USO EXPERIMENTAL DE UN UMBRAL DE CONFIANZA DINÁMICO EN EL PROCESO DE RECONOCIMIENTO DEL HABLA

R. López-Cózar, A.J. Rubio, A. de la Torre, J.M. López-Soler
Dpto. Electrónica y Tecnología de Computadores, Facultad de Ciencias. Campus Universitario de Fuentenueva, s/n, Universidad de Granada. 18071 GRANADA (SPAIN)
Tel.: +34 958 243193, FAX: +34 958 243230 E-mail: {ramon,rubio,atv,juanma}@hal.ugr.es

Resumen

Los sistemas de diálogo mediante el habla suelen usar uno o dos umbrales de confianza durante el reconocimiento de las palabras. El valor de confianza asignado a cada palabra representa la confianza del módulo de reconocimiento respecto al reconocimiento correcto de la misma. Si el valor asignado a una palabra no supera un cierto umbral, la palabra se puede considerar reconocida incorrectamente, o bien, reconocida con reservas, según sea una u otra la estrategia utilizada. En este caso, el sistema deberá solicitar al usuario la confirmación o repetición de la misma. Dado que las condiciones ambientales y las peculiaridades de la voz de cada usuario pueden cambiar de un diálogo a otro, cabe preguntarse qué valor elegir para el umbral de confianza. Un valor demasiado bajo puede provocar que el sistema considere correctamente reconocidas las palabras insertadas erróneamente por el módulo de reconocimiento. Por contra, un valor excesivamente alto puede provocar que incluso las palabras realmente pronunciadas por los usuarios sean consideradas incorrectamente reconocidas, o bien, reconocidas con reservas. En este trabajo vamos a presentar una estrategia para seleccionar el valor más adecuado para el umbral. Hemos aplicado esta estrategia al sistema de diálogo que hemos desarrollado, el cual tiene como finalidad atender por teléfono a los clientes de los restaurantes de comida rápida. Vamos a mostrar los resultados obtenidos en laboratorio y vamos a indicar algunas líneas futuras de investigación.

1. Introducción

La tecnología del Reconocimiento Automático del Habla (RAH) aún dista bastante de la capacidad humana para reconocer el lenguaje, si bien, se han producido avances importantes en los últimos años. Los sistemas que hacen uso de esta tecnología deben cumplir varios requisitos, a fin de ser aceptados ampliamente por los usuarios. En primer lugar, deben presentar alguna ventaja respecto a los sistemas tradicionales de comunicación, como por ejemplo, incremento de productividad o mayor facilidad de uso. Para ello, es crucial que funcionen en tiempo real y proporcionen una tasa de reconocimiento suficientemente alta, teniendo en cuenta los requerimientos de la tarea. Los sistemas de RAH están siendo utilizados actualmente en tareas de control, procesos de entrada de datos, consultas a bases de datos, realización de dictados, sistemas de diálogo, etc. Las aplicaciones de control requieren vocabularios reducidos y altas tasas de exactitud. En la mayor parte de los casos, las órdenes de los locutores son lo suficientemente cortas como para permitir reconocimiento de palabras aisladas o conectadas. Los sistemas de dictado requieren grandes diccionarios. Los sistemas iniciales se basaban en el reconocimiento de palabras aisladas y requerían la adaptación al usuario. Los sistemas actuales pueden reconocer la voz continua y son independientes del locutor. Estos sistemas permiten ahorrar tiempo en la generación de informes, así como reducir los costes de transcripción de documentos. Además, permiten que personas incapacitadas puedan crear sus propios documentos.

Los sistemas de diálogo constituyen una tecnología aparecida a finales de la década de los ochenta, impulsada principalmente por los

proyectos DARPA SLS (*Spoken Language Systems*) en Estados Unidos, y SUNDIAL (*Speech UNDerstanding and DIALog*) en Europa. El objetivo principal de ambos proyectos es desarrollar sistemas informáticos capaces de proporcionar información de viajes a los usuarios, haciendo uso del habla. El proyecto SLS se limita únicamente a los viajes en avión, mientras que el proyecto SUNDIAL abarca la información de viajes tanto en avión como en tren. Los sistemas de diálogo hacen uso de las tecnologías de reconocimiento, comprensión, y síntesis del habla, y deben acomodarse a un amplio rango de variantes acústicas y del lenguaje, las cuales pueden degradar considerablemente el proceso de reconocimiento [López-Cózar et al. 97]. Para mejorar el funcionamiento de estos sistemas, se suele realizar una verificación de las frases reconocidas. Este proceso permite confirmar palabras clave y determinar qué partes de las frases probablemente han sido reconocidas correctamente. Es importante usar una medida de confianza en el reconocimiento de las palabras o frases. Entre otras cuestiones, esta medida se puede usar para decidir cuáles han de ser las respuestas del sistema tras el análisis de las frases de los usuarios. Así, por ejemplo, el sistema puede generar mensajes que indiquen a los usuarios la necesidad de corregir posibles errores de reconocimiento.

2. Algunos conceptos iniciales

Varias secuencias de sonidos pueden representar múltiples secuencias de palabras. Las personas desambiguan inconscientemente estos sonidos usando información sintáctica y otros tipos de información no fonética. Los sistemas de reconocimiento suelen usar grafos de palabras. Dichos grafos son representaciones de todas las secuencias alternativas de palabras, generadas en caso de que no sea posible identificar una secuencia única. Los nodos del grafo representan los instantes de tiempo. Los nodos situados más a la izquierda representan los instantes en los cuales puede comenzar una frase, y los situados más a la derecha representan los instantes en los que la frase puede concluir. Cada camino desde el nodo inicial al final representa una posible frase. Los arcos del grafo se etiquetan con palabras y con medidas de probabilidad de cada palabra. Cuando mayor sea el valor de estas medidas, mayor será la posibilidad de que las palabras se correspondan con los sonidos recibos entre los dos instantes de tiempo. La probabilidad de la frase se estima en función de las probabilidades de las palabras.

Los sistemas de reconocimiento generalmente producen como salida una lista de hipótesis, ordenada según una cierta medida de probabilidad. Sin embargo, los valores de esta medida no proporcionan información acerca de la calidad del proceso de reconocimiento, ni acerca de la confianza del sistema en la decodificación correcta de las palabras. Para los sistemas de diálogo es deseable que los módulos de reconocimiento proporcionen información respecto a la calidad del proceso de reconocimiento, a fin de rechazar las frases (o palabras) reconocidas incorrectamente [Kemp y Schaaf 97]. Podemos hablar de dos tipos de medidas de confianza. Por una parte, las relacionadas con el reconocimiento de las frases, y por otra, las relacionadas con el reconocimiento de las palabras de las frases. Es posible obtener medidas de confianza a partir de los modelos acústicos y del lenguaje conjuntamente, o bien, a partir de cualquiera de ellos de forma independiente. Por ejemplo, en [Uhrik 97] podemos encontrar una medida de confianza basada en una trigramática. La medida considera que las probabilidades provenientes de las trigramáticas son más fiables que las provenientes de las bigramáticas, y que éstas, a su vez, son más fiables que las procedentes de las unigramáticas. No obstante, esta medida está basada únicamente en el modelo del lenguaje, el cual constituye sólo una parte de la información utilizada durante el proceso de reconocimiento.

El reconocimiento de voz continua presenta muchas más dificultades que el reconocimiento de palabras aisladas. Por una parte, los límites entre las palabras y frases pueden ser poco claros. Esta imprecisión impide la división del discurso en unidades claramente diferenciadas que puedan ser tratadas individualmente (palabras, por ejemplo), con el consiguiente aumento de la complejidad de los sistemas de reconocimiento. Por otra parte, los efectos coarticulatorios son mucho más acusados en el discurso continuo. La entonación inherente al discurso continuo añade un nuevo factor de

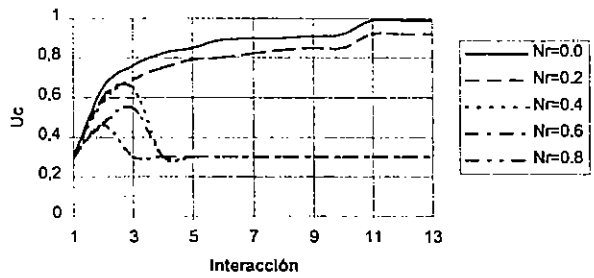
complejidad, pues la pronunciación de cada palabra depende de su posición en la frase [Lee y Alleba 91]. Como consecuencia de todo ello, las tasas de error obtenidas para el reconocimiento de voz continua son considerablemente mayores que las correspondientes a palabras aisladas. El conjunto de frases que pueden ser reconocidas viene determinado por la gramática utilizada. Si se utiliza una gramática muy simple, el proceso de reconocimiento se realiza muy rápidamente, y la tasa de error se mantiene pequeña. Sin embargo, el número de frases que pueden ser reconocidas es muy reducido, y por tanto, la interacción con los usuarios deja de ser espontánea. Si se utiliza una gramática muy compleja aumenta el tiempo de reconocimiento y la tasa de error. No obstante, el número de frases que pueden ser reconocidas es mucho mayor, lo que puede redundar en un diálogo más espontáneo. Generalmente es necesario llegar a un compromiso respecto a la complejidad de la gramática que se va a usar durante el reconocimiento.

3. Umbral de confianza dinámico

Los sistemas de diálogo suelen usar un umbral de confianza para decidir qué palabras probablemente han sido reconocidas correctamente. El valor del umbral de confianza es crucial para realizar la comprensión robusta de las frases. Dicho valor debe ser lo suficientemente alto como para evitar que las palabras insertadas erróneamente por el módulo de reconocimiento puedan confundir al módulo que debe realizar el análisis lingüístico. Al mismo tiempo, el valor del umbral debe ser lo suficientemente bajo como para permitir que las palabras realmente pronunciadas por el usuario puedan llegar al analizador lingüístico. Por tanto, es necesario llegar a un compromiso. Nuestra propuesta consiste en utilizar un umbral de confianza dinámico, el cual se adapte a las características de cada conversación. La idea consiste en determinar el valor más adecuado para el umbral conforme se desarrolla el diálogo, mediante un proceso de actualización sucesivo. A tal fin, el sistema de diálogo que hemos desarrollado dispone de un buffer en el que almacena el valor de confianza asignado a cada una de las n últimas interacciones del usuario. El valor de confianza asignado a una interacción del usuario se corresponde con la media de los valores de confianza asignados a las palabras de la interacción. Al principio de cada diálogo, el sistema introduce en el buffer el valor de un umbral de confianza considerado mínimo, y a lo largo del diálogo, introduce el valor de confianza asignado a cada interacción. Durante el diálogo, el valor del umbral se calcula como la media de los valores de confianza existentes en el buffer. Si el valor obtenido fuera menor que el valor de confianza mínimo, entonces el nuevo valor del umbral de confianza sería el valor de confianza mínimo. Esta estrategia de actualización conlleva un incremento lento del valor del umbral de confianza. Si el sistema ha comprendido correctamente la frase anterior, el valor actual del umbral se considera adecuado. En caso contrario, el valor actual se considera inadecuado. Durante los experimentos, hemos observado que en un gran número de ocasiones, la incomprensión de las frases es consecuencia de estar excesivamente alto el umbral de confianza. Por tanto, ante una incomprensión debemos reducir su valor. Para ello, retiramos del buffer los valores de confianza mayores o iguales que el valor actual del umbral, y volvemos a calcular su valor en función de los valores que queden en el buffer. En este momento, consideramos que el umbral está "fijado". Una vez fijado el umbral, no volvemos a actualizar su valor a lo largo del diálogo a no ser que se produzca alguna otra incomprensión, en cuyo caso, volvemos a reducir su valor de la forma que acabamos de describir.

Mostramos en la siguiente gráfica el proceso de actualización del umbral de confianza (Uc) para varios valores del parámetro Nr (nivel de ruido). Este parámetro determina la proporción de la energía de la señal que corresponde a ruido. Nr=0 representa el caso ideal, en el cual toda la energía de la señal corresponde a la voz del usuario. Suponemos que en este caso no se producen errores de reconocimiento. El proceso de actualización mostrado en la gráfica corresponde a uno de los diálogos utilizados durante los experimentos. En el eje de abscisas indicamos las sucesivas interacciones, y en el eje de ordenadas mostramos los valores alcanzados por el umbral de confianza como

consecuencia de la actualización. El valor mínimo del umbral de confianza es 0.3.



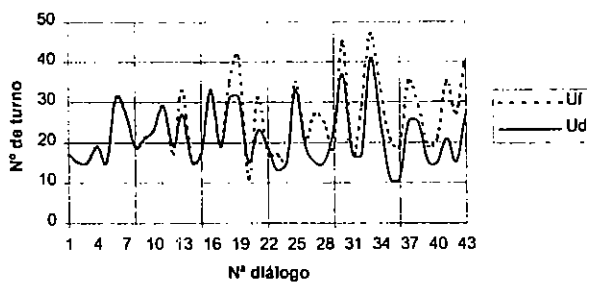
Gráfica 1. Actualización del umbral de confianza dinámico

Como podemos observar, en el caso ideal (Nr=0) el valor del umbral de confianza se incrementa lentamente hasta quedar fijado en el valor máximo (Uc=0.99) en la interacción número 11. Cuando Nr=0.2 ocurre algo similar, pero en este caso, el incremento del umbral se realiza más lentamente debido a que el valor de confianza asignado a cada interacción es menor, por existir ruido. En este caso, el umbral queda fijado al valor 0.92. Cuando Nr=0.4 se producen sólo dos incrementos de Uc, pues en la tercera interacción se produce una incomprensión, lo que provoca que el valor del umbral quede fijado al valor mínimo (Uc=0.33). Cuando Nr=0.6 se vuelven a producir sólo dos incrementos de Uc, pues en la tercera interacción se vuelve a producir una incomprensión, lo que provoca que el valor del umbral quede fijado al valor mínimo. Finalmente, cuando Uc=0.8 sólo se produce un incremento de Uc, pues en la segunda interacción se produce una incomprensión, lo que provoca que el umbral quede fijado al valor mínimo.

4. Experimentos

Para evaluar el efecto del umbral dinámico, hemos usado 43 diálogos realizados previamente por usuarios de test del sistema que hemos desarrollado [López-Cózar et al. 98]. Estos diálogos se han llevado a cabo en un restaurante de comida rápida usando diferentes valores del umbral de confianza, pero en todos los casos, el valor del umbral permanece fijo. Los diez primeros diálogos seleccionados corresponden al caso ideal (Uc=0) en el cual se supone que no existen errores de reconocimiento. Los diálogos 11-20 corresponden al grupo de Uc=0.6, los diálogos 21-30 corresponden al grupo de Uc=0.7, los diálogos 31-40 corresponden al grupo de Uc=0.8, y finalmente, los diálogos 41-43 corresponden al grupo de Uc=0.9. Este último grupo es el más reducido, pues sólo tres de los usuarios entrevistados no abandonaron el diálogo. Hemos reproducido en laboratorio los 43 diálogos seleccionados utilizando un módulo de simulación de errores de reconocimiento, el cual permite la inserción, cambio o supresión de palabras en las frases de los usuarios [López-Cózar et al. 98]. Hemos usado un umbral de confianza dinámico, cuyo valor se determina siguiendo el procedimiento descrito previamente.

Durante los experimentos de laboratorio, hemos medido la duración de los diálogos y la capacidad de comprensión por parte del sistema. Nuestro objetivo ha sido evaluar el funcionamiento del sistema cuando se usa el umbral fijo y el umbral dinámico. La siguiente gráfica muestra los resultados obtenidos con respecto a la duración de los diálogos, medida en número total de turnos necesarios para concluir cada diálogo. La línea discontinua corresponde a los diálogos realizados por los usuarios de test, en los cuales, se ha usado un umbral de confianza fijo (Uf). La línea continua corresponde a los diálogos reproducidos en el laboratorio mediante el umbral de confianza dinámico (Ud).

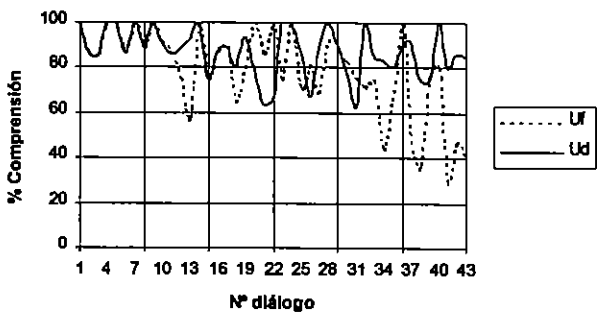


Gráfica 2. Duración de los diálogos

Como podemos observar en esta gráfica, la duración es similar en los diez primeros diálogos. Este hecho se debe a que estos diálogos corresponden al caso ideal, en el que

no se producen errores de reconocimiento Las diferencias comienzan a aparecer a partir del diálogo número 12. Como podemos observar en la gráfica, la duración de los diálogos suele ser menor cuando se usa el umbral dinámico. La mayor diferencia se alcanza en los diálogos del último grupo (a partir del número 41), como consecuencia de la aparición de un excesivo número de errores de reconocimiento.

La siguiente gráfica muestra los resultados obtenidos con respecto a la comprensión de las frases por parte del sistema. La línea discontinua corresponde a los diálogos realizados por los usuarios de test, en los cuales, se ha usado un umbral de confianza fijo (Uf). La línea continua corresponde a los diálogos reproducidos en el laboratorio mediante el umbral dinámico (Ud).



Gráfica 3. Capacidad de comprensión por parte del sistema

Como podemos observar en esta gráfica, el porcentaje de comprensión por parte del sistema es similar en los diez primeros diálogos. Como ocurre en el caso anterior, las diferencias comienzan a aparecer a partir del diálogo número 12. Generalmente, el porcentaje de comprensión por parte del sistema es mayor cuando es usa el umbral dinámico. La mayor diferencia se alcanza para los diálogos de los dos últimos grupos (a partir del número 31), como consecuencia del aumento de los errores de reconocimiento.

5. Conclusiones y trabajo futuro

A partir de los resultados obtenidos, podemos concluir que el funcionamiento del sistema mejora cuando usamos el umbral de confianza dinámico. El porcentaje medio de comprensión en los diálogos cuando se usa el umbral fijo es 78.69%, mientras que cuando se usa el umbral dinámico es 86.96%. La duración media de los diálogos cuando se usa el umbral fijo es 25.04 turnos, mientras que cuando se usa el umbral dinámico es 21.34 turnos. Si bien los resultados obtenidos son satisfactorios, debemos tomarlos con cautela. La idea nos parece interesante, y creemos que la técnica puede resultar de interés en ciertas circunstancias. No obstante, debemos realizar experimentos en situaciones reales para obtener resultados más fidedignos. Además, debemos investigar diversas técnicas de actualización del umbral, y comparar los resultados obtenidos.

Actualmente, el sistema que hemos desarrollado usa un único valor de confianza para aceptar o rechazar las palabras reconocidas. Una alternativa más elaborada consiste en usar dos umbrales de confianza. En este caso, cada palabra se podría considerar como no reconocida correctamente, reconocida con reservas, o reconocida correctamente. Sea *conf(w)* el valor de confianza asociado a la palabra *w*. Podríamos usar dos umbrales, Uc_1 y $Uc_2 \in (0,1)$, $Uc_1 < Uc_2$, de forma que si $conf(w_i) < Uc_1$ entonces w_i se consideraría no reconocida correctamente, si $conf(w_i) = Uc_1$ y $conf(w_i) < Uc_2$ entonces w_i se consideraría reconocida con reservas, y finalmente, si $conf(w_i) = Uc_2$ entonces w_i se consideraría reconocida correctamente. Por ejemplo, supongamos que ante la pregunta del sistema: *"¿DE QUE QUIERES EL BOCADILLO?"*, el usuario responde: *"DE_LOMO"*. El valor de confianza asociado a esta palabra podría ser: *conf("DE_LOMO")*=0.75. Si supuestamente los umbrales de confianza son Uc_1=0.7 y Uc_2=0.8, entonces la respuesta del usuario se consideraría reconocida con reservas. En este caso, el sistema debería generar una pregunta para obtener la confirmación de la respuesta del usuario, por ejemplo *"¿HAS DICHO QUE QUIERES UN BOCADILLO DE LOMO?"*.

Varios miembros del grupo de investigación han desarrollado un módulo de reconocimiento de voz continua que ha sido utilizado con éxito en el sistema STACC [Rubio et al. 97]. La finalidad de este sistema es proporcionar información a los alumnos sobre las calificaciones obtenidas en diversas titulaciones de la Universidad de Granada. Este sistema realiza una gestión del diálogo muy simple, no

realiza comprensión de las frases, y realiza la síntesis de voz a partir de frases pregrabadas. No requiere valores de confianza asociados a cada una de las palabras reconocidas, sino un único valor de probabilidad de la frase que se está reconociendo en cada momento, el cual se utiliza para realizar la poda de los caminos menos probables. Estamos trabajando actualmente en la adaptación de este módulo para que pueda ser usado eficientemente en el nuevo sistema de diálogo desarrollado, cuya finalidad es atender por teléfono a los clientes de los restaurantes de comida rápida [López-Cózar y Rubio 97]. La gestión del diálogo que hemos desarrollado en el nuevo sistema es mucho más compleja. El módulo de gestión del diálogo usa los valores de confianza asociados a las palabras, en función de los cuales, puede decidir entre requerir al usuario la repetición de una frase completa, o bien, la repetición únicamente de los datos que no han sido reconocidos de forma correcta.

6. Referencias

[Lee y Alleba 91] Lee K. F., Alleva F., "Advances in Speech Signal Processing", capítulo Continuous Speech Recognition, pp. 623-650, Dekker

[Kemp y Schaaf 97] Thomas Kemp, Thomas Schaaf, "Estimating Confidence Using Word Lattices", Eurospeech '97, pp. 827-830

[López-Cózar et al. 97] Ramón López-Cózar, Pedro García, J. Díaz, Antonio J. Rubio. "A Voice Activated Dialog System for Fast-Food Restaurant Applications", Eurospeech '97.

[López-Cózar et al. 98] R. López-Cózar, A. J. Rubio, P. García, J. C. Segura, "A Spoken Dialogue System Based on a Dialogue Corpus Analysis", First International Conference on Language Resources and Evaluation (LREC'98), pag. 55-58

[López-Cózar y Rubio 97] Ramón López-Cózar Delgado, Antonio J. Rubio Ayuso. "SAPLEN: Un Sistema de Diálogo en Lenguaje Natural para una Aplicación Comercial". III Jornadas de Informática '97. AEIA

[Rubio et al. 97] A.J. Rubio, P. García, A. de la Torre, J.C. Segura, J. Díaz-Verdejo, M. C. Benítez, V. Sánchez, A.M. Peinado, J.M. López-Soler, J.L.Pérez-Córdoba: "STACC: an automatic service for information access using continuous speech recognition through telephone line". Eurospeech '97, vol. 4, pp. 1779-82, Sept. 1997

[Uhrik 97] C. Uhrik, "Confidence Metrics Based on N-Gram Language Model Backoff Behaviors", Eurospeech '97, pp. 2771-2774

[Williams y Renals 97] Gethin Williams and Steve Renals, "Confidence Measures For Hybrid Hmm/Ann Speech Recognition", Eurospeech '97, pp. 1955-1958