PONDERACIÓN DE LAS PROBABILIDADES DE OBSERVACIÓN DE LOS HMM PARA VERIFICACIÓN DE LOCUCIONES EN SISTEMAS DE RAH

Marta Casar, José A. R. Fonollosa

Centre TALP, Universitat Politècnica de Catalunya
C. Jordi Girona 1-3, 08034, Barcelona
email: {mcasar,adrian}@gps.tsc.upc.edu

RESUMEN

Las tecnologías de reconocimiento del habla requieren sistemas fiables capaces de funcionar correctamente en diferentes tareas y entornos. Al no ser posible conseguir un reconocimiento libre de errores en la mayoría de casos, surge la necesidad de disponer de una medida de la fiabilidad del sistema. En este artículo se propone la ponderación de la probabilidad de observación de los modelos de Markov, equilibrando la diferencia entre rangos dinámicos de las probabilidades de transición y de observación. El parámetro definido *OPW* es independiente de la gramática utilizada, a diferencia de otras soluciones similares. Dado que se espera que las variaciones introducidas por este nuevo parámetro tengan mayor influencia en las palabras incorrectamente reconocidas, se propone un sistema de verificación de locuciones basado en una 'segunda opinión'. Para ello se utiliza la salida de decodificar utilizando diferentes $OPW \neq 1$. Resultados experimentales conseguidos hasta el momento muestran la validez de la propuesta.

1. INTRODUCCIÓN A LOS HMM

La mayoría de sistemas de reconocimiento del habla actuales se basan en un conjunto de modelos estadísticos (también llamados modelos acústicos) que relacionan las características observables de la señal de voz con el conjunto de unidades fonéticas teóricas. La implementación más usual de estos modelos estadísticos se basa en la utilización de los Modelos Ocultos de Markov (HMM) [1].

Los HMM resultan una potente herramienta estadística para modelar señales de voz. Consisten en una cadena de Markov, la probabilidad de salida de la cual es una variable aleatoria X generada mediante una función de probabilidad de salida asociada a cada estado. Así pues, no existe una correspondencia unitaria entre la secuencia de observaciones y la secuencia de estados, de forma que esta secuencia de estados permanece 'oculta'. Formalmente, un HMM se define como:

$$\lambda = (A, B, \pi)$$

Este trabajo ha sido parcialmente subvencionado por el MCyT a través del proyecto TIN-2005-08852

siendo A la distribución de probabilidad de transición entre estados $A = a_{ij}$, B la distribución de probabilidad de observación de símbolos en el estado j, $B = b_j(k)$, y π la distribución de probabilidad inicial para cada estado $\pi = \pi_i$.

En procesado del habla, la secuencia de estados subyacente asociada a los modelos HMM se caracteriza por su configuración temporal (izquierda-a-derecha): a medida que se incrementa la variable temporal, el índice de estados se incrementa a su vez o bien se mantiene igual. En un instante de tiempo, cada trama de voz puede estar en un único estado.

En HMM semicontinuos (SCHMM) cada estado del modelo contiene un conjunto de funciones de probabilidad de emisión, generalmente funciones densidad de probabilidad de mezclas de gaussianas. Para cada estado, estas funciones proporcionan la probabilidad con la cual ese estado puede generar cada trama. Las funciones de densidad de probabilidad están ligadas de forma conjunta para todos los modelos, formando un codebook que permite mapear el vector de características de entrada continuo X y transformarlo en O_k. A partir de este vector, podremos utilizar una distribución de probabilidad de salida discreta de la forma $b_j(k)$.

La forma más intuitiva de calcular la probabilidad $P(\mathbf{O}|\lambda)$ de la secuencia de observaciones $\mathbf{O} = (O_1, O_2, \ldots, O_T)$ dado un HMM λ, consiste en sumar las probabilidades de todas la secuencias de estados posibles:

$$P(\mathbf{O}|\lambda) = \sum_{allS} P(S|\lambda)P(\mathbf{O}|S,\lambda) \qquad (1)$$

Para una secuencia de estados particular $S = (s_1, s_2, \ldots, s_T)$, la probabilidad de salida conjunta se puede expresar como:

$$P(\mathbf{O}|S,\lambda) = \prod_{i=1}^{T} P(O_t|s_t,\lambda) = b_{s_1}(O_1)b_{s_2}(O_2)\ldots b_{s_T}(O_T)$$

y así, podemos reescribir la ecuación (1):

$$P(\mathbf{O}|\lambda) = \sum_{allS} a_{s_0 s_1} b_{s_1}(O_1)\ldots a_{s_{T-1}s_T} b_{s_T}(O_T) \quad (2)$$

2. PONDERACIÓN DE LAS PROBABILIDADES DE OBSERVACIÓN DE LOS MODELOS

En la mayoría de sistemas de reconocimiento del habla, la decodificación consiste en un proceso de búsqueda de la secuencia de palabras $\hat{\mathbf{W}} = \mathrm{w}_1\mathrm{w}_1 \dots \mathrm{w}_m$ de máxima probabilidad a posteriori $P(\mathbf{W}|\mathbf{O})$ correspondiente a una determinada observación acústica $\mathbf{O} = O_1 O_2 \dots O_n$. Expresado formalmente:

$$\hat{\mathbf{W}} = \arg\max_{\mathrm{w}} P(\mathbf{W}|\mathbf{O}) = \arg\max_{\mathrm{w}} \frac{P(\mathbf{W})P(\mathbf{O}|\mathbf{W})}{P(\mathbf{O})}$$
$$= \arg\max_{\mathrm{w}} P(\mathbf{W})P(\mathbf{O}|\mathbf{W}) \quad (3)$$

donde $P(\mathbf{W})$ representa la probabilidad del modelo de lenguaje, y $P(\mathbf{O}|\mathbf{W})$ la probabilidad del modelo acústico.

A partir de esta ecuación parece que ambas probabilidades puedan ser combinadas mediante una simple multiplicación. Sin embargo, al trabajar con los HMM como modelos acústicos las probabilidades acústicas suelen estar subestimadas (ver [1]). Al combinarlas con las probabilidades de los modelos de lenguaje, se acaba dando a estos modelos un peso insuficiente. Además, los rangos dinámicos de ambas probabilidades difieren al trabajar con modelos continuos o semicontinuos. Por consiguiente, es práctica corriente ponderar la probabilidad del modelo de lenguaje mediante un peso LW (*language model weight*) determinado de forma empírica (típicamente > 1), transformando $P(\mathbf{W})$ en $P(\mathbf{W})^{LW}$.

De forma equivalente, la idea de ponderar el valor de las probabilidades de observación de los HMM surge de la definición (y construcción) de estos modelos, centrándonos en la contribución de los diferentes tipos de probabilidades. Como ya se ha señalado en la introducción, se puede entender la probabilidad de una secuencia de observaciones como la suma de contribuciones de las probabilidades de transición entre estados y observación de cada estado. La ecuación (2) se puede reescribir para una determinada secuencia S_i mediante logaritmos:

$$logP(\mathbf{O}|\lambda) = log(a_{s_0 s_1} b_{s_1}(O_1) \dots a_{s_{T-1} s_T} b_{s_T}(O_T))$$
$$= P_A + P_B$$

$$\text{donde} \quad P_A = log(a_{s_0 s_1} a_{s_1 s_2} \dots a_{s_{T-1} s_T})$$
$$P_B = log(b_{s_1} b_{s_2} \dots b_{s_T})$$

Las probabilidades de transición están asociadas a la gramática utilizada, de forma que la relación entre los rangos dinámicos de P_A y P_B se puede ver desequilibrada por restricciones impuestas por dicha gramática. Introduciendo un factor de ponderación de las probabilidades de observación *OPW* (*Observation Probability Weight*) la contribución de las probabilidades de observación de los símbolos ser verán modificadas, pasando a ser $\mathrm{OPW} \cdot P_B$. *OPW* será determinado de forma empírica con el objetivo de corregir la diferencia de rangos dinámicos que pudiese existir entre ambas probabilidades. *OPW* se asemeja así al factor *LW* presentado, con la diferencia de ser independiente de la gramática.

3. RECONOCIMIENTO DEL HABLA UTILIZANDO OPW

Partiendo de un conjunto de HMMs convencionales se han ponderado las probabilidades de observación de estos modelos utilizando el factor *OPW* definido. Los modelos resultantes han sido utilizados para el reconocimiento de cadenas de dígitos. Las pruebas de reconocimiento realizadas nos han permitido estudiar la viabilidad de utilizar una decodificación alternativa basada en *OPW* para verificación de locuciones.

3.1. Sistema de reconocimiento del habla

El sistema de reconocimiento del habla utilizado para estos experimentos es RAMSES [2], basado en SCHMM. Las principales características de este sistema son:

- La señal de voz es enventanada cada 10ms con ventanas de 30ms de longitud. Cada trama se parametriza utilizando coeficientes cepstrales melfrequency (MFCC) y su primera y segunda derivadas, junto a la primera derivada de la energía.
- Los parámetros espectrales son cuantificados utilizando 512 centroides, y la energía utilizando 64.
- Las unidades acústicas utilizadas en los HMM son semidígitos. Se entrenan 40 modelos de semidígitos, además de un modelo ruidoso para cada dígito, modelados mediante 10 estados. Se utilizan también modelos de silencio y de relleno, modelados mediante 8 estados.
- Se utiliza un algoritmo de Viterbi en decodificación, implementando búsqueda en haz para limitar el número de caminos. Para ello, las tramas se cuantifican utilizando 6 centroides para los parámetros espectrales, y 2 para la energía.

3.2. Bases de datos

Dos bases de datos diferentes han sido utilizadas a lo largo de estos experimentos. Primero, a partir de la base de datos en español de SpeechDat [3] se ha creado un subconjunto de entrenamiento con 11443 frases y un subconjunto de test con 3405 frases, todas ellas formadas por secuencias de dígitos. Los resultados obtenidos a partir de este primer subconjunto de test se han utilizado para seleccionar (de forma empírica) el valor óptimo de *OPW*.

A continuación, los nuevos modelos se han puesto a prueba utilizando una base de datos independiente, que llamaremos DigitVox, obtenida mediante una aplicación telefónica de reconocimiento de voz. Esta base de datos contiene 5317 frases correspondientes a DNIs (cadenas de 8 dígitos) grabadas en condiciones ruidosas. Estos experimentos tienen como objetivo probar la independencia de los modelos obtenidos, trabajando en condiciones 'reales'. Además, se pretende comprobar si los resultados

Configuración	Tasa de reconocimiento de FRASES	Tasa de reconocimiento de PALABRAS	Sustituciones	Inserciones	Borrados
modelos originales (*OPW* =1)	93.304 %	98.73 %	0.24 %	0.97 %	0.06 %
OPW = 0.2	93.831 %	99.09 %	0.23 %	0.40 %	0.28 %
OPW = 0.5	93.511 %	98.76 %	0.23 %	0.84 %	0.17 %
OPW = 2.0	90.856 %	98.26 %	0.25 %	1.45 %	0.03 %

Tabla 1. Tasas de reconocimiento con diferentes valores de *OPW* en la ponderación de las probabilidades de observación

obtenidos para ciertos valores de *OPW* en los primeros tests se deben al sobreentrenamiento o adaptación de los modelos, que serian efectos a evitar.

3.3. Resultados

A partir de los HMM originales entrenados utilizando RAMSES, se han construido nuevos conjuntos de modelos acústicos con diferentes valores de *OPW* entre 0.2 y 5. Los resultados obtenidos nos han permitido analizar la contribución del peso de las probabilidades de observación a las prestaciones de los modelos.

Los modelos ponderados han sido evaluados utilizando DigitVox. La tabla 1 resume los resultados obtenidos, comparándolos con los modelos originales (*OPW* = 1).

Estos resultados muestran una mejora importante para valores de $OPW < 1$, confirmando la hipótesis de que existe un desequilibrio entre las probabilidades de observación y las probabilidades de transición. Cabe recordar que el objetivo de esta propuesta no es la mejora en los resultados de reconocimiento por si misma, sino la obtención de una segunda opinión en vistas a implementar un sistema de verificación de locuciones.

4. VERIFICACIÓN DE LOCUCIONES

En todo sistema de reconocimiento hay un cierto grado de incertidumbre inherente a la decodificación obtenida. Por tanto, parece necesario obtener una medida que nos informe de la correspondencia entre la secuencia de palabras resultante y la señal de voz de entrada. Esta medida nos permitirá decidir si la salida obtenida se considera correcta o incorrecta, de una forma fiable.

Los sistemas estadísticos de reconocimiento del habla se basan en la probabilidad a posteriori $P(\mathbf{W}|\mathbf{O})$ de una palabra $\mathbf{W}$ dada una secuencia de observaciones acústicas $\mathbf{O}$. La secuencia de palabras W_{opt} que maximiza esta probabilidad a posteriori también minimiza la probabilidad de error en la frase reconocida. Por tanto, una estimación precisa de $P(\mathbf{W}|\mathbf{O})$ (ver eq. 3) nos proporciona una medida de confianza fiable, representando el ratio entre la probabilidad asociada a la hipotesis obtenida, $P(\mathbf{W})P(\mathbf{O}|\mathbf{W})$, y la probabilidad acústica. No obstante, en la mayoría de implementaciones se tiende a evitar esta costosa estimación teórica, definiendo otros parámetros y umbrales de aceptación que determinen cuándo aceptar o rechazar la salida del decodificador.

La validez de los resultados de reconocimiento ha sido un tema ampliamente estudiado en la literatura de reconocimiento del habla y enfocado desde diferentes perspectiva, ya sea utilizando modelos de hipótesis alternativas, algoritmos N-best, o grafos de palabras [4]. Uno de los principales problemas a solucionar consiste en hallar las características o parámetros que aportan más información sobre el reconocimiento o las locuciones a verificar. En este sentido, un parámetro muy útil ha sido el *LMJitter* [5] (también llamado *Acoustic Stability* [6]), que se basa en la hipótesis de que las palabras mal reconocidas son más sensibles a las variaciones del *Grammar Scale Factor* (o *LW*, según se le ha llamado en el apartado 2) [7].

Uno de los hándicaps de implementar el *LMJitter* es el elevado coste computacional que representa, para lo cuál se han propuesto algunos algoritmos eficientes [7]. Además, la implementación de sistemas de verificación basados en medidas de confianza suele realizarse utilizando varias de ellas y combinando las salidas mediante algoritmos de decisión, de forma que aumenta la complejidad del sistema.

En lugar de la utilización del *LMJitter* como una medida de confianza, otra opción anteriormente planteada consiste en modificar el peso LM entre decodificaciones. En [8] los autores realizan múltiples procesos de reconocimiento aplicando diferentes pesos para ponderar las probabilidades de los modelos de lenguaje frente las probabilidades de los modelos acústicos. En [6], esta estrategia se compara con un sistema basado en obtener las N-best soluciones de un proceso de reconocimiento trabajando con word-lattices, obteniendo prestaciones casi similares a la vez que se reducía el coste computacional.

No obstante, el coste computacional no representa un obstáculo insalvable, de forma que la utilización de un doble reconocimiento se plantea como una solución eficaz. Nuestro enfoque se centra en la hipótesis de que, igual que ocurre con el *LMJitter*, las palabras incorrectamente reconocidas serán más sensibles a las variaciones del factor de ponderación de las probabilidades de observación definido (*OPW*). La propuesta presentada consiste en desarrollar un sistema de verificación usando una estrategia basada en una 'segunda opinión' [9]. A partir de las decodificaciones obtenidas por dos sistemas de reconocimiento el sistema decidirá sobre la validez de la frase reconocida por el sistema de referencia, utilizando como segunda opinión la decodificación proporcionada por los modelos ponderados con $OPW \neq 1$.

El reconocimiento de cadenas de dígitos, que continúa siendo una aplicación de gran interés práctico, se presenta como una primera tarea sencilla a la que enfrentarse para probar la validez de nuestra propuesta.

Sistema y configuración	Exactas	Errores	Detectadas	Rechazadas	Garbage	TRR	FRR	CER
phone-based filler models	92.55 %	1.28 %	0.30 %	1.02 %	4.85 %	80.09 %	1.09 %	2.30 %
verificación OPW, *OPW* = 0.2	92.57 %	0.79 %	0.49 %	0.73 %	5.36 %	88.10 %	0.78 %	1.52 %
verificación OPW, *OPW* = 0.5	93.19 %	1.15 %	0.19 %	0.11 %	5.36 %	82.84 %	0.12 %	1.26 %

Tabla 2. Resumen de los resultados de verificación con filtrado a nivel de frase

4.1. Verificación basada en una segunda opinión

En nuestro día a día, cuando no estamos seguros de la validez de una hipótesis concreta buscamos una segunda opinión que nos haga sentir más seguros. Ello nos puede conducir a una mayor inseguridad si la nueva opinión difiere de la primera, e incluso inducirnos a un error si ambas apuntan a una misma dirección incorrecta. Por ello, resulta imprescindible tener una cierta confianza en la fiabilidad de nuestra segunda fuente.

Como ya hemos indicado, la arquitectura de verificación propuesta se basa en una doble decodificación: una primera utilizando HMM convencionales (sin ponderación), y otra utilizando modelos ponderados. Comparando ambas salidas, las frases serán clasificadas como "aceptadas" si hay consenso, o "rechazadas" en caso contrario. Para evaluar la fiabilidad de esta decisión, se etiquetan las frases resultantes en cuatro categorías: *exacta* cuando es una frase correctamente aceptada, *error* cuando ha sido incorrectamente aceptada, *detectada* si se trata de una frase incorrecta que ha sido rechazada, y *falso rechazo* si es una frase rechazada que era correcta. Para ello, las dos decodificaciones se clasifican previamente como correctas o incorrectas mediante un alineado. Así, las fases exactas son aquellas correctamente reconocidas por ambos sistemas; errores las incorrectamente reconocidas por ambos; detectadas las frases incorrectamente reconocidas por el sistema de referencia; y falso rechazo aquellas incorrectamente reconocidas utilizando los modelos ponderados.

Puesto que la salida del reconocedor consiste en una cadena de palabras reconocidas, puede realizarse un primer "filtrado a nivel de frase" previamente a la comparación entre ambas decodificaciones. Se obtiene así una primera decisión de aceptación/rechazo de las hipótesis en base a diferentes reglas (p.ej. longitud de la frase, presencia de palabras no presentes en el vocabulario, etc.). Aquellas frases rechazadas en función de esta decisión se etiquetarán como *garbage*.

4.2. Sistema de referencia de verificación

Con la finalidad de evaluar los posibles beneficios aportados por nuestra propuesta, ésta se ha comparado con un algoritmo de verificación estándar basado en modelos de fonemas de relleno (o *phone-based filler models* [10]). Este método consiste en normalizar las probabilidades de salida del reconocedor mediante una decodificación basada en fonemas independientes, sin utilizar gramáticas ni otros modelos de lenguaje. Una vez normalizadas, dichas probabilidades se convierten en una medida de la calidad del reconocimiento al proporcionar una estimación de la semejanza, a nivel acústico, entre los modelos de fonemas y las modelos utilizados para el reconocimiento.

Los modelos de fonemas de relleno han demostrado un mejor comportamiento que otras soluciones independientes del vocabulario, como serían modelos de relleno de palabras, o anti-modelos [11]. Otras soluciones más complejas presentes en la literatura, como modelos de transformación de parámetros, o lattice-based combination models [6], superan en eficiencia estos métodos, aunque al coste de ser más dependientes de la tarea y el entorno. Nuestro objetivo, en cambio, es definir una solución de verificación de baja complejidad e independiente de la aplicación.

4.3. Resultados de verificación

Comencemos por definir [7] *TRR (True Rejection Rate)*, o tasa de rechazos correctos, como la relación entre el número de hipótesis incorrectas que han sido detectadas por la verificación, y el número total de hipótesis incorrectas: $TRR = D/I$. De forma similar, *FRR (False Rejection Rate)*, o tasa de falsos rechazos, es la relación entre el número de hipótesis correctas rechazadas y el número total de hipótesis correctas: $FRR = R/C$.

A partir de estos conceptos, se define el *CER (Classification Error Rate)*, o tasa de errores de clasificación, como el porcentaje de hipótesis incorrectamente clasificadas por el sistema de verificación respecto al número total de hipótesis reconocidas (ver [7, 4]). Se expresa en función de los dos tipos de errores que se pueden cometer: clasificar una frase correctamente reconocida como incorrecta (*R*), y clasificar una hipótesis incorrectamente reconocida como correcta (*I - D*).

$$CER = \frac{R + (I - D)}{I + C} \cdot 100$$

Los resultados obtenidos en verificación utilizando DigitVox se resumen en la tabla 2. La relación entre los parámetros C, D, I y R, y las etiquetas utilizadas por nuestro sistema de verificación es la siguiente: R = *rechazadas*, C = *exactas* + *R*, D = *detectadas* (o D = *detectadas* + *garbage* si se implementa filtrado a nivel de frases), y I = *errores* + D. Podemos observar como la solución basada en una segunda opinión utilizando *OPW* supera claramente los resultados de referencia, tanto considerando valores absolutos (tasa de reconocimientos exactos) como medidas relativas (valores TRR y FRR).

En términos de los parámetros *TRR* y *FRR*, se define la curva ROC (Receiver Operating Characteristic) [12] como la representación gráfica de la relación entre TRR y FRR para cada umbral utilizado, con FRR en el eje x. La

figura 1 muestra la representación de la ROC para nuestro sistema de verificación y para el sistema de referencia. Se observa de nuevo la mejora introducida por el sistema basado en una segunda opinión, demostrándose el interés de la propuesta para su utilización en tecnologías de reconocimiento de voz.

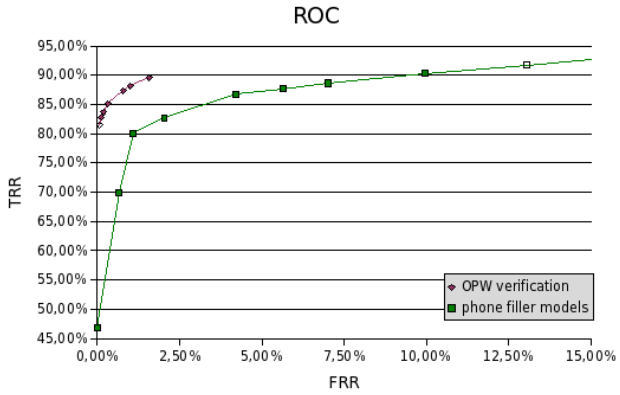

Figura 1. TRR vs. FRR

5. CONCLUSIONES

La mayoría de implementaciones estándar de HMMs para reconocimiento del habla utilizan un factor de ponderado de la gramática (o *LW*) para equilibrar la diferencia de rangos dinámicos entre las probabilidades acústicas y del modelo de lenguaje. En este artículo se propone la implementación de un factor de ponderado de las probabilidades de observación (*OPW*) con el objetivo de suavizar la diferencia en orden de magnitud entre las probabilidades de observación y de transición entre símbolos de los HMM. Los valores óptimos de *OPW* para reconocimiento del habla se han determinado de forma empírica como $OPW < 1$, de forma que se mejoran los resultados de reconocimiento de referencia obtenidos mediante HMM normales (sin ponderar).

A partir de la hipótesis de que las palabras incorrectamente reconocidas serían más sensibles a las variaciones de *OPW* (como ocurre con *LW*) se propone la implementación de un sistema de verificación de locuciones utilizando un sistema basado en una segunda opinión. Así, la etapa de verificación consistirá en comparar dos decodificaciones: una de referencia, resultado del reconocimiento con modelos no ponderados, y la segunda obtenida utilizando los HMM ponderados, considerando diferentes $OPW \neq 1$.

Observando los valores TRR y FRR de los resultados obtenidos, así como su representación gráfica mediante la curva ROC, se observa una importante mejora de nuestra propuesta en comparación con los resultados obtenidos utilizando modelos basados en fonemas de relleno. Podemos concluir, pues, que el sistema de reconocimiento y verificación propuesto se perfila como una solución interesante para tecnologías basadas en el reconocimiento del habla.

6. BIBLIOGRAFÍA

[1] X. Huang, A. Acero and H.W. Hon, *Spoken Language Processing*, Prentice Hall PTR, 1st edition, 2001.

[2] A. Bonafonte et alter, "RAMSES: el sistema de reconocimiento del habla continua y gran vocabulario desarrollado por la UPC," *VIII Jornadas de Telecom I+D*, 1998.

[3] A. Moreno, R. Winksky, "Spanish fixed network speech corpus," *SpeechDat Project. LRE-63314*, 1999.

[4] F. Wessel, R. Schlüter, K. Macherey and H. Ney, "Confidence measures for large vocabulary continuous speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 3, pp. 288–298, March, 2001.

[5] L. Chase, "Word and acoustic confidence annotation for large vocabulary speech recognition," *Proceedings of European Conf. on Speech Technology (EUROSPEECH)*, pp. 815–818, 1997.

[6] T. Schaaf and T. Kemps, "Confidence measures for spontaneous speech recognition," *Proceedings of IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. II, pp. 875–878, 1997.

[7] A. Sanchis, "Tesis doctoral. estimación y aplicación de medidas de confianza en reconocimiento automático del habla," *Universidad Politécnica de Valencia. Departamento de Sistemas Informáticos y Computación*, 2004.

[8] D. Willett, A. Worm, C. Neukirchen and G. Rigoll, "Confidence measures for hmm-based speech recognition," *Proceedings of IEEE Int. Conf. on Spoken Language Processing (ICSLP)*, 1998.

[9] G. Hernández-Ábrego and J. B. Mariño, "A second opinion approach for speech recognition verification," *Proceedings of the VIII SNRFAI*, vol. I, pp. 85–92, 1999.

[10] S. R. Young, "Detecting misrecognitions and out-of-vocabulary words," *Proceedings of IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. I, pp. 21–24, 1994.

[11] L. Jiang and X.D. Huang, "Vocabulary-independent word confidence measure using subword features," *Proceedings of IEEE Int. Conf. on Spoken Language Processing (ICSLP)*, 1998.

[12] J.P. Egan, *Signal detection theory and* ROC *analysis*, Academic Press, 1975.