Una Nueva Técnica para Evaluar Sistemas Conversacionales Basada en la Generación Automática de Diálogos

R. López-Cózar, J. C. Segura, A. De la Torre, A. J. Rubio
Dpto. Electrónica y Tecnología de Computadores, 18071 Universidad de Granada
Tel.: +34 958243271, FAX: +34 958243230 E-mail: {rlopezc,segura,atv,rubio}@ugr.es

Resumen

En este artículo se presenta una nueva técnica que permite evaluar sistemas conversacionales mediante la simulación de las interacciones que realizan los usuarios de los mismos. Creemos que la técnica puede ser útil para conocer aspectos de los sistemas conversacionales que se deben mejorar antes de que se encuentren operativos en condiciones reales. Utilizando esta técnica, se ha evaluado el funcionamiento de un sistema conversacional desarrollado en nuestro laboratorio, el cual se encuentra en fase experimental actualmente. La evaluación se ha realizado considerando diversas condiciones de ruido. Los resultados obtenidos muestran que el uso de una técnica de compensación de ruido mejora el funcionamiento del sistema. La técnica propuesta ha permitido detectar errores de la estrategia utilizada por el sistema para procesar las confirmaciones de los usuarios.

1. Introducción

Los sistemas conversacionales (también conocidos como *sistemas de diálogo*) son unos sistemas basados en el procesamiento de la voz humana cuya finalidad es proporcionar diversos tipos de servicios a los usuarios mediante el habla [1], [2], [3]. Estos sistemas utilizan principalmente las tecnologías de reconocimiento de voz, comprensión del lenguaje natural, gestión del diálogo y generación de voz. El incremento en el uso comercial y en la sofisticación de estos sistemas ha originado la necesidad de desarrollar nuevas técnicas que faciliten el desarrollo y la evaluación de los mismos. Por ejemplo, una técnica muy utilizada durante las fases iniciales de desarrollo es la denominada *Mago de Oz* (*Wizard of Oz*) [4]. Cuando se usa esta técnica, los diseñadores hacen creer a usuarios de test que están hablando con sistemas inteligentes automáticos, pero realmente son operadores humanos quienes deciden las respuestas de los sistemas y dirigen la conversación en cada momento. Esta técnica es útil pues permite evaluar desde el principio el funcionamiento de los sistemas y probar nuevas ideas [5].

2. Generación automática de diálogos

En la bibliografía pueden encontrarse diversos trabajos que hacen referencia a técnicas de simulación para desarrollar sistemas basados en el procesamiento de la voz. Por ejemplo, en [6] se presenta una técnica que permite evaluar algoritmos de reconocimiento de voz usando simulaciones de datos de voz. En el trabajo de referencia, los datos se obtienen a partir de modelos acústicos y frases arbitrarias en formato de texto. Por otra parte, en [7] se presenta una técnica de simulación para aprender estrategias de diálogo. Esta técnica se basa en el uso de un modelo estocástico que recibe el estado del diálogo y el *prompt* (petición de información) del sistema conversacional, y produce una representación semántica idéntica a la que podría provenir de una frase generada por un usuario del sistema.

La nueva técnica de evaluación propuesta en este artículo también está basada en una técnica de simulación. Concretamente, está basada en el uso de un segundo sistema conversacional denominado *simulador*, cuya finalidad es comportarse de forma similar a como lo harían los usuarios del sistema conversacional que se desea evaluar. La idea consiste en utilizar el simulador para generar automáticamente diálogos que permitan evaluar el comportamiento del sistema. De esta forma, las conversaciones se pueden generar a partir de la interacción entre el sistema y el simulador, sin que sea necesaria la interacción de los usuarios reales del sistema.

La técnica tiene la ventaja de permitir la evaluación del sistema bajo diversas condiciones, las cuales pueden ser generadas fácilmente en laboratorio variando el corpus de frases utilizado para llevar a cabo la simulación. Por ejemplo, se puede grabar un corpus de frases en condiciones limpias, y posteriormente se

pueden añadir artificialmente diferentes tipos y niveles de ruido para comprobar el funcionamiento del sistema en condiciones de ruido mediante la simulación. La técnica propuesta es útil para encontrar los puntos débiles del sistema, pudiendo realizar mejoras antes su implementación en condiciones reales.

Para implementar la técnica es necesario construir dos corpora. Por una parte, un corpus de escenarios representativo de posibles objetivos de los usuarios del sistema. Por otra parte, un corpus de frases grabadas por diversos locutores representativas de los posibles objetivos considerados en los escenarios [8]. A partir de ambos corpora, el simulador puede generar los diálogos automáticamente. Para ello, debe tratar de lograr los objetivos fijados en los escenarios durante la interacción con el sistema, usando las frases grabadas previamente por los locutores. Durante la interacción, el simulador debe analizar las respuestas del sistema conversacional y responderlas apropiadamente, como si se tratara de un usuario del sistema.

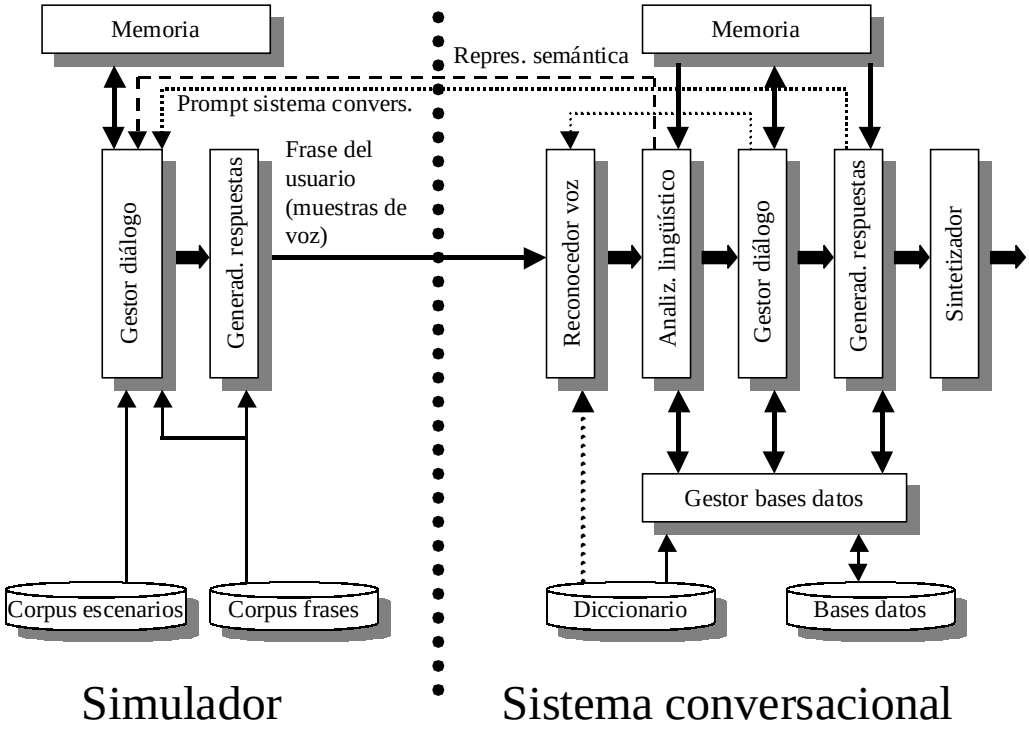

Figura 1. Interconexión entre un simulador y un sistema conversacional

2.1 Funcionamiento del simulador

Los usuarios de los sistemas conversacionales suelen interactuar teniendo en cuenta los *prompts* de los sistemas y la información contextual. A fin de comprender las frases, estos sistemas suelen crear representaciones semánticas (típicamente *frames* [9]) que extraen la información relevante de las frases de los usuarios que reciben como entradas. Por tanto, una forma de lograr que el simulador pueda generar coherentemente las respuestas en el diálogo consiste en usar los *prompts* del sistema conversacional con el que debe interactuar así como las representaciones semánticas que éste crea a partir de las frases, como muestra la Figura 1. Los *prompts* se pueden usar para conocer la información que el sistema conversacional solicita en cada momento, y las representaciones semánticas se pueden utilizar para determinar si el sistema conversacional comprendió correctamente o no la anterior frase que el simulador generó a modo de respuesta. Por consiguiente, a la hora de crear un simulador para un sistema conversacional, es necesario conocer todos los posibles tipos de *prompts* del sistema, y definir una acción del simulador que le permita generar una respuesta apropiada para cada uno de ellos.

Las acciones a definir dependen del dominio de aplicación del sistema. Por ejemplo, un simulador en el dominio ATIS (*Air Travel Information Service*) debe ser capaz de realizar reservas de vuelos. No obstante, existen dos tipos de acciones que son comunes a todos los

dominios de aplicación: la *recuperación de errores* y la *confirmación*. La corrección de errores tiene como finalidad intentar corregir los errores cometidos por el sistema conversacional, generalmente derivados de la inserción, sustitución o borrado de palabras durante la fase de reconocimiento de la voz, o bien, derivados de errores de comprensión de las frases. Ambas acciones se pueden implementar haciendo uso de las representaciones semánticas (o de las transcripciones fonéticas) creadas por el sistema conversacional tras el análisis de las frases que recibe como entradas. Así, si tras el análisis de una determinada frase la representación semántica (o transcripción fonética) obtenida no coincide con la representación semántica (o transcripción fonética) correcta, se puede concluir que la frase no ha sido correctamente procesada por el sistema conversacional. En este caso, el simulador debe indicar al sistema que se ha producido un error, a fin de iniciar un *subdiálogo* encaminado a corregirlo. Asimismo, en caso de ser requerido por el sistema conversacional, el simulador puede generar una confirmación positiva o negativa basada en la comparación de las representaciones semánticas (o transcripciones fonéticas) obtenidas con las correctas.

2.2 El corpus de frases

El corpus de frases contiene todas las frases que el simulador puede necesitar para interactuar con el sistema conversacional. Las frases deben ser grabadas por varios usuarios, a fin de verificar la independencia del locutor del reconocedor de voz del sistema. El simulador debe usar las frases de forma apropiada a fin de que el diálogo con el sistema sea coherente. Para ello, proponemos crear una transcripción fonética y una representación semántica para cada frase del corpus. De esta forma, las transcripciones fonéticas se pueden utilizar para generar trazas (ficheros *log*, por ejemplo) que almacenen los diálogos generados automáticamente, y las representaciones semánticas se pueden usar para realizar la selección de las frases en base a su significado. Como diferentes frases pueden tener el mismo significado, esta metodología permite generar los diálogos de una forma flexible, ya que en un momento dado de la conversación, se podrá usar cualquiera de las frases que comparten el significado deseado.

2.3 El corpus de escenarios

El corpus de escenarios es un conjunto de escenarios que representan posibles objetivos de los usuarios del sistema conversacional que se desea evaluar. Cada escenario representa los objetivos de un hipotético usuario interactuando con el sistema conversacional. Los objetivos se pueden representar de diversas formas en los escenarios. Por ejemplo, se pueden utilizar transcripciones fonéticas, representaciones semánticas, etc. A modo de ejemplo, mostramos en la Figura 2 uno de los escenarios utilizados durante los experimentos, diseñado para realizar el pedido de comida rápida a domicilio. En este escenario los objetivos se han representado mediante representaciones semánticas (frames).

```
Cantidad=1                          # el usuario pide un
Comida=BOCADILLO                    # bocadillo de jamón
Contenido=JAMÓN

Cantidad=1                          # el usuario pide una
Bebida=CERVEZA                      # cerveza grande
Tamaño=GRANDE

Numero_teléfono=958275360           # el usuario dice su
                                    # número de teléfono

Código_postal=18001                 # el usuario dice su
                                    # código postal

id_dirección=CALLE                  # el usuario dice los
nombre=ANDALUCÍA                    # datos de su dirección
número_edificio=58
planta=PRIMERA
letra=E
```

Figura 2. Un ejemplo de escenario

Cuando se usa este escenario, la finalidad del simulador es intentar conseguir estos objetivos durante la interacción con el sistema conversacional, utilizando para ello las frases del corpus cuyos significados se corresponden con los presentes en el escenario. La representación de objetivos mediante representaciones semánticas permite extraer fácilmente los diversos ítems de cada objetivo. De esta forma, el simulador puede responder a los *prompts* del sistema conversacional generados para reparar posibles errores de comprensión de las frases,

posiblemente derivados de errores de reconocimiento. Por ejemplo, supongamos que un sistema conversacional que atiende pedidos de comida rápida por teléfono genera el *prompt*: *"¿Qué deseas pedir?"* en un momento dado de una conversación. A partir de este *prompt*, el simulador debe buscar un objetivo en el escenario cuya representación semántica sea "realizar un pedido" (el primer objetivo del escenario mostrado en la Figura 2 podría ser el elegido, por ejemplo). Una vez determinado el objetivo, el simulador debe seleccionar alguna frase del corpus cuya representación semántica coincida con la de dicho objetivo. Por ejemplo, alguna de las siguientes frases podría ser seleccionada: *"Un bocadillo de jamón"*, *"pues, ponme un bocadillo de jamón por favor"*, *"me gustaría uno de jamón"*, etc. Una vez seleccionada la frase, el simulador debe proporcionar al sistema conversacional el fichero de muestras de voz de la frase en cuestión, que debe ser procesado por el sistema como si proviniera de un usuario real. Dado que el objetivo considerado en este ejemplo se compone de tres ítems (Cantidad=1, Comida=BOCADILLO, Contenido=JAMÓN), el simulador puede utilizar cualquiera de estos ítems para responder a los posibles *prompts* del sistema conversacional destinados a la corrección de errores. Por ejemplo, el sistema podría generar el *prompt* de corrección: *"¿Cuántos bocadillos de jamón has dicho que quieres"?*, y el simulador podría utilizar el primer ítem para responder mediante la frase *"uno"*, pues la representación semántica de esta frase coincide con la del ítem en cuestión. Por consiguiente, además de grabar previamente al menos una frase correspondiente a cada objetivo de los escenarios, es necesario también grabar al menos una frase por cada ítem existente en los objetivos. Claramente, para cada ítem se debe crear además una transcripción fonética y una representación semántica, a fin de poder utilizar el ítem durante la generación automática de los diálogos.

3. Experimentos

Se ha aplicado la técnica propuesta en este artículo para realizar la evaluación del sistema SAPLEN [10], cuya finalidad es atender telefónicamente a los clientes de restaurantes de comida rápida. La técnica ha permitido obtener una estimación del funcionamiento del sistema en condiciones de ruido blanco y *babble* (conversaciones de fondo). El sistema SAPLEN utiliza un reconocedor de voz continua independiente del locutor basado en HTK [11], que usa modelos acústicos entrenados a partir del corpus fonético de la base de datos Albayzín, constituida por 4767 frases [12]. La siguiente tabla muestra los parámetros de reconocimiento utilizados.

Tipo de ruido	SNR (dB)	Entorno de reconocimiento
Blanco	30,2	MFCC
Babble	26,5	MFCC + VTS
	21,4	
	15,6	

Tabla 1. Condiciones de reconocimiento

A fin de evaluar el funcionamiento del sistema en diversas condiciones de ruido, se han utilizado cuatro niveles de ruido blanco y *babble* para contaminar artificialmente las frases del corpus. Por tanto, el sistema conversacional procesaba las versiones contaminadas de las frases previamente grabadas por los diversos locutores. La Tabla 1 muestra los valores de la SNR media del corpus de frases contaminado con los cuatro niveles de ruido. La evaluación se ha realizado utilizando dos entornos de reconocimiento: uno basado en parámetros estándar MFCC (*Mel Frequency Cepstral Coefficients*) [13], y otro basado en la aplicación de la técnica VTS (*Vector Taylor Series*) de compensación de ruido [14].

Para construir el corpus de frases, se han seleccionado aleatoriamente 500 frases de un corpus de 523 diálogos previamente obtenido en un restaurante de comida rápida. Las transcripciones fonéticas de las 500 frases se han creado manualmente, y las representaciones semánticas correspondientes se han creado automáticamente, utilizando el analizador lingüístico del sistema conversacional utilizado en los experimentos. Nueve locutores han grabado varias versiones de las 500 frases seleccionadas. Cuatro locutores hablaban Castellano estándar, cuatro hablaban con acento andaluz, y un locutor era una mujer japonesa que habla Castellano. En total, se han grabado 1651 frases de forma espontánea, las cuales constituyen el corpus de frases utilizado en los experimentos. Las frases se han grabado en condiciones "limpias" de laboratorio, utilizando un ordenador PC, y usando 16 bits por muestra a 8 KHz.

Seleccionando de forma aleatoria diversos objetivos compatibles con el corpus de frases

se han creado 18 escenarios, cada uno de los cuales contiene los objetivos del usuario (de 1 a 5 pedidos de productos) y los datos de su dirección (número de teléfono, código postal y dirección). Para representar los objetivos en los escenarios se han utilizado las representaciones semánticas de las frases correspondientes. A fin de llevar a cabo la evaluación, se han generado 100 diálogos para cada uno de los escenarios y para cada nivel y tipo de ruido, aplicando y sin aplicar la técnica VTS. En total, se han generado de forma automática 2 x 4 x 2 x 100 x 18 = 28,800 diálogos.

Como métrica de evaluación, se ha utilizado el logro de objetivos (*Task Completion*) [15]. Se ha considerado que se producía el logro de los objetivos en un determinado diálogo si el simulador lograba que el sistema conversacional comprendiera todos los objetivos del escenario utilizado. En una situación real, los usuarios se comunicarían con el sistema a través del teléfono. En esta comunicación, los diálogos excesivamente largos serían rechazados por los usuarios, pues colgarían el teléfono si el sistema conversacional no funcionara satisfactoriamente. A fin de tener presente este hecho, se ha utilizado un parámetro en el simulador que permite cancelar la conversación tras un determinado número de interacciones. Durante los experimentos, los diálogos que alcanzaban las 30 interacciones se cancelaban y se consideraba que en éstos no se producía el logro de los objetivos.

La Figura 3 muestra los resultados obtenidos con respecto al porcentaje de logro de objetivos en los 18 escenarios considerados. En media, los escenarios número 2, 3, 4, 5 y 12 fueron los de menor tasa de logro de objetivos, considerando todas las condiciones de reconocimiento mostradas en la Tabla 1. Los escenarios 7, 8, 10, 17 y 18 alcanzaron la mayor tasa media de logro de objetivos. Cuando no se usa VTS, la tasa media de logro de objetivos es 31,43% para el ruido blanco y 29,64% para el ruido babble. En cambio, cuando se usa la técnica VTS, la tasa media de logro de objetivos asciende hasta el 57,35% para el ruido blanco y hasta el 54,44% para el ruido babble. Ello se debe a que la aplicación de la técnica VTS mejora el funcionamiento del reconocedor de voz, y por tanto, del sistema conversacional. Los resultados muestran que el sistema funciona mejor ante la presencia del ruido blanco que ante la presencia de ruido *babble* para niveles similares de ruido.

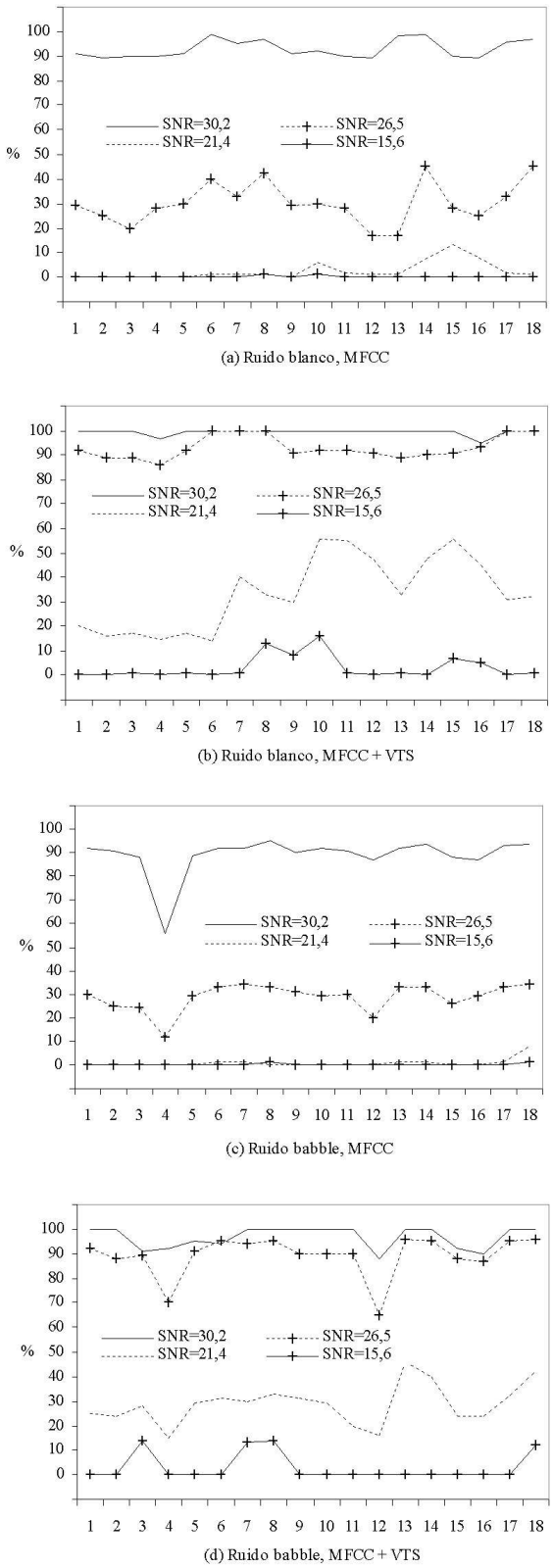

Figura 3. Logro de objetivos para los 18 escenarios considerados, bajo las diferentes condiciones de reconocimiento mostradas en la Tabla 1.

4. Análisis de los diálogos generados

El análisis de la traza de los diálogos generados automáticamente ha permitido detectar errores en el funcionamiento del sistema conversacional. Para localizar los errores, se han seleccionado aquellos diálogos con una baja tasa de logro de objetivos. El análisis de estos diálogos ha permitido comprobar que el reconocedor de voz suele cometer errores a la hora de analizar determinar palabras usadas en ciertos escenarios. Este problema es inevitable, y en condiciones reales siempre hay que tener presente la posible existencia de errores de reconocimiento. Sin embargo, la estrategia de confirmación implementada en el sistema conversacional para subsanar dichos errores no funciona correctamente en algunas ocasiones. Por ejemplo, a veces el sistema conversacional intenta realizar una confirmación, y el dato a confirmar fue reconocido incorrectamente. En este caso, el simulador responde *"no"* para generar una confirmación negativa. Sin embargo, en ocasiones la salida del reconocedor es *"sí"* en lugar de *"no"* , lo que provoca que algunos datos incorrectamente reconocidos sean considerados correctos por el sistema conversacional. En otras ocasiones, el sistema intenta confirmar un dato correctamente reconocido y el simulador responde *"sí"* para generar una confirmación positiva. Sin embargo, a veces la salida del reconocedor es *"no"* en lugar de *"sí"* , lo que provoca que el sistema tenga que utilizar más interacciones para obtener y confirmar el dato. Este hecho provoca que algunos diálogos excedan el límite máximo de interacciones, y por consiguiente, finalicen sin que se haya producido el logro de los objetivos del escenario.

5. Conclusiones y trabajo futuro

La técnica propuesta en este artículo ha permitido evaluar el funcionamiento de un sistema conversacional en diversas condiciones de ruido. Los resultados obtenidos muestran que el uso de la técnica VTS mejora el funcionamiento del sistema conversacional en condiciones de ruido. Asimismo, muestran que el sistema es más sensible al ruido *babble* que al ruido blanco.

La evaluación ha hecho posible detectar problemas en el funcionamiento del sistema conversacional, derivados principalmente del funcionamiento incorrecto de la estrategia de confirmación implementada. Estos problemas sugieren que se debe mejorar dicha estrategia. Una posible mejora podría consistir en utilizar confirmaciones redundantes. Se ha observado que el número de errores de reconocimiento aumenta cuando la SNR es baja. Por consiguiente, podría ser útil implementar una estrategia de confirmación que tenga en cuenta este hecho. Por ejemplo, cuando la SNR sea baja, el sistema podría volver a solicitar la confirmación de un cierto dato en caso de que la salida del reconocedor fuera una confirmación negativa, y podría realizar una confirmación de la confirmación previa (como por ejemplo, *"¿Has dicho que sí?"*) en caso de que la salida fuera una confirmación positiva. Aunque esta nueva estrategia implicaría la aparición de turnos adicionales del usuario, podría ayudar a evitar los errores en las confirmaciones. En un próximo trabajo implementaremos esta estrategia y mediremos su efecto en la tasa de logro de objetivos.

6. Referencias

[1] Zue V. et al., "JUPITER: A telephone-based conversational interface for weather information", IEEE Trans. on Speech and Audio Processing; pp. 85-95, 2000

[2] Rubio A.J., García P., De la Torre A., Segura J.C., Díaz-Verdejo J.E., Benítez M.C., Sánchez V., Peinado A.M., López-Soler J.M., Pérez-Córdoba J.L., "STACC: An Automatic Service for Information Access Using Continuous Speech Recognition Through Telephone Line", Eurospeech '97, pág. 1779-1782

[3] Chao H., Xu P., Zhang X., Zhao S., Huang T., Xu B., "LODESTAR: A Mandarin Spoken Dialogue System for Travel Information Retrieval", Eurospeech '99, pág. 1159-1162

[4] Bernsen N. O., Dybkjaer L., Heid U.; 1999; Current Practice in the Development and Evaluation of Spoken Language Dialogue Systems; Proc. of Eurospeech; pp. 1147-1150

[5] Rosset S., Bennacef S., Lamel L., "Design Strategies for Spoken Language Dialog Systems", Eurospeech '99, pág. 1535-1538

[6] McAllaster D., Gillick L., "Studies in Acoustic Training and Language Modeling Using Simulated Speech Data", Eurospeech '99, 1787-1790

[7] Levin E., Pieraccini R., Eckert W., "A stochastic model of human-machine interaction for learning dialog strategies", IEEE Trans. on Speech and Audio Processing, vol. 8, Enero 2000, 11-23

[8] Ziegenhain U., Harengel S., Kaiser J., Wilhem R. "Creating Large Pronunciation Lexica for Speech Applications", First International Conference on Language Resources and Evaluation, pág. 1039-1043

[9] Allen J., Natural language understanding, The Benjamin/Cummings Publishing Company Inc., 1995

[10] López-Cózar R., Rubio A. J., García P., Segura J. C., "A New Word-Confidence Threshold Technique to Enhance the Performance of Spoken Dialogue Systems", Eurospeech '99, pág. 1395-1398

[11] Hain T. Et al., "The 1998 HTK System for Transcription of Conversational Telephone Speech", ICASSP 99

[12] Casacuberta F. et al., "Development of Spahish Corpora for Speech Research (ALBAYZIN)", Workshop on International Cooperation and Standardization of Speech Databases and Speech I/O Assesment Methods, Chiavari, Italy; pp.26-28, 1991

[13] Rabiner L. R., Juang B. H., "Fundamentals of Speech Recognition", Prentice-Hall, 1993

[14] Moreno P. J.; 1996; Speech recognition in noisy environments; Ph Thesis, CMU

[15] M. Danieli, E. Gerbino, "Metrics for evaluating dialogue strategies in a spoken language system", AAAI Spring Symposium on Empirical Methods in Discourse Interpretation and Generation, 1995, pág. 34-37