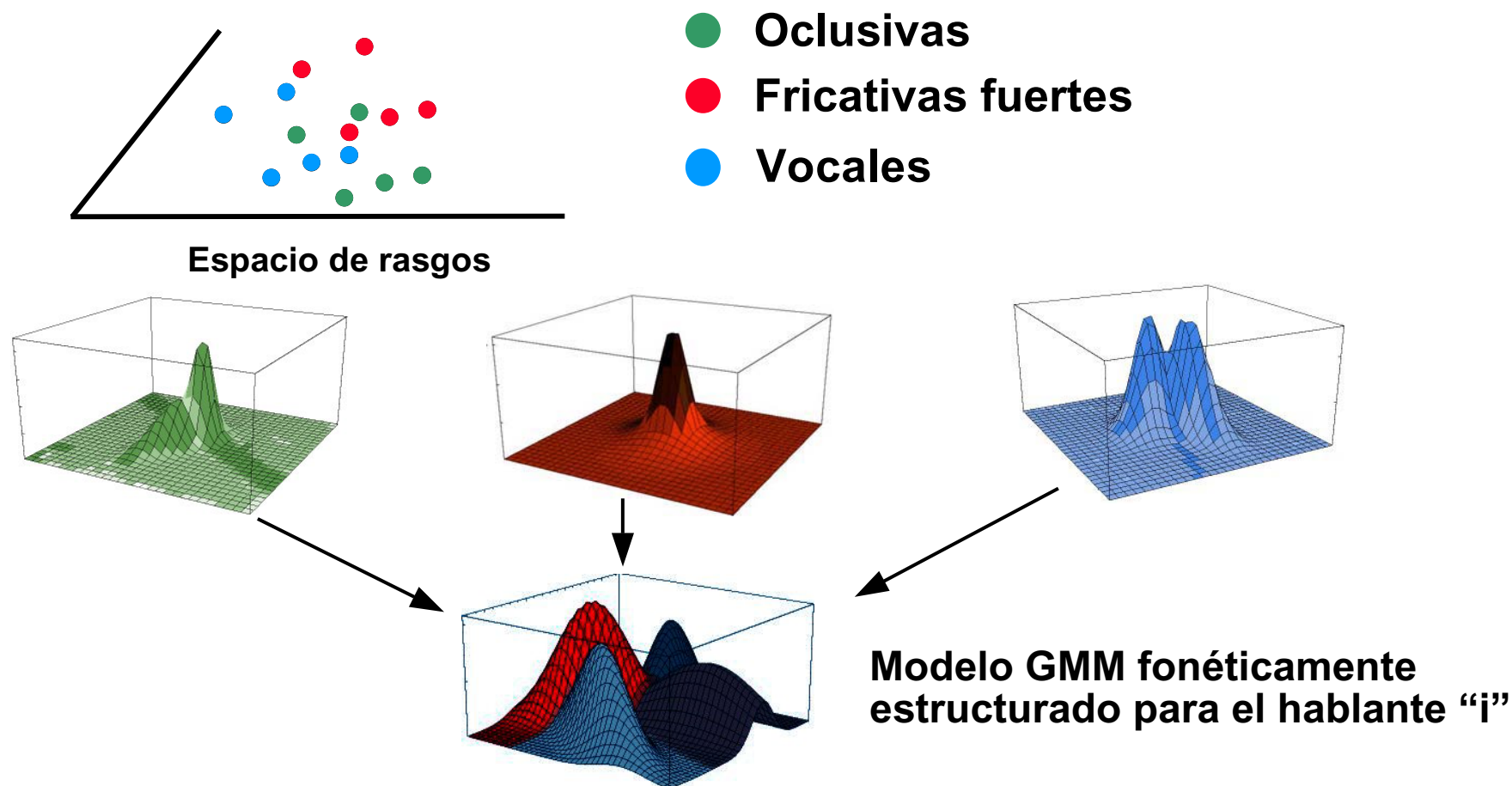


Enunciado de prueba

$$p(x_1 | S_i) + p(x_2 | S_i) = \text{puntos para hablante "i"}$$

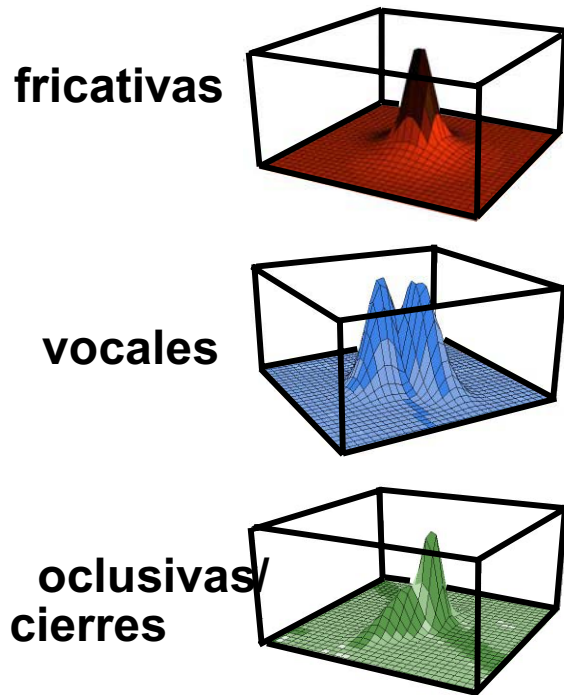
Modelo GMM fonéticamente estructurado

- Durante el entrenamiento, utilizar transcripciones fonéticas para entrenar a los modelos GMM de clase fonética para cada hablante
- Combinar modelos GMM de clase en un modelo simple "estructurado" que se usa después para puntuar como en el sistema de punto de referencia

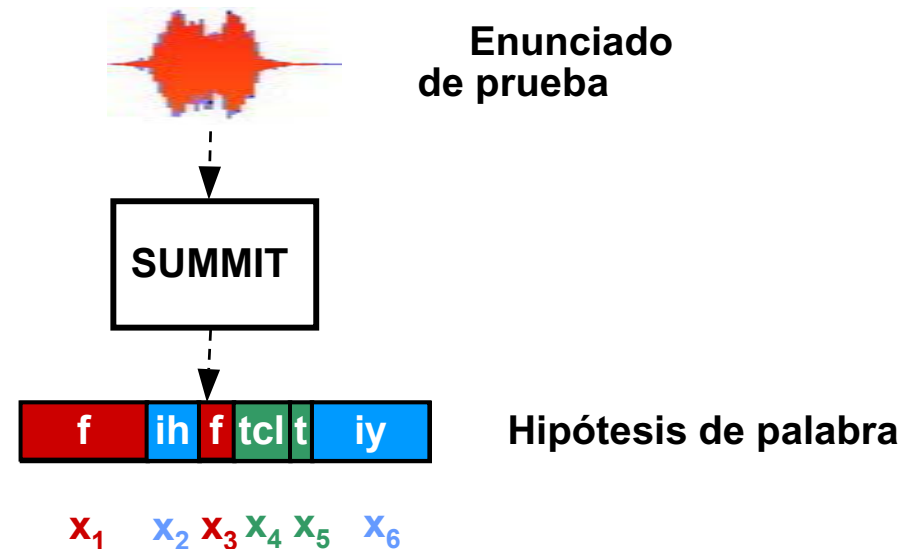


Clasificación fonética

- Entrenar modelos GMM de clases de fono independientes sin combinación
- Generar hipótesis de fonos/palabras desde el reconocedor
- Marcos de puntuación con modelos de clase de fonos hipotéticos



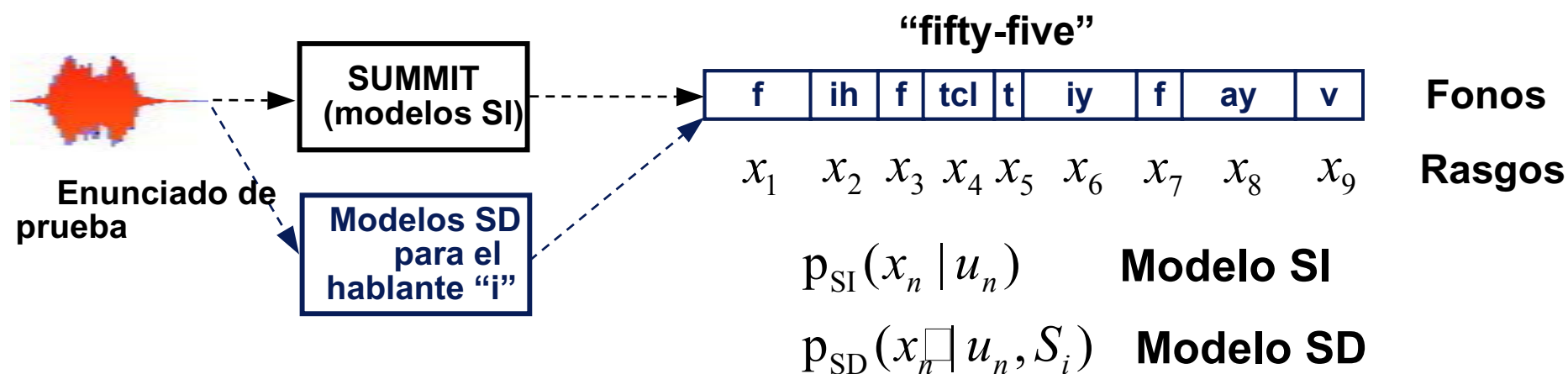
Modelos GMM de clase fonética para hablante “i”



$$\sum p(x_n | S_i, \text{clase}(x_n)) = \text{puntuación para hablante "i"}$$

Puntuación adaptada al hablante

- Entrenar modelos dependientes del hablante (SD) para cada hablante
- Obtener la mejor hipótesis del reconocedor usando modelos independientes del hablante (SI)
- Repuntuar la hipótesis con modelos SD
- Calcular la puntuación total adaptada al hablante interpolando puntuación SD con puntuación SI



$$p_{SA}(x_n | u_n, S_i) = \frac{\lambda_n p_{SD}(x_n | u_n, S_i) + (1 - \lambda_n) p_{SI}(x_n | u_n, S_i)}{p_{SI}(x_n | u_n, S_i)} \quad \lambda_n = \frac{c(u_n)}{c(u_n) + K}$$

Dos corpus experimentales

Corpus	YOHO	Mercury
Descripción	Corpus LDC para evaluación de verificación del hablante	Corpus SLS de un sistema de vuelos aéreos
Tipo de discurso	Texto apuntado Frases de "combinación cerrada" (ej. "34-25-86")	Discurso conversacional espontáneo en el dominio de vuelos aéreos
Nº hablantes	138 (106M, 32F)	38 (18M, 20F)
Condiciones de grabación	Auricular fijo Entorno tranquilo de oficina 8kHz de banda limitada	Teléfono variable Entorno variable Canal del teléfono
Datos de entrenamiento	96 enunciados Desde 4 sesiones (menos de 3 segundos cada una)	50-100 enunciados De 2-10 sesiones (longitud variable)
Tamaño conj. pruebas	5520	3219

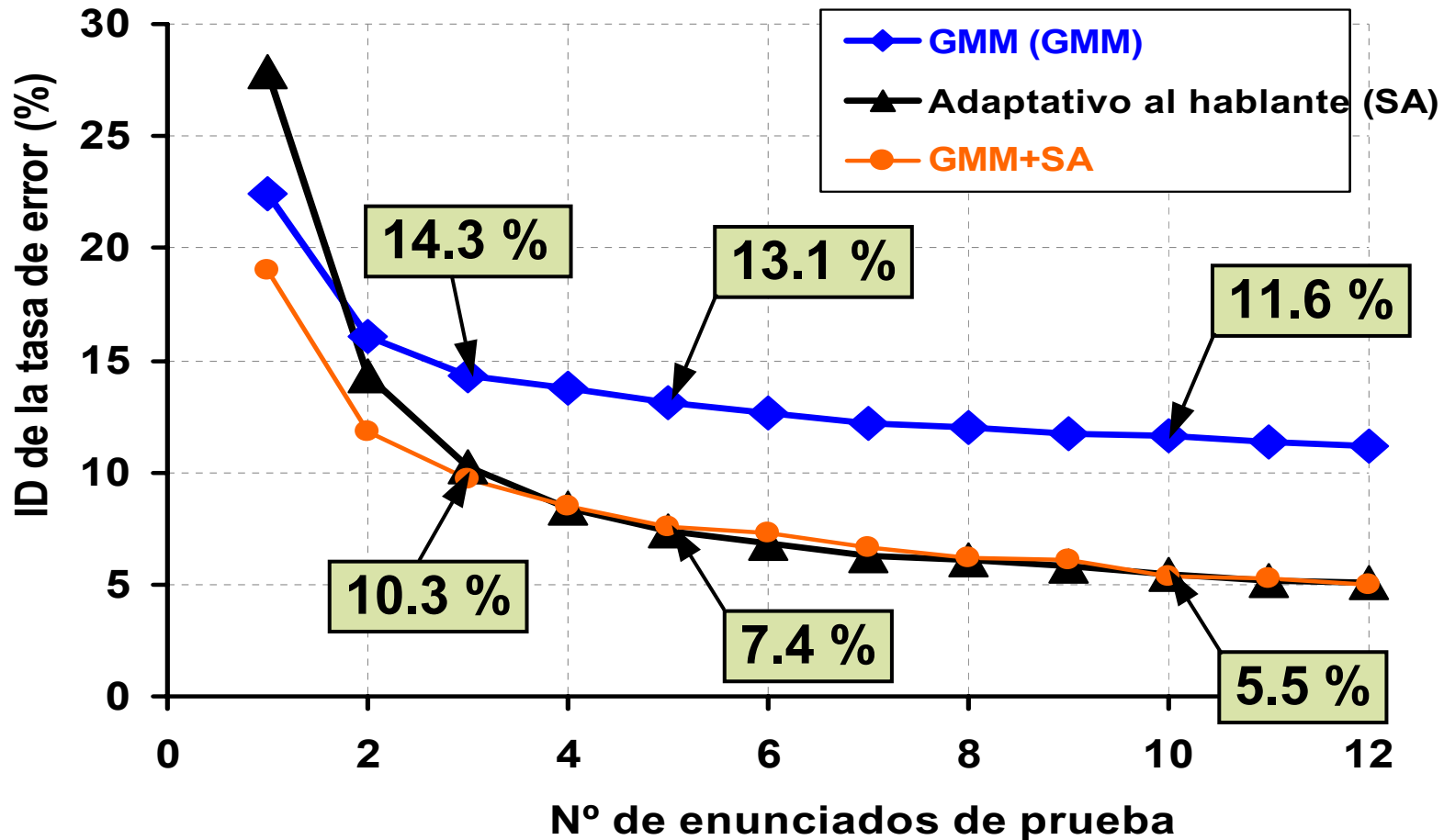
Resultados de un enunciado sencillo

- Experimento: reconocimiento del hablante en grupo cerrado con enunciados sencillos
- Resultados:

Sistema	Tasa error por ID de hablante %	
	YOHO	Mercury
GMM estructurado(SGMM)	0.31	21.3
Clasificación de fonos	0.40	21.6
Adaptativo al hablante (SA)	0.31	27.8
SA+SGMM	0.25	18.3

- Todos los enfoques casi iguales con el corpus YOHO
- El enfoque adaptativo al hablante tiene el rendimiento más pobre en Mercury
 - Los errores de reconocimiento de ASR (RAH) pueden degradar el rendimiento de la ID del hablante
- La combinación del clasificador produce mejoras sobre el mejor sistema

Resultados en enunciados múltiples de Mercury



- En enunciados múltiples, la puntuación adaptable al hablante consigue tasas de error más bajas que el método individual del próximo mejor
- Reducciones del 28%, 39 % y 53% de la tasa de error relativa en los enunciados 3, 5 y 10, comparado con el punto de referencia

MIT

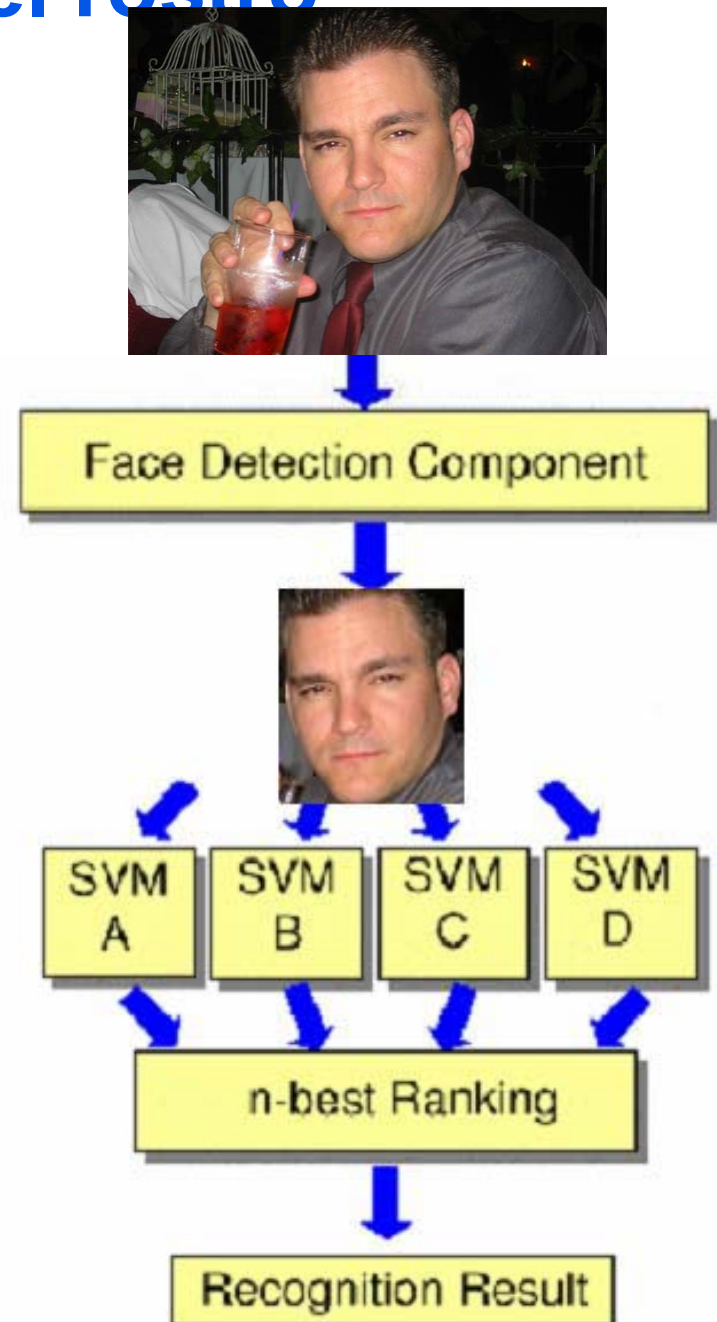
Interfaces multimodales

- **Las interfaces multimodales permitirán una interacción entre el humano y la máquina más natural, flexible, eficiente y robusta**
 - **Natural:** No precisa entrenamiento especial
 - **Flexible:** Los usuarios seleccionan las modalidades preferidas
 - **Eficiente:** El lenguaje y los gestos pueden ser más simples que en las interfaces unimodales (ej., Oviatt y Cohen, 2000)
 - **Robusto:** Las entradas son complementarias y sistemáticas
- **Las señales de audio y video contienen información sobre:**
 - **Identidad de la persona:** ¿Quién habla?
 - **Mensaje lingüístico:** ¿Qué es lo que dicen?
 - **Emoción, humor, estrés, etc.:** ¿Cómo se sienten?
- **La integración de estas señales puede conducirnos a un aumento de las capacidades de futuras interfaces de humanos con ordenadores**

- **Se ha desarrollado un iPaq de mano con entrada/salida de audio/video como parte del proyecto Oxygen de MIT**
- **La presencia de canales multientrada permite esquemas de verificación multimodal**
- **El sistema prototipo emplea un escenario de registro**
 - **Imagen frontal del rostro en una fotografía instantánea**
 - **Nombre del estado**
 - **Recitar la frase de la combinación cerrada apuntada**
 - **El sistema *acepta o rechaza* al usuario**

Enfoque de la identificación del rostro

- **Detección del rostro Compaq/HP (Viola/Jones, CVPR 2001)**
 - Cascada eficaz de clasificadores
- **Reconocimiento del rostro por MIT AI Lab/CBCL (Heisele et al, ICCV 2001)**
 - Basado en máquinas de vectores soporte (SVM)
 - Tiempo de ejecución del reconocimiento del rostro: puntuar imagen contra cada clasificador SVM
- **Implementado en un iPaq de mano como parte del proyecto Oxygen de MIT (E. Weinstein, K. Steele, P. Ho, D. Dopson)**



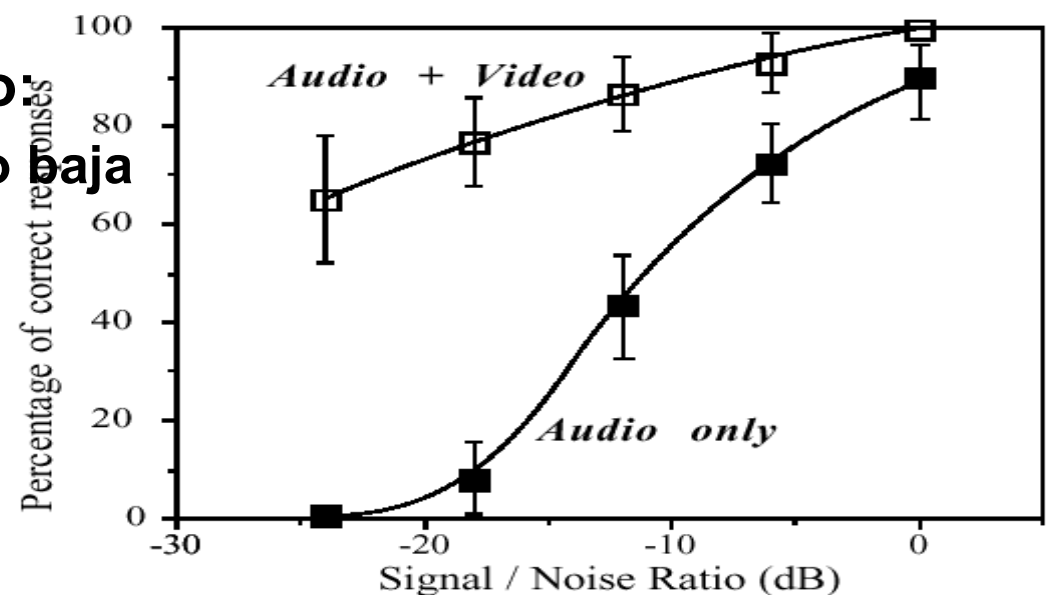
Combinación del rostro y la ID del hablante

- Experimento de verificación del registro de un usuario multimodal mediante iPaq
- Datos de inscripción:
 - Entrenamiento de datos recopilados de 35 usuarios inscritos
 - 100 imágenes faciales y 64 frases de combinación cerrada por usuario
- Datos de prueba:
 - 16 parejas rostro/imagen de 25 usuarios inscritos
 - 10 parejas rostro/imagen de 20 impostores no inscritos
- Métrica de evaluación: Tasa de error igual de verificación (EER)
 - Misma probabilidad de aceptaciones falsas y rechazos falsos
 - El sistema fusionado reduce la tasa de error igual en un 50%

Sistema	Tasa de error igual
Sólo ID del rostro	7.30%
Sólo ID del discurso	1.77%
Sistema fusionado	0.89%

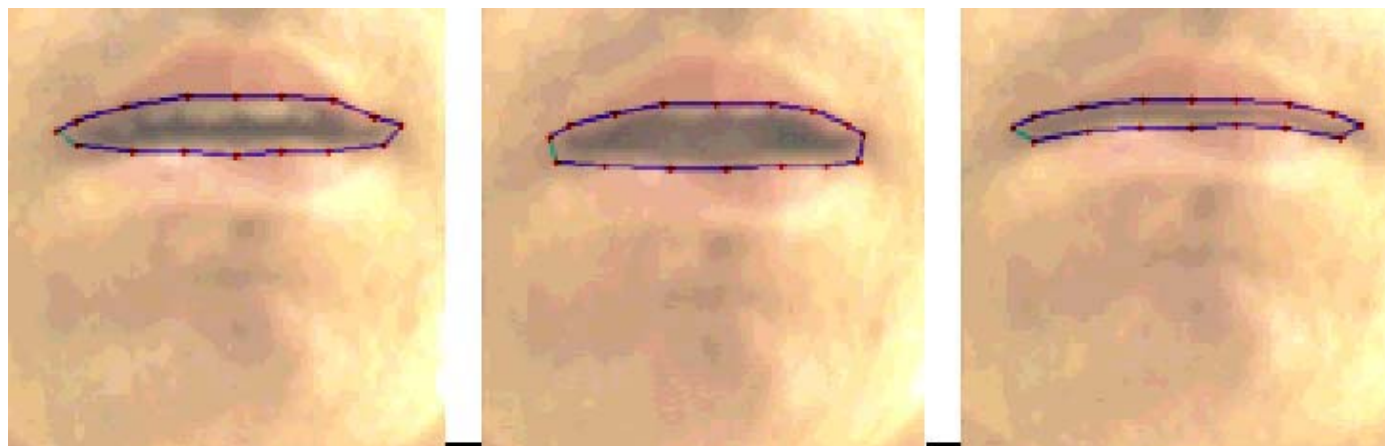
¿Cómo mejorar el rendimiento del ASR (RAH)?

- Los humanos emplean expresiones faciales y gestos para aumentar la señal de voz
- Las señales faciales pueden optimizar el reconocimiento de voz con ruido hasta aproximadamente 30 dB, dependiendo de la tarea
- El rendimiento del reconocimiento de voz puede mejorarse al incorporar señales faciales (ej., movimientos de labios y apertura de la boca)
- La figura muestra el func. del reconocimiento humano
 - Proporción de señal a ruido **baja**
 - Presentado con audio más video y con audio sólo
 - Referencia: Benoit, 1992



Reconocimiento de voz audiovisual (AVSR)

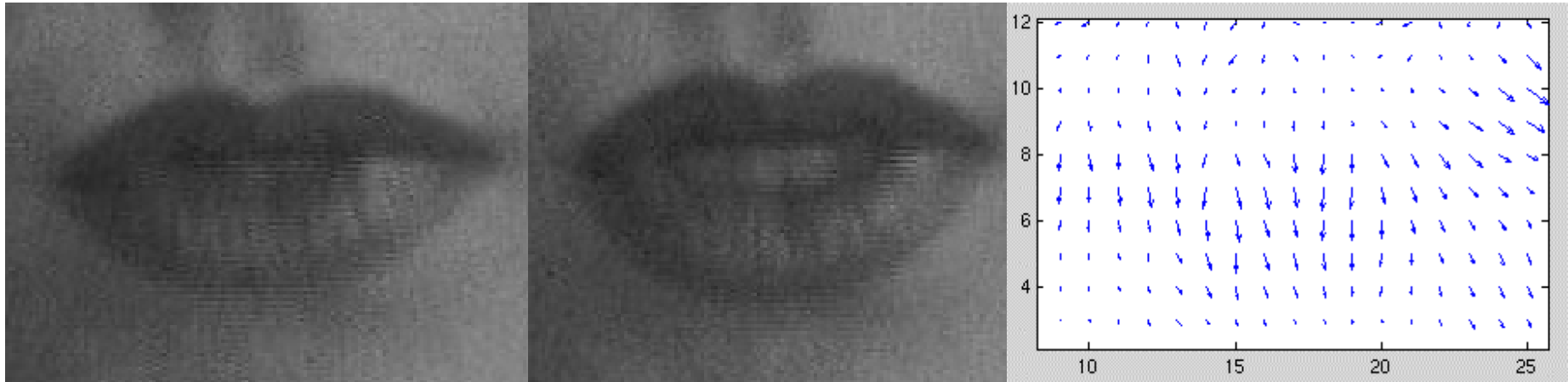
- Integrar información sobre rasgos visuales de la boca/labios/mandíbula, con rasgos extraídos de una señal de audio
- Extracción de rasgos visuales:
 - Región de interés (ROI): sobre todo labios y boca; algún registro
 - Rasgos: basados en píxel, en la forma o geométricos
 - Casi todos los sistemas deben localizar y seguir puntos de referencia
 - No se utiliza explícitamente la correlación y la información de movimiento



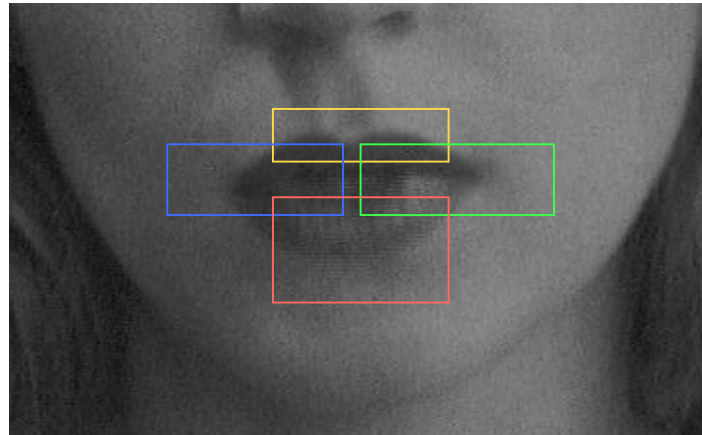
Ejemplo de rasgos basados en píxel (Covell & Darrell, 1999)

AVSR: Investigaciones preliminares

- **Objetivo: integración con el sistema de ASR SUMMIT**
- **Mediciones derivadas visualmente basadas en flujo óptico**



- **Los rasgos de dimensión baja representan apertura y alargamiento**



MIT

Cuestiones con integración audio/video

- **Integración temprana frente a tardía**
 - Temprana: concatenar vectores característicos desde modos distintos
 - Tardía: combinar salidas de clasificadores unimodales
 - * Puede estar en muchos niveles (fono, sílaba, palabra, enunciado)
- **Esquemas de ponderación del canal**
 - El canal de audio proporciona normalmente más información
 - Basado en una estimación SNR (proporción señal a ruido) para cada canal
 - Ponderaciones preestablecidas optimizando la tasa de error de un grupo de dispositivos
 - Estimar ponderaciones separadas para cada fonema o visema
- **Modelando la asincronía audio/visual**
 - Muchas señales visuales se dan antes de que el fonema se pronuncie realmente
 - Ejemplo: redondear los labios antes de producir el fonema de redondeado

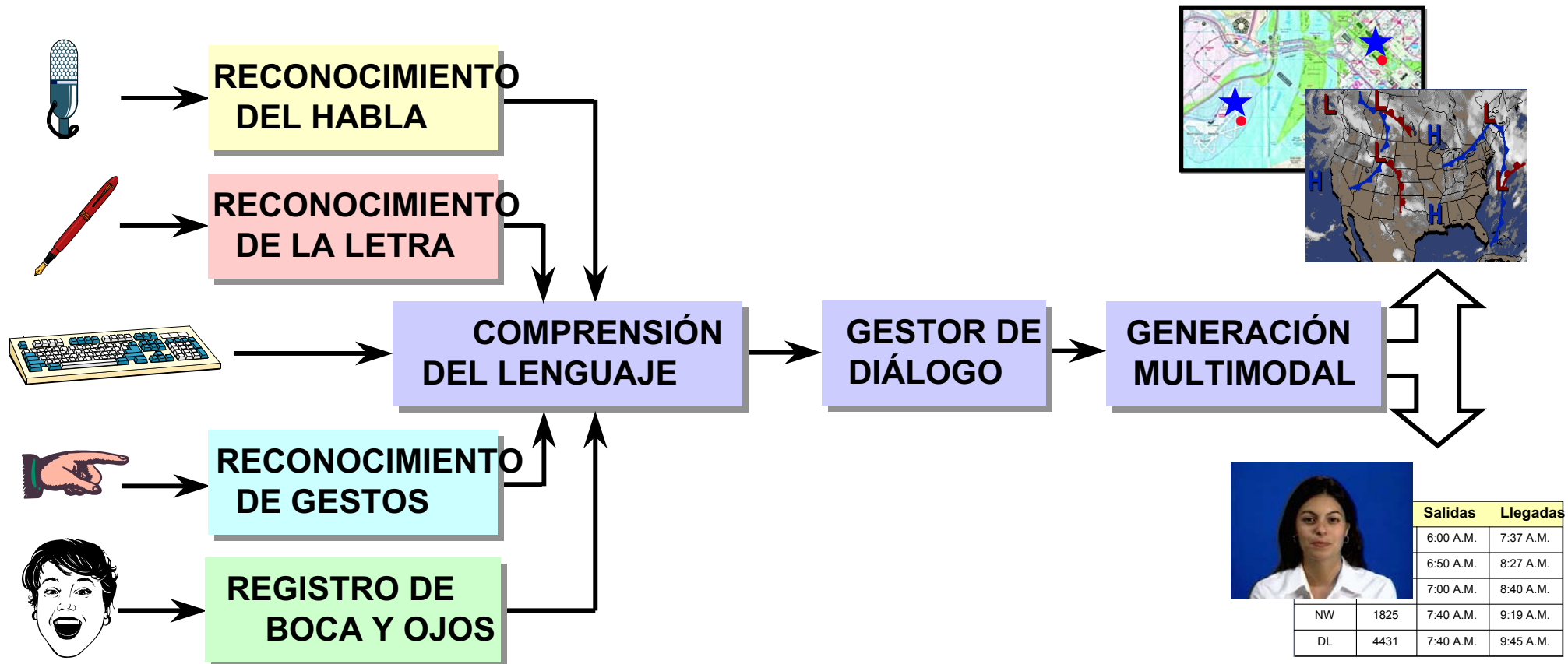
MIT AVSR: Estado del arte

- **Ejemplo: Neti *et al*, 2000 (Seminario de verano JHU)**
 - Vocabulario de palabras mayor a 10K
 - Entrenamiento y datos de desarrollo: 264 sujetos, 40 horas
 - Datos de prueba: 26 sujetos, 2.5 horas
 - Condiciones tranquilas (19.5 dB de SNR) y ruidosas (8.5 dB SNR)

Condiciones	Limpio WER (%)	Ruidoso WER (%)
Sólo audio	14.4	48.1
AVSR	13.5	35.3

Investigación de interacción multimodal

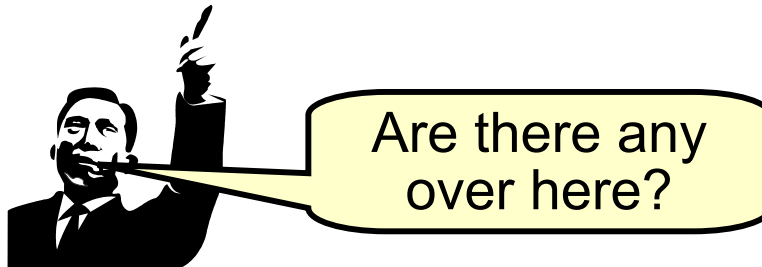
- **Comprensión de la ciencia**
 - ¿Cómo lo hacen los humanos (ej. expresando contexto de modalidad cruzada)?
 - ¿Cuáles son las señales importantes?
- **Desarrollo de una arquitectura que pueda describir adecuadamente las interacciones de modalidades**



MIT

Interfaces multimodales

- Es necesario que las entradas sean comprendidas en el contexto adecuado

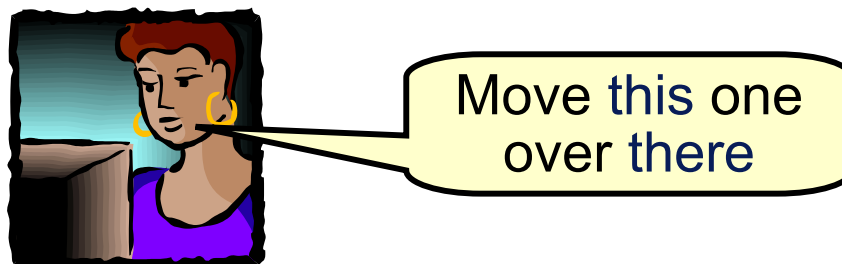


¿Qué quiere decir con “*any*” (alguno), y a qué está señalando?



¿Significa esto “*yes*” (sí), “*one*” (uno) u otra cosa?

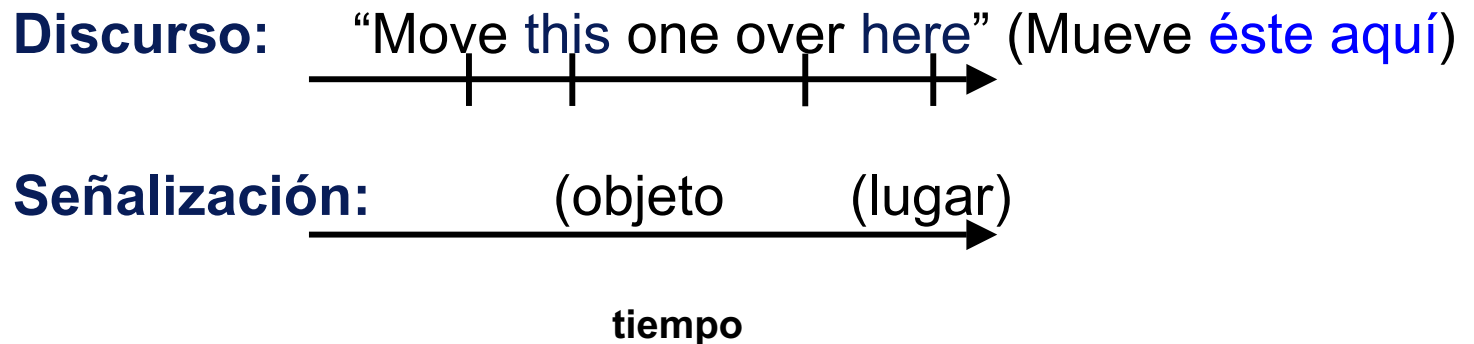
- La información de tiempo es un modo útil de relacionar entradas



¿A dónde está mirando o señalando mientras pronuncia “*this*” (esto) y “*there*” (aquí)?

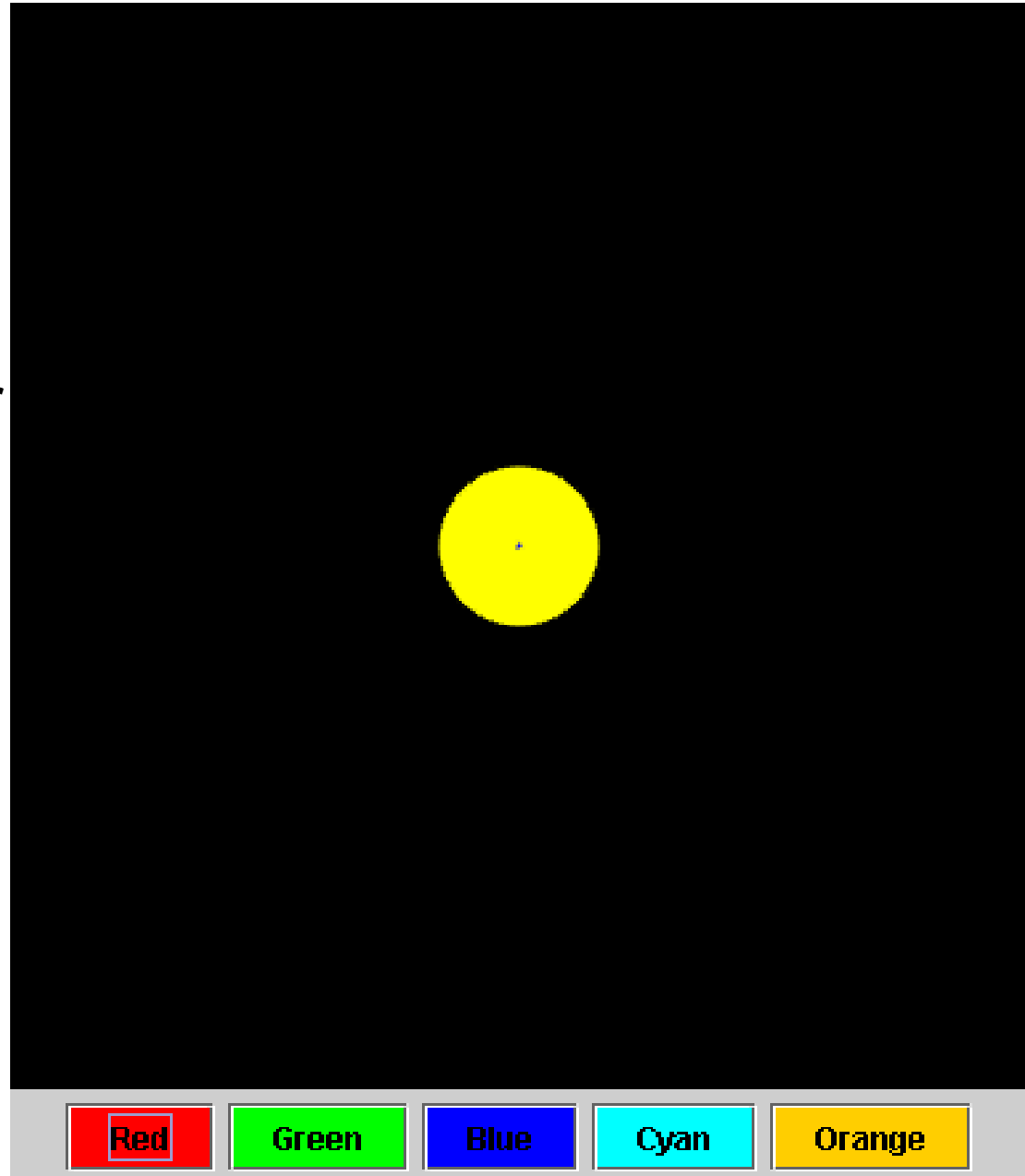
Fusión multimodal: Progreso inicial

- **Todas las entradas multimodales están sincronizadas**
 - El reconocedor de voz genera tiempos absolutos o palabras
 - Los movimientos de gestos y el ratón producen triples $\{x,y,t\}$
- **La comprensión del habla restringe la interpretación de los gestos**
 - El trabajo inicial identifica un objeto o lugar a partir de entradas de gestos
 - El discurso limita qué, cuándo y cómo se resuelven los elementos
 - La resolución del objeto también depende de la información de la aplicación

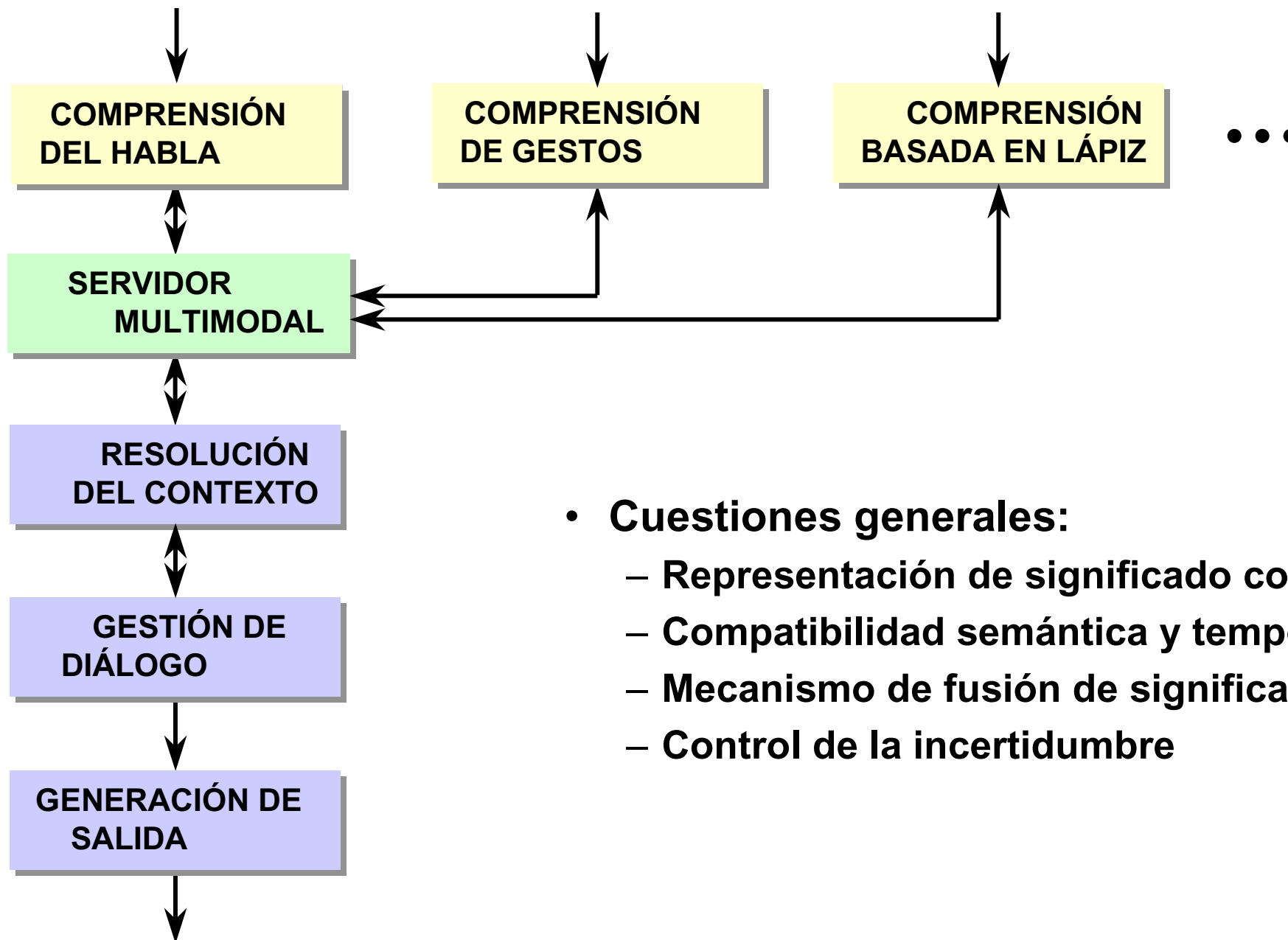


Demostración multimodal

- Manipulación de planetas en una aplicación del sistema solar
- Seguimiento continuo del ratón o de gestos de señalización
- Creado con la utilidad SpeechBuilder con pequeños cambios (Cyphers, Glass, Toledano & Wang)
- La versión autónoma se ejecuta con la entrada del ratón/lápiz
- Puede combinarse con gestos de **from determined from vision (Darrell & Demirdjien)**



Actividades recientes : Servidor multimodal



- **Cuestiones generales:**

- Representación de significado común
- Compatibilidad semántica y temporal
- Mecanismo de fusión de significado
- Control de la incertidumbre

- **El discurso conlleva contenido paralingüístico:**
 - **Prosodia, entonación, acento, énfasis, etc.**
 - **Emoción, humor, actitud, etc.**
 - **Características específicas del hablante**
- **Las interfaces multimodales pueden ser mejores que los sistemas destinados a voz**
 - **Mejor identificación del individuo mediante rasgos faciales**
 - **Mejor reconocimiento de voz por la lectura de labios**
 - **Interacción entre hombre y máquina natural, flexible, eficiente y robusta**

Referencias

- C. Benoit, “The intrinsic bimodality of speech communication and the synthesis of talking faces,” *Journal on Communications*, September 1992.
- M. Covell y T. Darrell, “Dynamic occluding contours: A new external-energy term for snakes,” *CVPR*, 1999.
- F. Dellaert, T. Polzin, and A. Waibel, “Recognizing emotion in speech,” *ICSLP*, 1998.
- B. Heisele, P. Ho, and T. Poggio, “Face recognition with support vector machines: Global versus component-based approach,” *ICCV*, 2001.
- T. Huang, L. Chen y H. Tao, “Bimodal emotion recognition by man and machine,” *ATR Workshop on Virtual Communication Environments*, April 1998.
- C. Neti, *et al*, “Audio-visual speech recognition,” *Tech Report CLSP/Johns Hopkins University*, 2000.
- S. Oviatt y P. Cohen, “Multimodal interfaces that process what comes naturally,” *Comm. of the ACM*, March 2000.
- A. Park y T. Hazen, “ASR dependent techniques for speaker identification,” *ICSLP*, 2002.
- D. Reynolds, “Speaker identification and verification using Gaussian mixture speaker models,” *Speech Communication*, August 1995.
- P. Viola y M. Jones, “Rapid object detection using a boosted cascade of simple features,” *CVPR*, 2001.
- C. Wang, “Prosodic modeling for improved speech recognition and understanding,” PhD thesis, MIT, 2001.
- E. Weinstein, *et al*, “Handheld face identification technology in a pervasive computing environment,” *Pervasive 2002*.