Procesamiento del Lenguaje Natural, núm. 35 (2005), pp. 221-227 recibido 29-04-2005; aceptado 01-06-2005

Verificación de tema en sistemas de diálogo mediante la aplicación de un test de hipótesis bayesiano

David Pérez-Piñar López
Universidad de Vigo
Campus Universitario - Vigo
dperez@gts.tsc.uvigo.es

Carmen García Mateo
Universidad de Vigo
Campus Universitario - Vigo
carmen@gts.tsc.uvigo.es

Resumen: Uno de los aspectos más complejos de cualquier sistema de diálogo consiste en identificar la intención del usuario y en distinguir los errores de reconocimiento. En este artículo presentamos una nueva aproximación a la tarea de decidir qué ha dicho el usuario y qué quiere hacer en cada momento del diálogo. Para ello empleamos clasificadores estadísticos dirigidos por medidas de confianza generadas en paralelo por varios reconocedores adaptados a los diferentes temas del diálogo. Esta aproximación se ha mostrado especialmente adecuada para temas difíciles, como los nombres propios o las confirmaciones. La arquitectura utilizada mejora sustancialmente la tasa de reconocimiento y permite identificar la intención del usuario y detectar los cambios de tema.
Palabras clave: Sistemas de diálogo, medidas de confianza, verificación de tema, detección de intención, reconocimiento de voz.

Abstract: One of the most difficult aspects of any dialogue system is the identification of user intention and recognition errors. In this paper, we present a novel approach to the task of deciding what the user has said and what she wants to do next. We use statistic classifiers driven by confidence measures which are generated in parallel by several topic-adapted speech recognizers. This approach has shown to be especially suited for difficult topics, such as proper names or confirmations. Recognition performance is greatly enhanced through the use of this architecture, which helps also in the identification of user intention and user-initiated topic change.
Keywords: Dialogue systems, confidence measures, topic verification, intention detection, speech recognition.

1 Introducción

Los sistemas de diálogo han sido y siguen siendo objeto de un profundo estudio. Su implementación en aplicaciones reales se ha extendido a campos muy variados, y muchos de ellos emplean la voz como único canal de comunicación. Esta limitación los hace más susceptibles de error, y obliga a implementar técnicas de recuperación que resuelvan los problemas del diálogo sin reducir la satisfacción del usuario.

Estas técnicas, sin embargo, son muy complicadas. En muchos casos, los errores se producen por una interpretación incorrecta de la locución del usuario, que obliga al gestor de diálogo a llevar la conversación en una dirección equivocada. En este nuevo estado, el usuario se da cuenta del error con facilidad y responde en consecuencia. El sistema, sin embargo, debe iniciar un proceso de recuperación muy complicado, muy susceptible de errores y, en muchos casos, muy tedioso para el usuario.

Por ello, la verificación del tema es un elemento muy importante para resolver estas situaciones (Carberry 2001). Se trata de dotar al sistema de la capacidad para garantizar con cierta fiabilidad que el tema expresado por el usuario es el que se espera, o para detectar el cambio de tema si éste se produce.

ISSN: 1135-5948

En este artículo presentamos una variante a la arquitectura tradicional de reconocimiento de los sistemas de diálogo que dota al sistema de esa capacidad. La figura 1 muestra un esquema de esta arquitectura. El sistema se basa en dos elementos previos al gestor de diálogo: un módulo de reconocimiento en paralelo y un módulo de detección de tema y de cambio de intención.

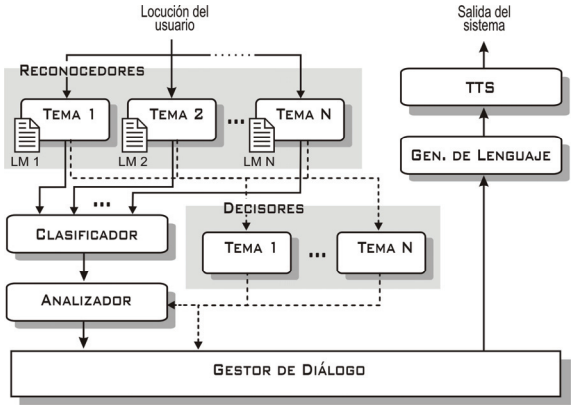

Figura 1: Sistema de diálogo con reconocedores en paralelo y detección de tema

El módulo de reconocimiento está formado por reconocedores en paralelo adaptados a cada uno de los temas que constituyen la aplicación del sistema. Cada uno se especializa en el reconocimiento de un tema mediante el uso de mezclas de modelos de lenguaje.

Los reconocedores adaptados son necesarios para generar las medidas de confianza que se utilizarán en el módulo de detección de intención (Cox y Dasmahapatra, 2002). Éste contiene un conjunto de decisores, también adaptados a cada tema, que reciben las medidas de confianza necesarias y generan un indicador de fiabilidad para cada tema. Este score es enviado al gestor de diálogo, que dispone así de información sobre el tema correspondiente a la transcripción recibida del módulo de reconocimiento.

Esta arquitectura permite verificar que el tema es el esperado y también detectar los cambios de tema iniciados por el usuario. Además, la información disponible es suficiente para que el gestor de diálogo pueda identificar el nuevo tema en este segundo caso. Esta capacidad es importante, porque evita la necesidad de algoritmos dialogados de recuperación complejos.

El resto del artículo está organizado del siguiente modo. A continuación se describe con detalle el mecanismo de verificación de tema y de detección de intención. La sección 3 describe con detalle el marco experimental empleado, y la sección 4 presenta de modo resumido la arquitectura de reconocimiento en paralelo. Los resultados se presentan en la sección 5, y finalmente se exponen las conclusiones.

2 Verificación de tema

El problema general de conocer la intención del usuario en un sistema de diálogo se ha resuelto tradicionalmente con diversos métodos (Carberry, 2001). Todos ellos funcionan correctamente cuando no hay errores, pero el diálogo y su gestión se complican mucho cuando se presenta una situación inesperada para el sistema.

La arquitectura que proponemos trata de aliviar el problema ofreciendo más información al gestor de diálogo y simplificando los procesos de recuperación. La idea básica consiste en obtener, a partir del módulo de reconocimiento, indicadores fiables del tema expresado por el usuario. De este modo, el sistema puede conocer con cierta seguridad el tema, detectar los cambios de tema y evitar situaciones de error originadas por defectos propios.

2.1 Test de hipótesis

El problema general de la detección de la intención del usuario puede descomponerse en dos partes: la verificación del tema y la detección del cambio de tema. Ambas pueden plantearse y analizarse como un test de hipótesis (García-Mateo et al., 1999), en el que se parte de la suposición de que el sistema espera una respuesta perteneciente a un tema prefijado. H_0 es la hipótesis de que la frase pronunciada por el usuario pertenece a dicho tema (hipótesis nula) y H_1 es la hipótesis de que no pertenece al mismo (hipótesis alternativa).

La implementación de este mecanismo de verificación se realiza mediante un grupo de clasificadores adaptados a cada tema. Cada clasificador recibe los parámetros de los reconocedores adaptados, y en función de ellos decide si la frase reconocida pertenece al tema o no.

En nuestro estudio hemos empleado dos aproximaciones para evaluar la hipótesis alternativa (H_1). La primera consiste en obtener H_1 a partir de un modelo universal, evaluado

mediante las medidas de confianza generadas por el reconocedor adaptado al tema que se está verificando. Este decisor lo hemos denominado *"clasificador de verificación"*.

La segunda aproximación utiliza antimodelos, es decir, obtiene un *score* a partir de un conjunto de modelos de los demás temas, excluyendo el tema que se trata de verificar. El *score* se evalúa a partir de las medidas de confianza de los reconocedores alternativos (todos excepto el del tema supuesto). Así obtenemos lo que denominamos *"decisor de contraste"* para cada tema.

Los dos decisores trabajan como verificadores. Cada uno genera un indicador de fiabilidad del tema correspondiente a la hipótesis nula que se evalúa mediante curvas DET donde se representa la tasa de falsas alarmas (verificación positiva de frases de otros temas), frente a la tasa de falsos rechazos (verificación negativa de frases del tema). Un punto de trabajo posible es aquel que proporciona una tasa de falsas alarmas igual a la tasa de falsos rechazos. Este punto se denomina EER ("Equal Error Rate").

Esta arquitectura requiere la implementación de un módulo de reconocimiento modificado, que se detalla en el apartado 4. Está formado por un conjunto de reconocedores adaptados a cada tema, que generan los parámetros necesarios para obtener las medidas de confianza empleadas en la clasificación. Un post-clasificador selecciona la transcripción de salida más probable. En la figura 2 se muestra el esquema de los decisores de verificación y contraste para un tema concreto.

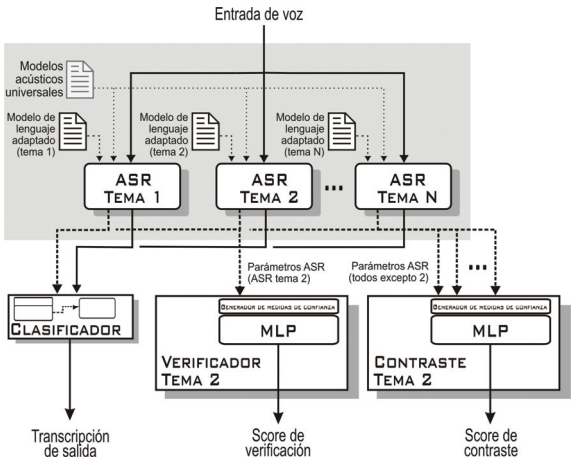

Figura 2: Módulo de detección de intención (decisores correspondientes a un tema)

Cada uno de los temas restantes tiene un esquema similar. La relativa complejidad de esta arquitectura de decisión se compensa con la información adicional que se ofrece al gestor de diálogo y con la simplificación del módulo de corrección de errores.

2.2 Implementación de los decisores

Los decisores de verificación y contraste son clasificadores implementados con perceptrones multicapa (MLP). La red neuronal tiene 3 entradas (9 para el decisor de contraste) que reciben las medidas de confianza, 5 neuronas de tipo sigmoide tangente hiperbólica en la capa oculta (6 en el de contraste) y una salida que ofrece el *score* del tema supuesto.

El entrenamiento de la red se realiza con la partición de validación extraída de SpeechDAT, como se indica en el apartado 3.2. En este proceso se han eliminado algunas entradas del corpus que presentaban valores discordantes (outlayers) en la verosimilitud acústica.

Por otro lado, se implementa un algoritmo genético sencillo (Khare y Yao, 2002), que optimiza los pesos y el número de los elementos ocultos de la red.. En nuestro entorno experimental, se ha obtenido una reducción media del error del 3.8%.

3 Marco experimental

3.1 Sistema de reconocimiento

En el experimento empleamos un reconocedor de habla continua y grandes vocabularios basado en HMM continuos (CHMM). El motor de reconocimiento funciona en dos fases: un algoritmo de Viterbi síncrono con búsqueda en haz y un algoritmo A* (Diéguez-Tirado et al., 2005).

Los modelos acústicos se generan a partir de las bases de datos SpeechDAT en gallego y castellano. Se trata de un corpus de voz telefónica, muestreada a 8 KHz y codificada mediante ley A con 8 bits por muestra. El entrenamiento de los modelos emplea 15 horas en gallego y 25 horas en castellano, y se obtienen 627 unidades acústicas, que son demifonemas formados HMMs de dos estados. Cada estado se modela mediante una mezcla de entre 4 y 8 distribuciones gaussianas en un espacio de 39 dimensiones: 12 coeficientes MFCC, la energía normalizada y sus derivadas primera y segunda.

Los modelos de lenguaje se basan en trigramas, y se entrenan mediante las

herramientas SRILM (Stolcke, 2002) con un suavizado Katz.

3.2 Bases de datos de pruebas

Todas las fases del experimento utilizan un subconjunto de la base de datos SpeechDAT en castellano (Moreno, 1997). El corpus está formado por 5000 llamadas telefónicas de diferentes usuarios, y cada llamada incluye distintos temas: números, fechas, habla continua, etc.

No se trata de una base de datos de diálogo: cada llamada consiste en una batería de preguntas fijas con respuestas variables. Sin embargo, el alcance del experimento no requiere diálogos anotados. Las pruebas se realizan suponiendo un tema activo y alimentando el sistema con frases de todos los temas. En este sentido, SpeechDAT proporciona suficientes temas para trabajar y es un banco de pruebas adecuado a la tarea.

La aplicación se define como un conjunto de temas, y para ello se seleccionaron los siguientes de la base de datos:

- **Fechas**: Incluye locuciones que expresan fechas, tanto con formatos fijos (*tres de diciembre de dos mil dos*) como con expresiones coloquiales (*el próximo martes*).
- **Nombres**: Nombres propios, con nombre y apellidos.
- **Números**: Incluye diversos conceptos (números telefónicos, tarjetas de crédito y otros), expresados de modo natural (*treinta y seis mil doscientos*) o con dígitos aislados (*tres seis dos cero cero*).
- **Confirmaciones**: Frases de confirmación o rechazo de cierta información.

Estos temas tienen características lingüísticas y acústicas distintas, pero algunos comparten el vocabulario y ciertas expresiones. Se han seleccionado por su amplio uso en sistemas reales y por la importancia de las confirmaciones en la gestión del diálogo.

La aplicación queda, por tanto, definida por estos cuatro temas. El gestor de diálogo será capaz de manejarlos, pero cualquier otro tema no estará soportado por el sistema.

Las transcripciones ortográficas de las locuciones de cada tema se dividen en tres particiones: entrenamiento, validación y test. El subconjunto de SpeechDAT empleado contiene un total de 991 personas, 479 hombres y 512 mujeres, que son asignados a diferentes particiones.

La tabla 1 muestra la distribución de las transcripciones para cada tema y cada partición. La partición de entrenamiento es utilizada para generar los modelos de lenguaje, contiene el 75% de las transcripciones y excluye todas aquellas locuciones que presentan errores de pronunciación, palabras incompletas o ruido no estacionario (chasquidos, golpes, etc.) La partición de validación está formada por el 12.5% de las transcripciones, y se emplea para el entrenamiento del clasificador. Finalmente, la partición de prueba, con las transcripciones restantes, se usa para evaluar el reconocimiento y la detección de la intención.

Tema	Entrenamiento	Validación	Test
Fechas	2248	375	375
Nombre	1494	249	249
Números	3745	625	625
Conf.	1478	247	247

Tabla 1: Número de ficheros de voz y transcripciones para cada tema y partición

4 Reconocimiento de voz en paralelo

Como se ha indicado, el sistema requiere un módulo de reconocimiento que genere medidas de confianza significativas respecto al tema. Por ello, en nuestra propuesta dividimos este módulo en varios reconocedores adaptados a cada tema que funcionan en paralelo sobre la misma entrada de voz, como se muestra en la figura 3.

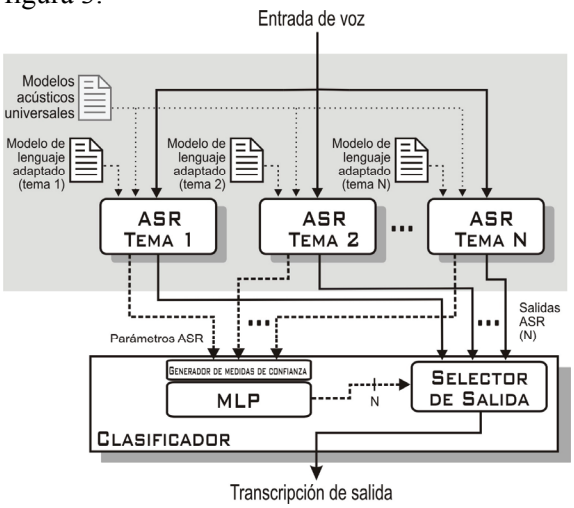

Figura 3: Módulo de reconocimiento

La salida correcta es seleccionada por un decisor a partir de las medidas de confianza. La adaptación se consigue mediante modelos de lenguaje y clasificación estadística.

4.1 Modelos de lenguaje adaptados al tema

Los reconocedores adaptados utilizan modelos de lenguaje generados a partir de las transcripciones ortográficas de SpeechDAT. La adaptación a cada tema se consigue mezclando los modelos individuales basados en n-gramas (Stolcke, 2002). Los vocabularios son de 115 palabras para las fechas, 581 para los nombres, 99 para los números y 66 para las confirmaciones. Estos modelos se mezclan con un modelo universal, entrenado a partir de texto periodístico con un vocabulario de unas 20000 palabras.

Por simplicidad, la mezcla realizada emplea pesos fijos en todos los temas: un 15% para el modelo del tema y un 85% para el modelo universal. Esta relación podría optimizarse minimizando la perplejidad del modelo de lenguaje (Diéguez-Tirado et al., 2005).

4.2 Medidas de confianza

Las medidas de confianza empleadas en el reconocimiento y la verificación de tema se obtienen a partir de las verosimilitudes acústicas, las probabilidades lingüísticas y la transcripción decodificada (García-Mateo et al., 1999). De estas características se derivan tres indicadores:

- Puntuación Acústica Normalizada de la Frase (NSAS). La verosimilitud acústica de las palabras de la frase se suman y se normalizan respecto al número de palabras reconocidas (NRW).
- Puntuación Lingüística Normalizada de la Frase (NSLS). Las probabilidades del modelo de lenguaje de las palabras se suman y normalizan del mismo modo que en el caso anterior.
- Número de Palabras Reconocidas (NRW). Es el número de palabras transcritas en la salida del reconocedor.

La relación de estas medidas de confianza con la corrección del reconocimiento se ha valorado calculando su correlación con los resultados de evaluación del reconocedor. Los resultados indican que las distribuciones son diferentes. Por tanto, las clases son separables, y un clasificador estadístico puede distinguirlas.

4.3 Clasificador de reconocimiento

Las medidas de confianza generadas por cada reconocedor se envían a un clasificador estadístico que selecciona la salida más fiable.

La arquitectura, entrenamiento y optimización de este clasificador son muy similares a los correspondientes a los decisores de detección de intención. En este caso, el clasificador tiene 12 entradas (tres por cada reconocedor) que reciben las medidas de confianza, 12 neuronas de tipo sigmoide tangente hiperbólica en la capa oculta y cuatro salidas correspondientes cada una a la detección de un tema.

Por otra parte, la reducción de la tasa de error de reconocimiento mediante optimización de la red es del 4.2%.

5 Resultados experimentales

5.1 Prestaciones del reconocedor

La evaluación del sistema se realiza mediante dos experimentos. El primero, previo a la definición y entrenamiento de los identificadores de tema, consistió en una evaluación del módulo de reconocimiento en paralelo, realizada con las particiones de test definidas previamente. La tabla 2-a muestra las tasas de reconocimiento (*correction*) de cada tema (columnas) usando cada reconocedor adaptado (filas). La tabla 2-b enumera la mejora porcentual respecto al reconocimiento con el modelo de lenguaje universal.

LM	Conf.	Fechas	Nombres	Números
universal	40,0	75,1	41,6	79,3
aplicación	57,5	88,2	82,1	90,7
confirmación	**79,6**	60,7	41,9	73,9
fechas	30,4	**88,1**	40,7	78,4
nombres	32,9	57,7	**82,9**	67,4
números	26,4	71,6	41,6	**90,9**

Tabla 2-a: Tasa de reconocimiento (%)

LM	Conf.	Fechas	Nombres	Números
aplicación	17,5	13,1	40,4	11,4
confirmación	**39,6**	-14,3	0,2	-5,4
fechas	-9,6	**12,9**	-0,8	-0,9
nombres	-7,1	-17,3	**41,3**	-11,8
números	-13,5	-3,4	0,0	**11,6**

Tabla 2-b: Mejora del reconocimiento (%)

Finalmente, la tabla 3 muestra los resultados globales de reconocimiento en forma de matriz de confusión, con los temas de prueba (las entradas) como filas y los temas escogidos (las transcripciones reconocidas) como columnas.

Tema	Conf.	Fechas	Nombres	Números
confirmación	**74,1**	10,7	1,5	13,7
fechas	4,9	**87,7**	5,7	1,7
nombres	0,7	6,1	**89,8**	3,4
números	8,5	6,7	2,5	**82,3**

Tabla 3: Matriz de confusión del reconocedor

Los resultados muestran una gran mejora en la corrección del reconocimiento para los temas individuales. El módulo de reconocimiento seleccionará el tema correcto con un error global del 16.45% sin contar con medidas de confianza de alto nivel. Comparando estos resultados con nuestro sistema de referencia, se observa que el reconocimiento mejora sensiblemente para las confirmaciones y los nombres.

5.2 Verificación de tema

El segundo experimento trata de evaluar la capacidad del módulo de detección de intención para verificar el tema. En este proceso se define un tema activo como hipótesis nula, y se emplean las particiones de test para enviar al módulo de reconocimiento frases de todos los temas posibles. La figura 4 muestra los resultados para los decisores de verificación y contraste adaptados a números.

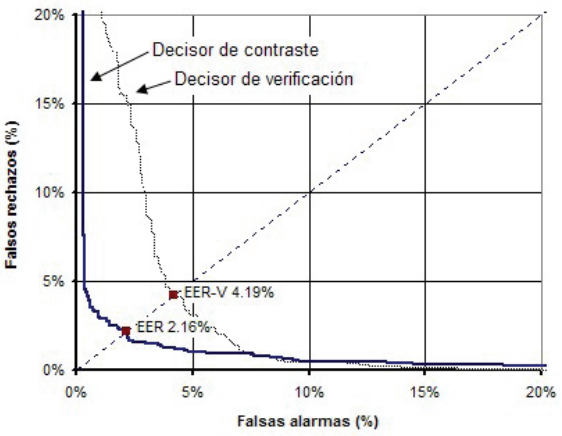

Figura 4: Curvas DET de los decisores adaptados a números.

En este caso, el decisor de verificación recibe las medidas de confianza del reconocedor adaptado a los números, y el decisor de contraste recibe las medidas de los demás reconocedores adaptados. La evaluación se realiza para cada tema por separado, y en cada caso se genera una curva DET para evaluar el punto de trabajo EER. Los resultados muestran claramente la capacidad de ambos decisores para detectar el tema. El rendimiento es similar en los demás temas; puede observarse que el decisor de contraste es ligeramente mejor en casi todos, como se muestra en la tabla 4.

Tema	EER (%)		Corrección (%)	
	Verif.	Cont.	Verif.	Cont.
Conf.	2.60	2.79	94.86	94.43
Fechas	3.45	4.09	93.11	91.83
Nombres	5.51	2.78	89.04	94.45
Números	4.19	2.16	91.62	95.69

Tabla 4: ERR y tasas de corrección (%) para los decisores de verificación y contraste.

La capacidad del sistema para verificar el tema es patente con estos datos. En todos los casos, las tasas de detección del tema son suficientes, y las falsas alarmas y falsos rechazos son muy reducidos.

El esquema puede ser empleado para detectar cambios de intención del usuario situando un verificador de cada uno de los temas posibles.

6 Conclusiones

La nueva aproximación al problema de la detección de la intención ha mostrado un rendimiento muy aceptable. La arquitectura empleada no utiliza gramáticas de estados finitos. Los métodos estadísticos que las sustituyen ofrecen una mejora notable en las tasas de reconocimiento y de detección de tema, sobre todo para algunos de particular importancia en los sistemas de diálogo: las confirmaciones juegan un papel esencial en los mecanismos de detección y corrección de errores del diálogo. Además, la arquitectura del sistema permite la detección del cambio de tema y la identificación del nuevo tema.

Las líneas de trabajo a partir de este sistema son muy numerosas. El reconocimiento puede mejorarse optimizando los pesos de la mezcla de los modelos de lenguaje. Por otra parte, entre las posibles mejoras del mecanismo de

verificación del tema cabe destacar dos: la combinación de los decisores de verificación y contraste y la introducción de medidas de confianza de alto nivel en los decisores.

Bibliografía

Chase, L.. 1997. Error-responsive feedback mechanisms for speech recognizers, PhD thesis, School of Computer Science, Carnegie Mellon University.

Carberry, S. 2001. Toward a Robust Dialogue System: Recognizing Dialogue Acts. Proc. of Pacific Association for Computational Linguistics (PACLING), Kitakyushu, Japan.

Cox, S. and Dasmahapatra, S. 2002. High-level Approaches to Confidence Estimation in Speech Recognition. IEEE Trans. on Speech and Audio, 7:10, páginas 460-471.

Dieguez-Tirado, J., García-Mateo, C. et al., Adaptation strategies for the acoustic and language models in bilingual speech transcription. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 833-836. March 2005

García-Mateo, C., Reichl, W., Ortmanns, S., 1999. On Combining Confidence Measures in HMM-based Speech Recognizers. Intern. Workshop on Automatic Speech Recognition and Understanding, ASRU99, Keystone, CO, USA, December 12–15, 1999.

Khare, V. y Yao, X. 2002. Artificial Speciation of Neural Network Ensembles. Proc. of the UK Workshop on Computational Intelligence (UKCI'02), pp. 96-103, Birmingham, UK.

Moreno, A. 1997. SpeechDAT Spanish Database for Fixed Telephone Networks. Corpus Design Technical Report, SpeechDAT Project LE2-4001.

San-Segundo, R., Pellom, B. et al. 2001. Confidence Measures for Spoken Dialogue Systems. Proc IEEE ICASSP, Pp 393-396, ISBN 0-7803-7041-4, Salt Lake City.

Stolcke, A. 2002. SRILM – An extensible language modelling toolkit. Proc. Int. Conf. Spoken Language Processing, vol. 2, pp. 901–904 Denver, CO.