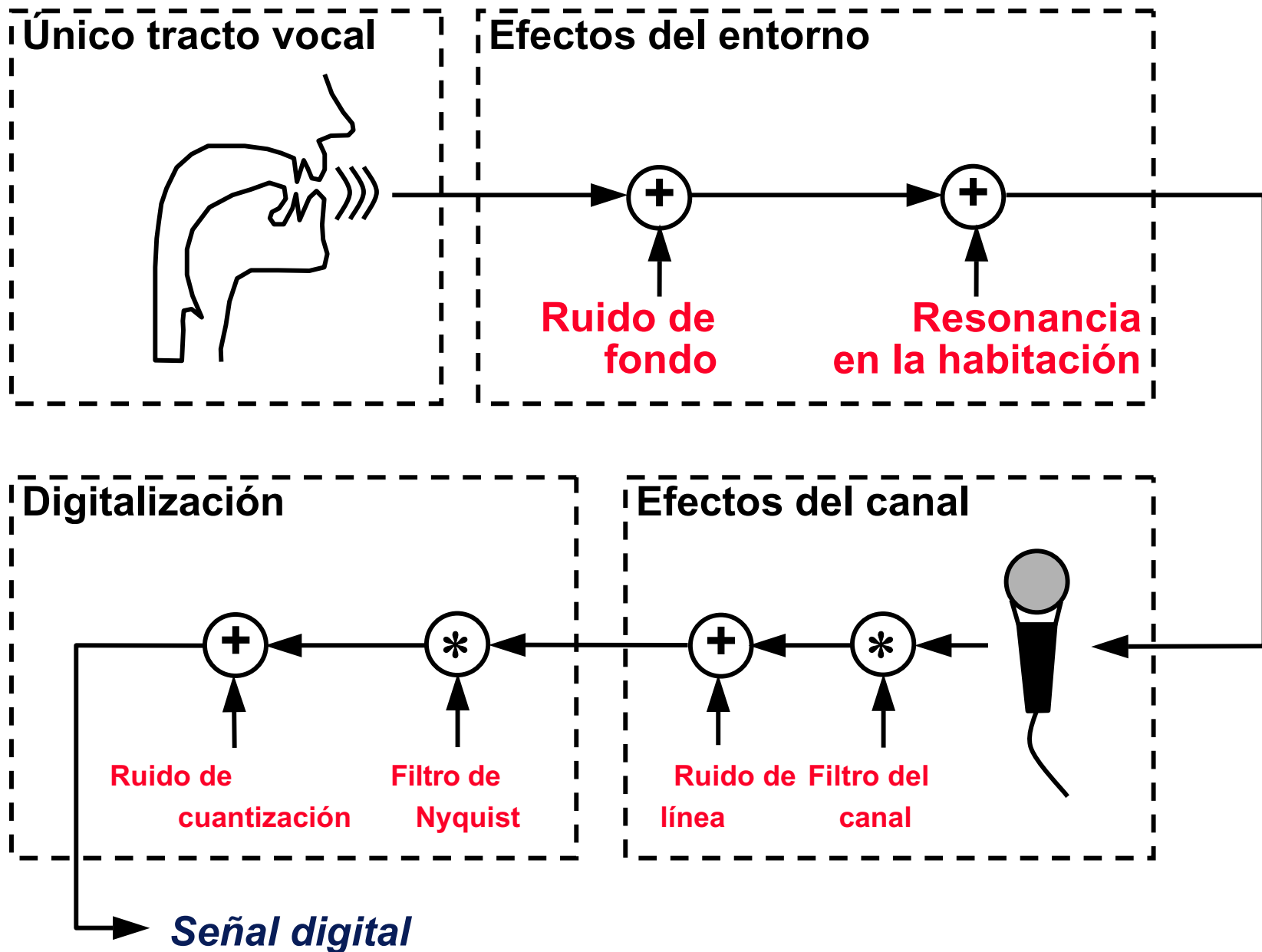


# Robustez de ruido y puntuación de seguridad

**Profesor: T. J. Hazen**

- **Manejo de la variabilidad en condiciones acústicas**
  - **Compensación del canal**
  - **Compensación del ruido de fondo**
  - **Ruidos de primer plano y artefactos no discursivos**
- **Computación y aplicación de la puntuación de seguridad**
  - **Puntuación de seguridad en el reconocimiento**
  - **Cuestiones relativas a la comprensión del lenguaje**
  - **Cuestiones relativas al modelado del diálogo**

# Grabación digital típica del discurso



- **Los reconocedores cometen errores**
- **Algunas razones para los errores:**
  - Presencia de palabras antes ocultas o eventos
  - Condiciones acústicas complejas o ruidos de fondo
  - Presencia de palabras muy prestadas a confusión
  - Cantidad insuficiente de datos de entrenamiento
  - Desequilibrio entre entrenamiento y datos de prueba
  - Modelos demasiado inflexibles para tratar la variabilidad
- **Métodos para manejar datos repletos de errores**
  - Ajustar o adaptar a las condiciones actuales
  - Identificar cuando ocurre el error y emprender acciones para la reactivación

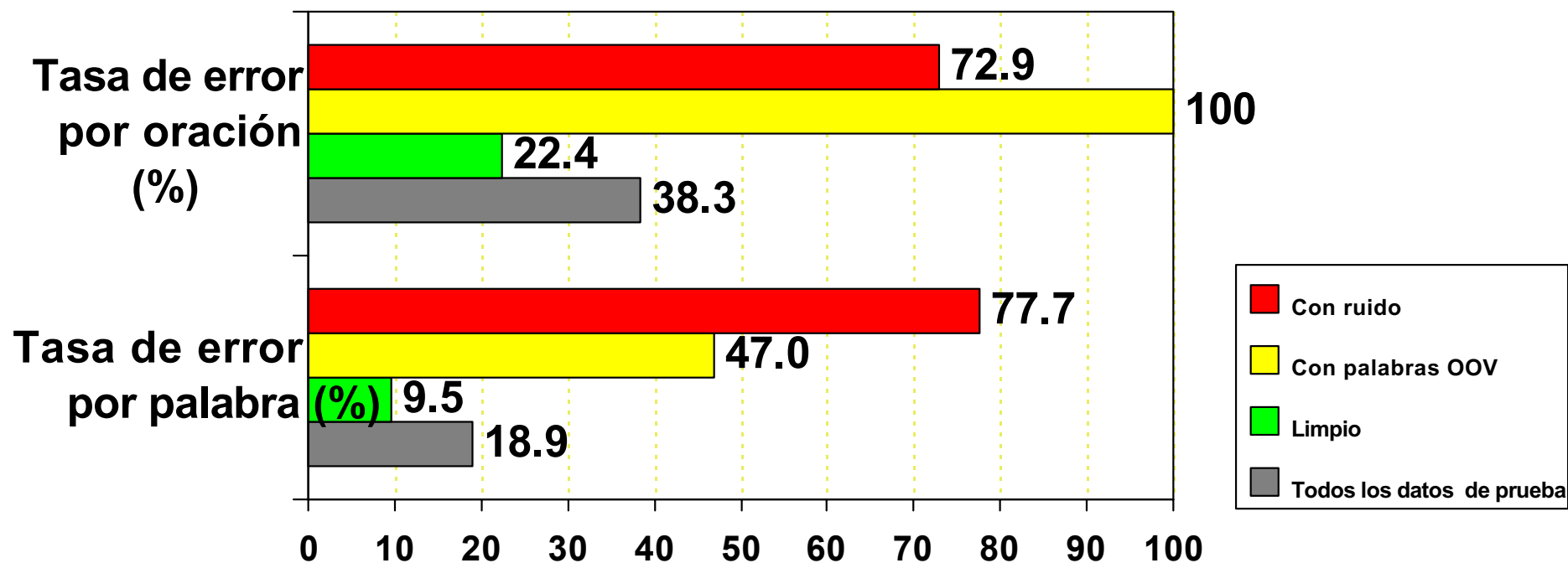
# Ruidos y artefactos no discursivos

- **Los artefactos no discursivos pueden ser sumamente variados**
  - Ruidos de fondo (música, ladridos, portazos, etc.)
  - Ruidos del micrófono y del canal (clics, silbidos, canal estático, etc.)
  - Ruidos del hablante sin contenido léxico (tos, risa, ruidos en la cavidad bucal, etc.)
- **Los ruidos se pueden simultanear con el discurso**



# Experimentos de reconocimiento

- Experimentos con el reconocedor JUPITER como punto de referencia
  - Limpio ① Sin palabras OOV y sin artefactos no discursivos
  - Con ruido ① Contiene al menos un artefacto no discursivo
  - Con palabras OOV ① Contiene al menos una palabra OOV



# Canal complejo y condiciones de ruido

- **Funciones variables del sistema**
  - Desde distintos canales (ej., línea de tierra, teléfono móvil, etc.)
  - Micrófonos distintos
- **Ruido de fondo constante**
  - Canal estático
  - Ruido del motor de un coche
  - Siseo del aire acondicionado
- **Primer plano intermitente o ruidos de fondo**
  - Tos
  - Risa
  - Portazo
  - Golpecitos o clics con los auriculares
  - Sonido del teléfono
  - Ladridos

# MIT Normalización cepstral media

- El canal de una grabación discursiva puede modelarse como un filtro lineal invariante temporal:

$$y[n] = s[n] * f[n]$$

discurso grabado      discurso original      filtro del canal

- En el dominio de la frecuencia, esto se convierte en:

$$Y(\omega) = S(\omega)F(\omega)$$

- En el dominio de la frecuencia logarítmica, esto se convierte en:

$$\log Y(\omega) = \log S(\omega) + \log F(\omega)$$

- En el dominio cepstral, esto se convierte en:

$$c[n] = \hat{s}[n] + \hat{f}[n]$$

# Normalización cepstral media (continuación)

- Durante el reconocimiento, el discurso se procesa en tramos
- Sea  $c[n, m]$  el coeficiente cepstral enésimo del tramo emésimo:

$$c[n, m] \equiv \hat{s}[n, m] \ominus \hat{f}[n, m]$$

- Dado que el filtro del canal es lineal invariante temporal:

$$\hat{f}[n, m] \equiv \hat{f}[n] \quad \Rightarrow \quad c[n, m] \equiv \hat{s}[n, m] \ominus \hat{f}[n]$$

- **Objetivo: Eliminar el efecto del filtro**
- Comenzar por calcular el promedio del cepstro por tramos:

$$\bar{c}[n] \equiv \frac{1}{M} \sum_{m=1}^M c[n, m] = \hat{f}[n] + \frac{1}{M} \sum_{m=1}^M \hat{s}[n, m]$$

# Normalización cepstral media (continuación)

- La normalización cepstral media es:

$$c'[n, m] = c[n, m] - \bar{c}[n]$$

$$= \left( \hat{s}[n, m] + \hat{f}[n] \right) - \left( \hat{f}[n] + \frac{1}{M} \sum_{m=1}^M \hat{s}[n, m] \right)$$

$$= \hat{s}[n, m] - \frac{1}{M} \sum_{m=1}^M \hat{s}[n, m]$$

Se eliminan las  
prop. del filtro

Se elimina también  
el cepstro medio  
del discurso

- Útil cuando la variación del filtro es mayor que la variación del hablante
  - Referencia: Furui, 1981

# MIT

## Manejo del ruido de fondo

- **Entrenamiento multiestilo**
  - Entrenar con datos desde una variedad de entornos ruidosos
  - Problema: Estimaciones pobres para entornos nuevos o inesperados
  - Referencia: Lippmann, *et al*, 1987
- **Sustracción espectral**
  - Estimar los componentes espectrales estáticos durante el silencio
  - Sustraer los componentes espectrales estáticos de espectros dinámicos
  - Problema: Estimaciones pobres del discurso en regiones con una proporción señal a ruido baja
  - Referencia: Boll, 1979
- **Reconocimiento basado en subbanda**
  - Ejecutar reconocedores de "subbanda" paralelos
  - Los reconocedores de subbanda operan en bandas espectrales distintas
  - Subbandas de pesos basadas en su proporción señal a ruido
  - Problema: Utilizar múltiples reconocedores es computacionalmente caro
  - Referencia: Boulard and Dupont, 1996

# Combinación paralela del modelo

- **Combinación paralela del modelo (PMC) para la compensación del ruido de fondo**
  - Entrenar modelos acústicos discursivos en el discurso limpio
  - Estimar el modelo de ruido para las condiciones actuales
  - Combinar modelos discursivos limpios con modelos de ruido estimados
- **El método supone que el espectro medio de la señal puede estimarse contrariamente al vector medio del modelo**
  - Modelo de discurso limpio para la unidad fonética  $u$ :

$$P(\vec{s} | u) \equiv N(\vec{\mu}_u, \Sigma_u) \implies S(\omega) = F^{-1}(\vec{\mu}_u)$$

- Modelo de ruido estimado a partir de la región no discursiva de las condiciones actuales:

$$P(\vec{n}) \equiv N(\vec{\mu}_n, \Sigma_n) \implies N(\omega) = F^{-1}(\vec{\mu}_n)$$

# Combinación paralela del modelo

- **Dadas las estimaciones de los valores espectrales medios del discurso limpio y del ruido, realizar la combinación:**

$$\vec{\mu}'_u = F(S(\omega) + N(\omega)) = F\left(F^{-1}(\vec{\mu}_u) + F^{-1}(\vec{\mu}_n)\right)$$

$$P_{\text{PMC}}(\vec{a} | u) \equiv N(\vec{\mu}'_u, \Sigma_u)$$

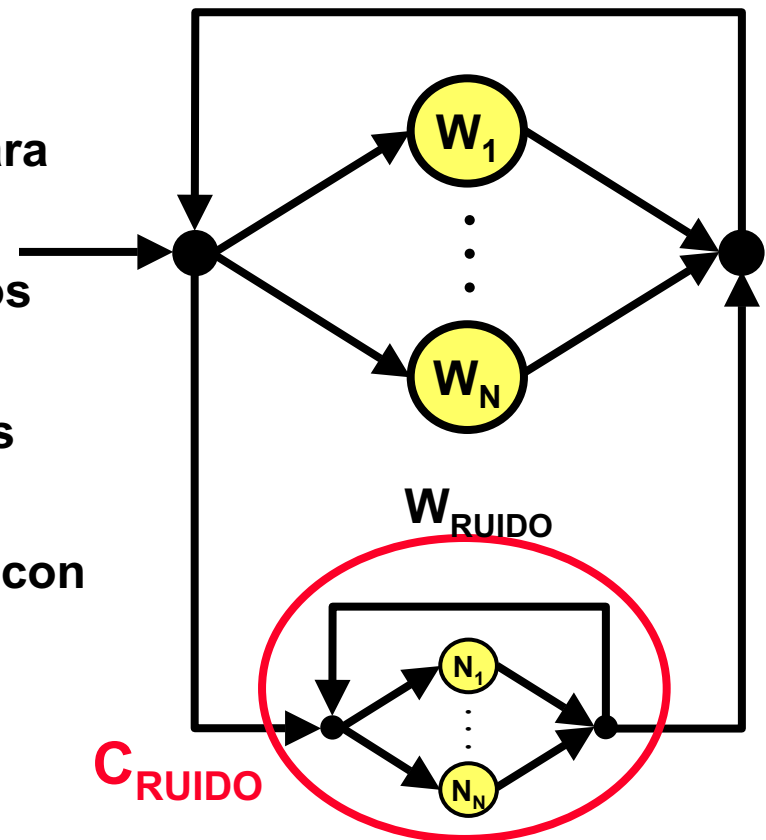
- **Cuestiones:**
  - Debe ser capaz de invertir el espectro de la estimación desde el modelo medio
  - Debe tener una estimación fiable de las condiciones de ruido actuales
- **Referencia: Gales, 1996**

# Manejo de ruidos de primer plano

- Construir modelos explícitos para ruidos distintos y artefactos no discursivos
  - Referencia: Ward, 1989

- Un posible enfoque:

- Construir una red del modelo acústico para cada modelo de ruido
- La red de ruido contiene múltiples estados para modelar los ruidos dinámicos
- Añadir redes de ruido a la red de palabras como nuevas palabras
- Controlar el índice de detección de ruido con coste  $C_{\text{RUIDO}}$



# Experimento de modelado no discursivo

- Se han añadido a JUPITER 5 modelos no discursivos
  - <tos>, <risa>, <ruido>, <fondo>, <colgar>
  - Referencia: Hazen, Hetherington y Park, 2001
- Resultados de la tasa de error por palabra:

Datos del grupo de pruebas	Punto de referencia	+ Modelos de ruido
Todos los datos	18.9%	17.1%
Datos c/ ruido	64.0%	45.1%
Datos IV c/ ruido	46.4%	28.2%
Datos IV c/ No ruido	9.4%	9.6%

**IV = únicamente datos presentes en el vocabulario**

# Perspectiva general de la puntuación de seguridad

- **Pregunta: ¿Cómo evaluamos si la hipótesis de un reconocedor es o no correcta?**
- **Objetivo: Generar puntuaciones de seguridad que estimen la probabilidad de que una hipótesis sea correcta**
- **Las puntuaciones pueden computarse en múltiples niveles:**
  - Puntuaciones fonéticas
  - Puntuaciones de palabra
  - Puntuaciones de enunciado
- **Un enfoque:**
  - Encontrar rasgos correlacionados con la correctitud
  - Construir un vector característico a partir de rasgos válidos
  - Construir un clasificador correcto/incorrecto para el vector característico

# Puntuaciones de probabilidad acústica

- Una puntuación de probabilidad acústica se computa como:

$$p(\vec{x} | u)$$

- Las puntuaciones de probabilidad acústica son buenas para comparar hipótesis distintas
  - Las puntuaciones son probabilidades de densidad relativa, no probabilidades
- Las puntuaciones de probabilidad no facilitan una estimación buena de la correctitud o fiabilidad

# Puntuaciones acústicas normalizadas

- La expresión de probabilidad *a posteriori* es:

$$p(u | \vec{x}) = \frac{p(\vec{x} | u)}{p(\vec{x})} p(u)$$

**puntuación de probabilidad acústica normalizada**

- En un marco probabilístico  $p(\vec{x})$  normalmente se ignora
- El reconocimiento no está afectado por la normalización
  - el modelo de normalización es independiente de la identidad del fono
  - las puntuaciones normalizadas pueden considerarse puntuaciones de seguridad

# Puntuaciones acústicas normalizadas

- Teóricamente el modelo de normalización es:

$$p(\vec{x}) = \sum_{\forall u} p(\vec{x} | u) p(u)$$

- En la práctica, la normalización se lleva a cabo con un modelo aproximado de  $p(\vec{x})$
- La aproximación de  $p(\vec{x})$  mediante agrupamiento ascendente:
  - Componentes gaussianos similares combinados
  - El modelo combinado es una aproximación de ML de componentes mezcla que serán combinados
  - La combinación continúa hasta que se haya alcanzado el tamaño deseado
  - El modelo de normalización típicamente presenta entre 50 y 100 componentes mezcla en reconocedores SLS

# Rasgos de seguridad de palabra

- **Se pretende extraer información de la computación del reconocimiento que va correlacionado con la correctitud**
- **Posibles rasgos de seguridad a nivel de palabra extraídos de puntuaciones acústicas:**
  - **Puntuación acústica media normalizada por palabra**
  - **Puntuación acústica mínima normalizada por palabra**
  - **Puntuación del modelo de normalización media**
- **Otras fuentes de información:**
  - **N-mejores puntuaciones de pureza**
  - **Puntuaciones del modelo de lenguaje**
  - **Número de hipótesis participantes**
  - **Diferencias de puntuación relativa entre hipótesis**
- **Referencia: Chase, 1997**

# MIT

## La medida de pureza n-mejor

- La pureza n-mejor es la fracción de las hipótesis N-mejores en las que aparece la hipótesis de palabra

---

(1) *what is the weather in new york*

1.0 0.8 0.6 1.0 1.0 0.6

---

(2) *what is the weather in newark*

1.0 0.8 0.6 1.0 1.0 0.4

---

(3) *what is <uh> weather in new york*

1.0 0.8 0.4 1.0 1.0 0.6

---

(4) *what is <uh> weather in newark*

1.0 0.8 0.4 1.0 1.0 0.4

---

(5) *what was the weather in new york*

1.0 0.2 0.6 1.0 1.0 0.6

---

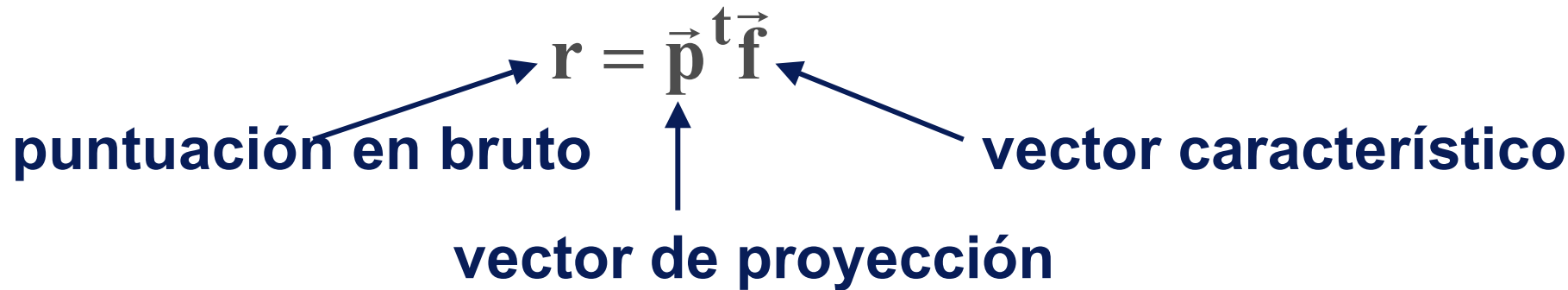
*newark* está en  
2 de 5 hipótesis  
∩ pureza =  $2/5 =$   
0.4

# MIT Clasificación de la seguridad

- Dado un vector característico de seguridad, queremos clasificar el vector como *correcto* o *incorrecto*
- Este es un problema de clasificación estándar de dos clases
- Posibles enfoques:
  - Proyección lineal discriminante (Pao, *et al*, 1998)
  - Clasificador de red neural (Wendemuth, *et al*, 1999)
  - Clasificador de mezcla gaussiano (Kamppari & Hazen, 2000)
  - Máquinas con vector de soporte (Ma, *et al*, 2001)

# Clasificador discriminante lineal

- **Proyección lineal discriminativa aplicada al vector característico de seguridad:**



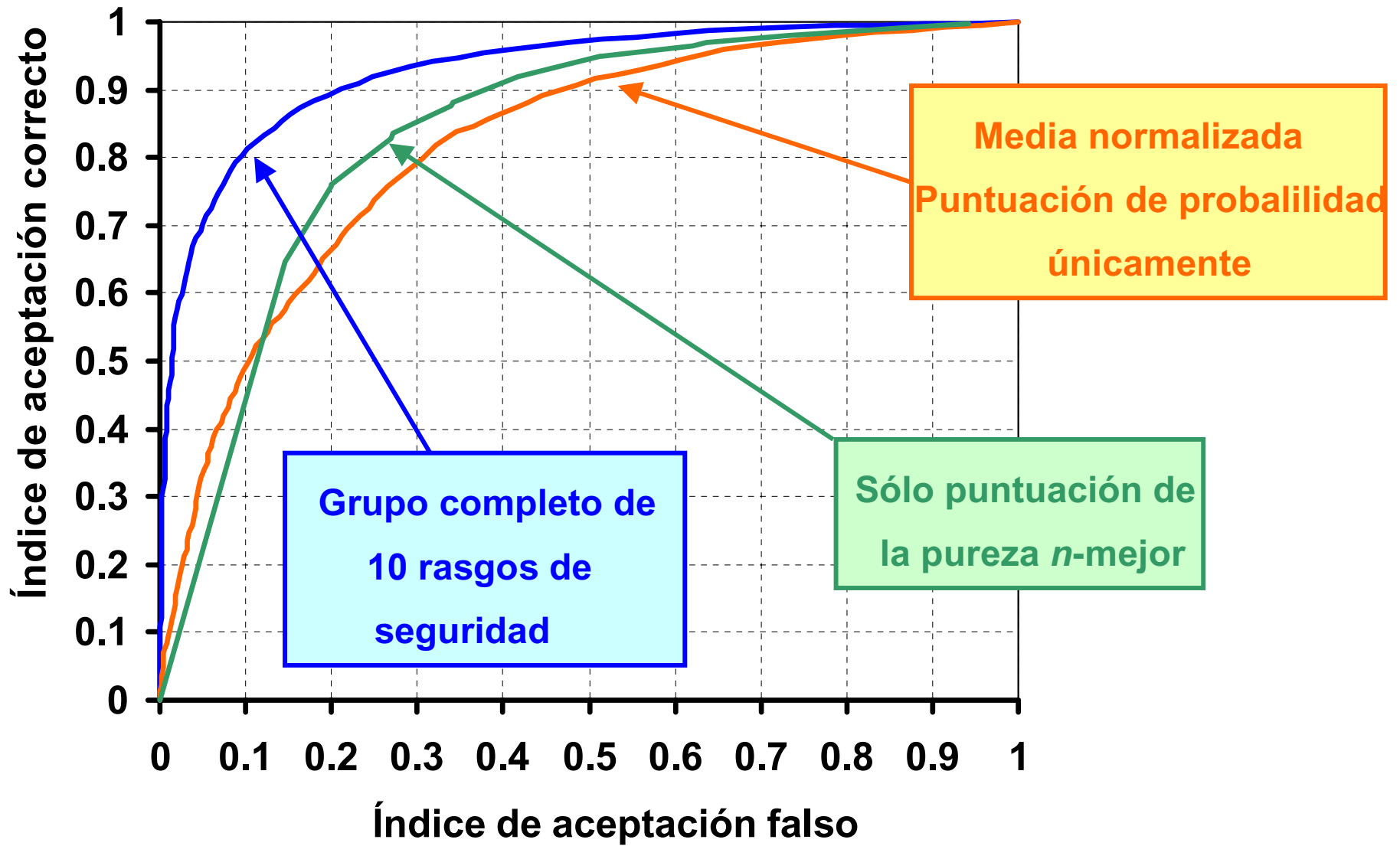
- **Vector de proyección:**
  - Entrenado con un grupo de desarrollo independiente
  - Entrenamiento con error de clasificación mínimo (MCE)
  - MCE realiza un entrenamiento de descenso gradiente en la tasa de error



# Experimento de seguridad de palabras

- **Se pretenden rechazar palabras hipotéticas para las que el reconocedor da una seguridad baja**
- **Entrenar al modelo de seguridad con datos de desarrollo independientes**
- **Probar con el grupo de pruebas independientes de datos de JUPITER**
- **Evaluar usando la curva ROC**
  - **Examina las aceptaciones correctas frente a las aceptaciones falsas**
  - **Se pretenden rechazar palabras hipotéticas incorrectas y aceptar palabras hipotéticas correctas**
  - **Resultados demostrados para dos rasgos individuales y para vectores característicos absolutos con 10 rasgos**
- **Referencia: Hazen, *et al*, 2002**

# Resultados de seguridad de palabra



# Empleo de puntuaciones de seguridad

- **Para ser útil, las puntuaciones de seguridad deben integrarse con la comprensión del lenguaje y el modelado del diálogo**
- **Las puntuaciones de seguridad normalmente se cuantizan en dos o tres regiones de decisión:**
  - **Aceptar o rechazar (dos regiones)**
  - **Aceptar, rechazar o dudar (tres regiones)**
- **El componente de comprensión del lenguaje puede adaptarse para el manejo de palabras rechazadas**
- **El componente de la gestión del diálogo puede realizar distintas acciones basadas en la puntuación de seguridad**
  - **Llevar a cabo una acción normal cuando todo se ha aceptado**
  - **Pedir confirmación cuando hay duda**
  - **Pedir al usuario que repita o reformule cuando se de el rechazo**
- **Referencia: Hazen, *et al*, 2002**

# Modificaciones de la lista de $N$ -mejores

*What is the forecast for Paramus Park, New Jersey?*

Lista estándar de  $N$ -mejores con puntuaciones de seguridad:

what_is	6.13	the	5.48	forecast	6.88	for	5.43	paris	-0.03	park	4.41	new_jersey	4.35
what_is	6.13	the	5.48	forecast	6.88	for	4.47	hyannis	-0.61	park	4.41	new_jersey	4.35
what_is	6.13	the	5.48	forecast	6.88	for	5.12	venice	-0.89	park	4.41	new_jersey	4.35
what_is	6.13	the	5.48	forecast	6.88	for	4.28	france	-1.12	park	4.41	new_jersey	4.35

Lista de  $N$ -mejores con *fuerte rechazo* de palabras con puntuación baja:

what_is	6.13	the	5.48	forecast	6.88	for	5.43	*reject*	0.00	park	4.41	new_jersey	4.35
what_is	6.13	the	5.48	forecast	6.88	for	4.47	*reject*	0.00	park	4.41	new_jersey	4.35
what_is	6.13	the	5.48	forecast	6.88	for	5.12	*reject*	0.00	park	4.41	new_jersey	4.35
what_is	6.13	the	5.48	forecast	6.88	for	4.28	*reject*	0.00	park	4.41	new_jersey	4.35

# Modificaciones de la lista de *N*-mejores (cont.)

## Lista de *N*-mejores con *rechazo opcional*:

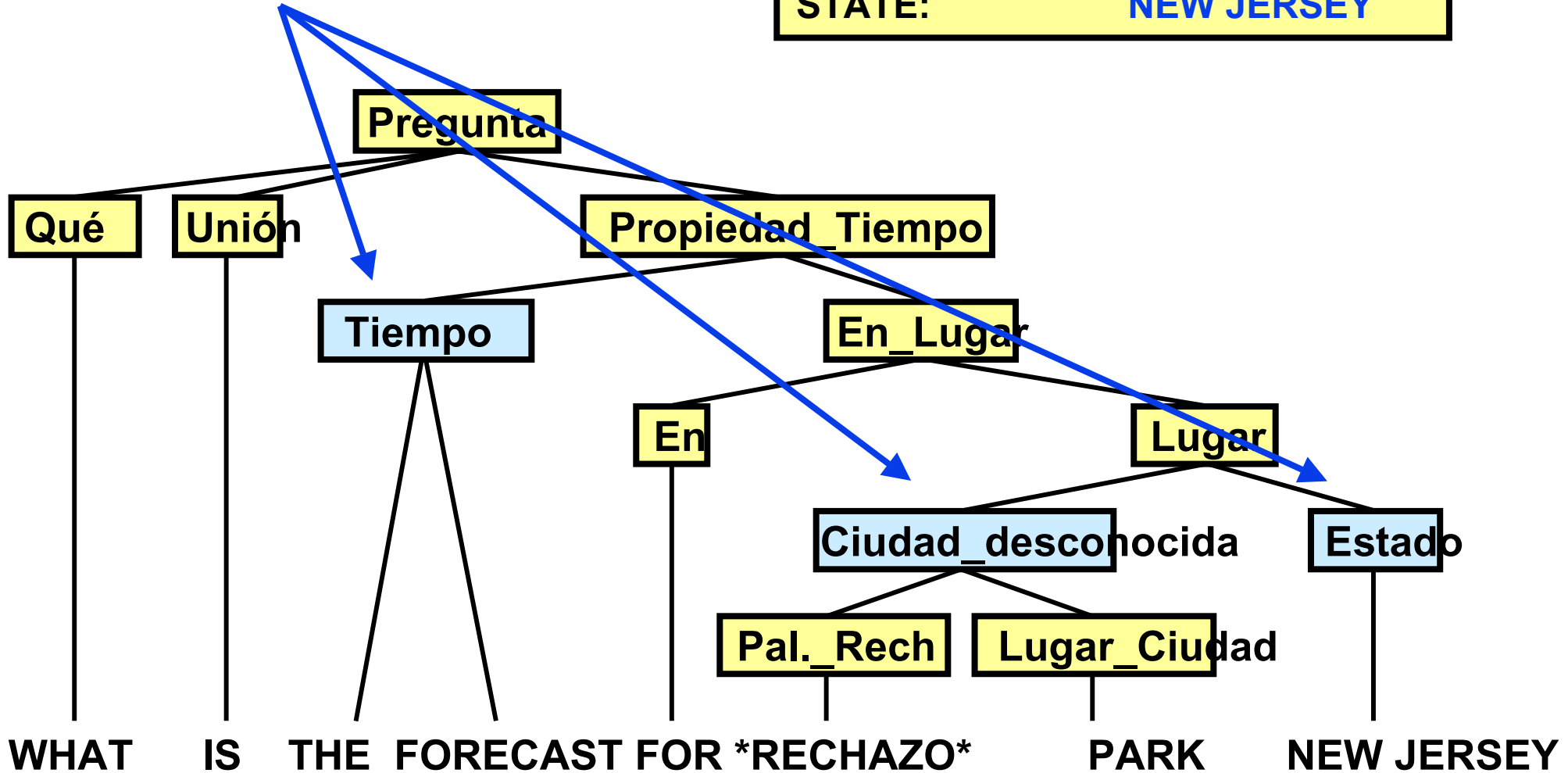
what_is	6.13	the	5.48	forecast	6.88	for	5.43	paris	-0.03	park	4.41	new_jersey	4.35
what_is	6.13	the	5.48	forecast	6.88	for	5.43	*reject*	0.00	park	4.41	new_jersey	4.35
what_is	6.13	the	5.48	forecast	6.88	for	4.47	hyannis	-0.61	park	4.41	new_jersey	4.35
what_is	6.13	the	5.48	forecast	6.88	for	4.47	*reject*	0.00	park	4.41	new_jersey	4.35
what_is	6.13	the	5.48	forecast	6.88	for	5.12	venice	-0.89	park	4.41	new_jersey	4.35
what_is	6.13	the	5.48	forecast	6.88	for	5.12	*reject*	0.00	park	4.41	new_jersey	4.35
what_is	6.13	the	5.48	forecast	6.88	for	4.28	france	-1.12	park	4.41	new_jersey	4.35
what_is	6.13	the	5.48	forecast	6.88	for	4.28	*reject*	0.00	park	4.41	new_jersey	4.35

Las palabras con puntuaciones de seguridad pobres, compiten con palabras rechazadas durante la búsqueda de comprensión del lenguaje natural

# Ejemplo de árbol de análisis de comprensión

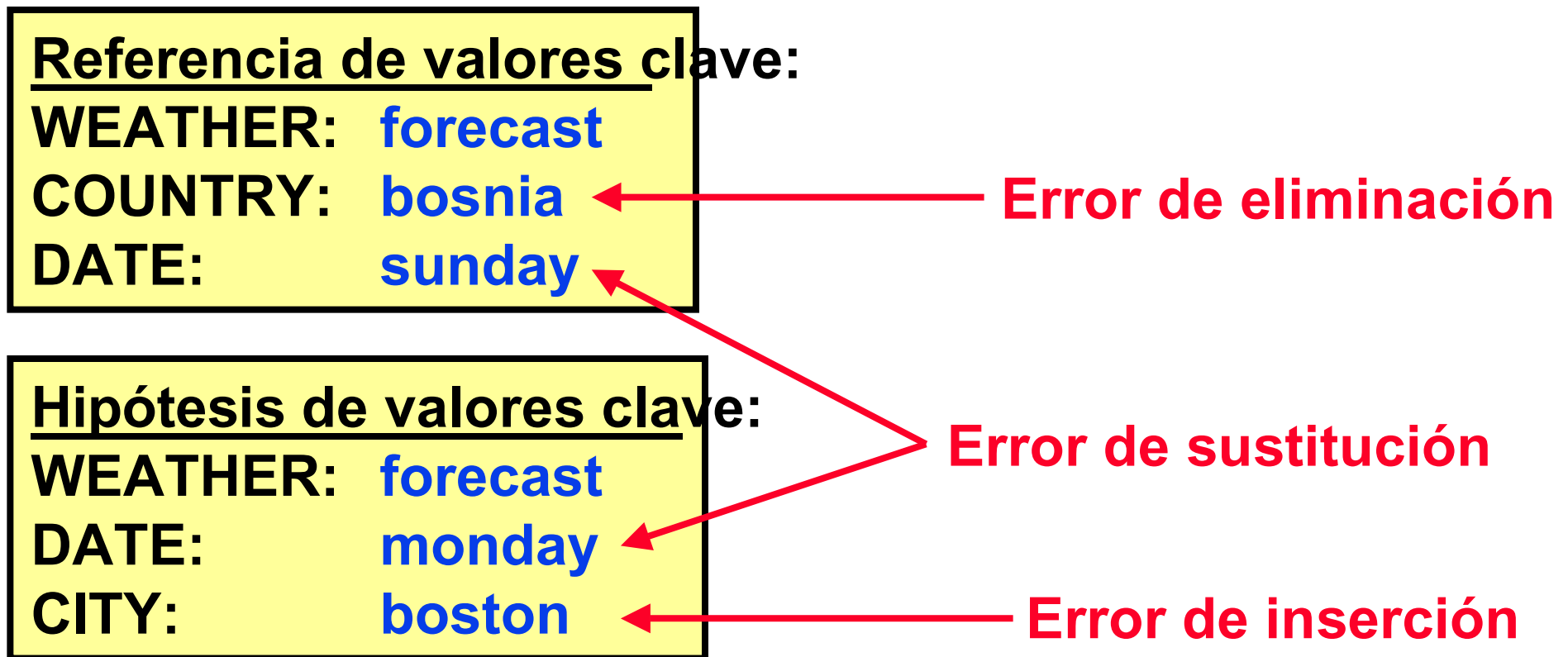
Conceptos semánticos extraídos expresados como pares clave-valor

WEATHER:	FORECAST
UNKNOWN_CITY:	*REJECT* PARK
STATE:	NEW JERSEY



# Evaluación de la comprensión

- Rechazo de puntuación de seguridad evaluado con la comprensión del *índice de error conceptual* (CER)
- El CER se calcula a partir de sustituciones, inserciones y eliminaciones de pares de concepto clave-valor



# Resultados de comprensión

- Resultados de comprensión experimental en JUPITER
- El sistema original no utilizaba puntuaciones de seguridad durante la selección de la hipótesis en el componente de comprensión

Condiciones experimentales	Índice de error conceptual (CER) (%)			
	Sub	Ins	Del	Total
Sistema original	2.2	19.9	6.3	28.5
Nuevo sistema/No rech	2.1	18.1	6.1	26.2
+ rechazo opcional	1.3	8.9	8.5	18.7
+ rechazo fuerte	1.0	7.0	10.6	18.6

35% de  
reducción  
del CER

Sub = Sustitución

Ins = Inserción

Del = Elimin.

# Ejemplos de modelado de diálogo

Ayudar al usuario proporcionándole información adicional:

— USUARIO: *what is the forecast for \*rechazar\* park new jersey*

— VALORES CLAVE: WEATHER: forecast  
UNKNOWN\_CITY: \*rechazar\* park  
STATE: new jersey

— RESPUESTA: *I may not know the city you are asking about.  
I know the following places in New Jersey: Atlantic City,  
Camden, Newark, Piscataway, Princeton, and Trenton.  
Please select one.*

# Más ejemplos de modelado de diálogo

Orientar al usuario hacia una pregunta que pueda ser controlada:

- USUARIO: *what is the forecast for \*rechazar\* \*rechazar\**
- VALORES CLAVE: WEATHER: *forecast*  
UNKNOWN\_CITY: *\*rechazar\* \*rechazar\**
- RESPUESTA: *I may not know the city you are asking about. Try asking for the state or country. For example: “What cities do you know about in Massachusetts?”*

# MIT Más ejemplos de modelado de diálogo

## Pidiendo confirmación:

- USUARIO : *what is the forecast for \*boston\**
- VALORES CLAVE: WEATHER: **forecast**  
CONFIRM\_CITY: **boston**
- RESPUESTA: *Did you say Boston?*

# Referencias

- S. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” *IEEE Trans. on Acoustic, Speech, and Signal Processing*, April 1979.
- H. Bourlard and S. Dupont, “A new ASR approach based on independent processing and recombination of partial frequency bands,” *ICSLP*, 1996.
- L. Chase, “Word and acoustic confidence annotation for large vocabulary speech recognition,” *Eurospeech*, 1997.
- S. Furui, “Cepstral analysis techniques for automatic speaker verification,” *IEEE Trans. on Acoustic, Speech, and Signal Processing*, April 1981.
- M. Gales y S. Young, “Robust continuous speech recognition using parallel model combination,” *IEEE Trans. on Acoustic, Speech, and Signal Processing*, September 1996.
- T. Hazen, L. Hetherington y A. Park, “FST-based recognition techniques for multi-lingual and multi-domain spontaneous speech,” *Eurospeech*, 2001.
- T. Hazen, J. Polifroni y S. Seneff, “Recognition confidence scoring for use in speech understanding systems,” *Computer Speech and Language*, January, 2002.

# Referencias

- **C. Pao, P. Schmid, y J. Glass, “Confidence scoring for speech understanding,” ICSLP, 1998.**
- **S. Kamppari y T. Hazen, “Word and phone level acoustic confidence scoring,” ICASSP, 2000.**
- **R. Lippman, E. Martin y D. Paul, “Multi-style training for robust isolated-word speech recognition,” ICASSP, 1987.**
- **C. Ma, M. Randolph y J. Drish, “A support vector machine-based rejection technique for speech recognition,” ICASSP, 2001.**
- **W. Ward, “Modelling non-verbal sounds for speech recognition,” DARPA Speech and Natural Language Workshop, Octubre, 1989.**
- **A. Wendemuth, G. Rose y J. Dolfing, “Advances in confidence measures for large vocabulary,” ICASSP, 1999.**