SISTEMA DE RECONOCIMIENTO AUTOMÁTICO DEL HABLA DISTRIBUIDO APLICADO A ENTORNOS LOGÍSTICOS

José Enrique García, Alfonso Ortega, Antonio Miguel y Eduardo Lleida.

Grupo de Tecnologías de las Comunicaciones (GTC)
I3A, Universidad de Zaragoza
{jegarlai,ortega,amiguel,lleida}@unizar.es

RESUMEN

Los sistemas de reconocimiento automático del habla (RAH) pueden llegar a ser muy útiles aplicados a procesos productivos en el sector industrial como los desarrollados en entornos logísticos. Tareas como el etiquetado de paquetes o la anotación de matrículas de transportista se pueden llevar a cabo usando únicamente la voz con la consiguiente ventaja de disponer de las manos libres para la ejecución de otras tareas. En este artículo se presenta un sistema de control por voz aplicado a la logística sobre un dispositivo móvil, a partir de una arquitectura distribuida cliente-servidor donde un PC convencional recibe los parámetros acústicos enviados por el dispositivo móvil, en este caso una PDA, realiza la decodificación acústica, y procede a la actuación. Además generar una respuesta oral a partir de un sintetizador de voz enviándola a la PDA. Además de esquemas de implementación del sistema, se ofrecen unos estudios de prestaciones que caracterizan el número de operarios que pueden actuar simultáneamente con una calidad de servicio adecuada.

1. INTRODUCCIÓN

Este artículo presenta un sistema de control por voz sobre dispositivos móviles aplicado a tareas de picking y transporte, basado en una arquitectura cliente-servidor y capaz de extenderse a cualquier otro tipo de tareas productivas dentro del sector industrial.

En la tarea de recogida de mercancías en almacén (picking) el operario logístico debe realizar la preparación de pedidos de acuerdo con un plan prefijado. De esta manera las instrucciones deben ser seguidas meticulosamente, verificando el cumplimiento de las mismas a la vez que se manipulan los bultos. La libertad de movimientos que un sistema de gestión con interfaz oral provee, permitiría al operario la recepción de órdenes y la verificación de las mismas sin tener que utilizar sus manos para la manipulación de un dispositivo de gestión de tareas, quedando éstas disponibles para la extracción, preparación y transporte de unidades y lotes.

Este trabajo ha sido financiado a través de la colaboración con la empresa Alerce Informática S.A. y el proyecto de investigación TIN2005-08660-C04

En cuanto a la recepción de mercancías en muelle, los operarios del almacén realizan una serie de acciones de desplazamiento y etiquetado de artículos con el objetivo de llevar a cabo una exhaustiva identificación de la mercancía recibida. Un etiquetado haciendo uso del RAH podría ayudar a hacer el proceso productivo más eficiente ya que las manos del operario quedarían libres para hacer el desplazamiento de los artículos de forma simultánea.

Dado que los operarios que hacen uso del RAH necesitan movilidad, para desplazar cajas, para manejar maquinaria específica, para controlar los camiones que llegan o salen del muelle, etc., es conveniente que el sistema de reconocimiento esté incorporado sobre un dispositivo móvil o PDA, en lugar de hacerlo sobre un emplazamiento fijo en el almacén al que tendrían que moverse los operarios cada vez que necesitasen anotar algún evento.

Los sistemas de RAH embebidos en dispositivos móviles suelen ofrecer problemas de funcionamiento en tiempo real para tareas medianamente complejas debido a su limitada capacidad de cálculo. La tarea sólo será mejor asistida a través de una interfaz oral si presenta ventajas en cuanto a su eficiencia con respecto al empleo de una interfaz visual-táctil. El tiempo consumido en cada preparación de pedido o en cada etiquetaje es crucial y siempre deberá ser menor a través del uso de la interfaz oral, en cualquier otro caso, su uso no será considerado por motivos de productividad. De ahí que se suelan emplear arquitecturas distribuidas cliente-servidor, donde el dispositivo móvil actúa como cliente realizando el acondicionamiento y extracción de vectores de parámetros acústicos, mientras que un ordenador convencional actúa como servidor, realizando la decodificación acústica que normalmente es el proceso más costoso computacionalmente en el reconocimiento del habla. Haciendo uso de la interfaz inalámbrica (de la que suelen disponer la gran parte de PDAs), cada operario es capaz de lanzar distintos programas de control por voz creando una conexión con un ordenador que actúa como el servidor del sistema. El ordenador personal tiene la misión de realizar la decodificación acústica de las tramas enviadas por la PDA, además de generar las respuestas orales y las

actuaciones correspondientes en función de los comandos de voz reconocidos.

La precisión del sistema de reconocimiento automático del habla también es un aspecto crucial. Una elevada tasa de error hace que el usuario deba corregir continuamente las acciones del sistema con la consiguiente pérdida de eficiencia.

El artículo se organiza de la siguiente manera: En la Sección 2 se muestra la problemática y las tareas asociadas al control por voz en el ámbito de la logística. Una descripción de la arquitectura del sistema de control por voz distribuido cliente-servidor se presenta en la Sección 3. En la Sección 4 se presentan dos estudios experimentales que caracterizan las prestaciones del sistema, de forma que es posible dimensionar el número de servidores y redes inalámbricas necesarios dependiendo del número de operarios que se prevea puedan estar actuando simultáneamente en el almacén. Finalmente, en la Sección 5 se encuentran las conclusiones.

2. INCORPORACIÓN DEL 'RAH' A TAREAS LOGÍSTICAS

El sistema que se presenta en este artículo está destinado a su utilización en tareas de picking (acciones llevadas a cabo por un operario en un almacén) y de transporte, más concretamente al llenado de formularios y albaranes a la llegada o salida de camiones de un muelle.

Dado que en los ambientes a los que va destinado el control por voz se encuentra presente una gran cantidad de maquinaria pesada (carretillas, carruseles, robots industriales,…), otros trabajadores, artículos que se caen o se golpean, etc. se puede hablar de un entorno acústico no controlado, con ruido de magnitud considerable proveniente de varias fuentes y no estacionario. Sin embargo, el uso de un micrófono situado en las proximidades de la boca del locutor (close-talk) puede, de forma considerable ayudar a reducir las impurezas de la señal de audio capturada.

Todos los procesos de llenado de formularios por voz siguen el mismo protocolo, consistente en un diálogo guiado en el que el dispositivo móvil pregunta un campo, el operario lo introduce con la voz y el dispositivo móvil confirma lo que ha reconocido, pasando al siguiente campo del formulario. Al producirse la confirmación de lo que el dispositivo móvil ha reconocido, el operario dispone de la posibilidad de repetir la entrada por voz en el caso de que se hubiese producido un error.

Cada formulario dispone de un número variable de campos configurable de acuerdo con la tarea a desarrollar. Algunas de las tareas de reconocimiento que se llevaron a cabo para los campos fueron las siguientes: reconocimiento de dígitos, direcciones, nombres de empresas, localidades, códigos postales, dimensiones, y pesos. Todas las tareas suponían un reconocimiento del habla de pequeño-mediano vocabulario.

3. ARQUITECTURA DISTRIBUIDA CLIENTE-SERVIDOR

La motivación de emplear un sistema de reconocimiento de voz distribuido reside en la baja capacidad de cálculo de la que hoy en día disponen los dispositivos móviles. Este hecho, sumado a la carencia de éstos de unidad de punto flotante, conlleva la dificultad añadida de tener que programar los algoritmos de reconocimiento como operaciones en coma fija para que éstos sean capaces de actuar en tiempo real. Por ello, los sistemas de RAH sobre dispositivos móviles suelen plantearse de forma distribuida, donde el dispositivo móvil realiza las tareas de acondicionamiento, procesado y extracción de vectores de observación acústicos para enviarlos a un PC que realiza el algoritmo de reconocimiento.

Los vectores de observación acústicos empleados en el sistema son los conocidos Mel Frequency Cepstral Coefficients (MFCC), definidos en el estándar del ETSI ES 201 108 V1.1.2 (2000-04) [1]. Estos vectores se cuantifican en una última etapa de procesado en el dispositivo móvil, de forma que el ancho de banda de envío se reduce a la tasa de 4.4 Kbps. Ya que el canal sobre el que se envían se corresponde con un canal seguro (TCP) sobre la interfaz inalámbrica, se eliminaron las cabeceras empleadas en el ETSI DSR para mitigación de errores, con lo que se reduce ligeramente el ancho de banda de envío.

El PC que actúa como servidor recibe los parámetros cuantificados, y procede a la de-cuantificación para introducir los vectores acústicos en el motor de reconocimiento del grupo de investigación basado en HMM, el cual hace uso de unidades acústicas contextuales, donde cada unidad es representada mediante un estado y modelada a partir de una GMM de 16 componentes. Una vez el PC ha decidido que se ha reconocido algún comando de voz, genera la respuesta oral que es sintetizada y codificada mediante la batería de codecs libre S*peex* [2] la cual hace uso de un codificador de voz perceptual CELP. Finalmente el bitstream generado es enviado al dispositivo móvil el cual procede a la decodificación y reproducción de dicha respuesta por sus altavoces. El codec empleado era capaz de conseguir un ancho de banda de subida para la respuesta oral de aproximadamente 6.2 kbps con una calidad similar a una codificación PCM con frecuencia de muestreo 16KHz y 16 bits por muestra.

El diagrama de bloques del sistema se puede ver en la Figura 1. El proceso que aparece en la Figura 1 se repite, mientras el operario permanece conectado al sistema, cada vez que aparece un nuevo comando oral reconocido. Se puede ver que tanto el ancho de banda de envío de tramas de voz para reconocimiento (4.4 kbps) como en ancho de banda de envío de respuestas

orales codificadas son muy reducidos (6.2 kbps), por lo que una red WIFI puede dar cabida a un número muy elevado de operarios simultáneamente, en cuanto a prestaciones de ancho de banda se refiere. En el apartado 3 de este artículo se presentan algunas estimaciones para el dimensionado de la red inalámbrica.

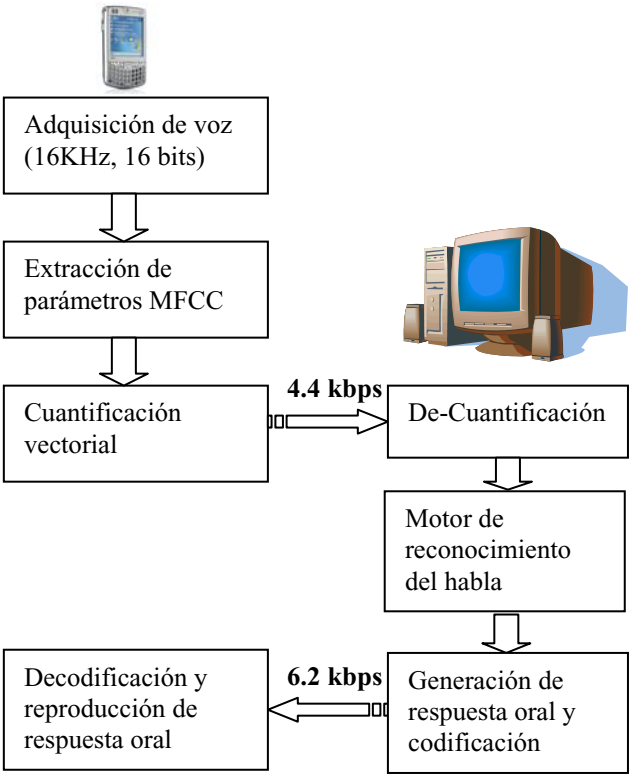

Figura 1. *Diagrama de bloques del sistema de control por voz distribuido.*

4. ESTUDIOS DE PRESTACIONES PARA DIMENSIONADO DEL SISTEMA

En cada PC que aparece como servidor es posible que se lancen tantos procesos de reconocimiento simultáneos como operarios conectados hay en el sistema. En principio, se podrían conectar un número muy elevado de operarios que disponen de su dispositivo móvil con un único PC, pero es posible que las prestaciones del sistema se viesen degradadas notablemente cuando el número de operarios conectados fuese demasiado grande. Cuando se hace referencia al término prestaciones, realmente podríamos evaluarlas en función del retardo en conocer la respuesta oral de los comandos de voz. Si este retardo es muy grande, el diálogo se ralentizaría de tal forma que sería imposible realizar un proceso de llenado de formularios o el etiquetado sería más rápido si se hiciese manualmente.

Por ello, el objetivo es realizar el dimensionado del sistema de forma que se coloquen los servidores necesarios para que todos los operarios, o una gran parte de ellos, obtengan respuestas orales a los comandos con retardos aceptables.

Los factores que pueden influir en el retardo de respuesta ofrecida por el servidor de reconocimiento son 2: El tiempo de procesado del ordenador que actúa como servidor y el ancho de banda de ocupación de la red inalámbrica.

Los sistemas de comunicaciones basados en WIFI, correspondientes a la familia de estándares 802.11b y 802.11g los cuales están implementados en la interfaz inalámbrica de la mayor parte de las PDAs, tienen velocidades de transmisión de 11 Mbps y 54 Mbps, respectivamente. Haciendo un cálculo simple, se puede ver como suponiendo el caso más desfavorable (de ocurrencia muy remota) en el que el ancho de banda total consumido por un cliente fuese 10.6 kbps (la suma del ancho de banda de envío de tramas de voz más el ancho de banda de recepción de respuestas orales) durante todo el tiempo, el número de clientes soportados para un funcionamiento en tiempo real sin retardos sería 1037 y 5094 clientes respectivamente, por red inalámbrica.

De ahí se puede observar cómo el factor más problemático que puede degradar las prestaciones es la capacidad de un servidor para procesar peticiones de clientes, y que salvo en entornos en los que el número de operarios fuese excesivamente grande, sería suficiente con un único punto de acceso.

4.1. Retardo debido al tiempo de procesado del servidor

El retardo de la respuesta oral ofrecida por el servidor se ve influenciado por el tiempo que le cuesta a éste procesar las peticiones de los clientes. Este tiempo de procesado depende principalmente de dos factores: el número de clientes conectados simultáneamente y la perplejidad de las tareas de reconocimiento con las que se está tratando en cada cliente.

Se fijaron dos experimentos, con tareas de RAH típicas del sistema de control logístico por voz. El ordenador elegido como servidor disponía de un procesador Intel Pentium IV a 3.4 GHz, bajo el sistema operativo Microsoft Windows XP. En ambos experimentos se procedió a desactivar el conversor de texto a voz, de forma que éste no influyese en los resultados ya que lo que realmente se quería evaluar eran las prestaciones del motor de reconocimiento.

El primer experimento consistió en lanzar simultáneamente varios procesos de reconocimiento en el PC, con las tareas de reconocimiento de dígitos primero, y reconocimiento del nombre de 40 calles después, midiendo el tiempo de ejecución para completar un número de procesos paralelos determinado, y obteniendo su cociente frente al tiempo real. En la Figura 2 se ve una gráfica que muestra los resultados para la tarea de reconocimiento de dígitos.

Se puede observar en la Figura 2 cómo a partir de 13 procesos paralelos el PC que actúa como servidor no sería capaz de llevar a cabo las peticiones en tiempo real, por lo que induciría un retardo en el sistema de control por voz. En la tarea de reconocimiento de calles, el número de procesos paralelos en los que se observó el corte fue 11, debido a la mayor perplejidad de la tarea. Estos valores obtenidos en los experimentos nos pueden dar una aproximación teórica de cuándo comenzarán a existir retardos no aceptables en las peticiones de los clientes para un servidor con unas características determinadas.

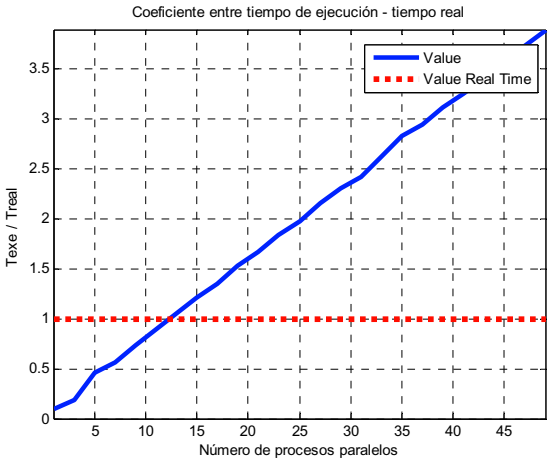

Figura 2. *Curva con el cociente Tiempo ejecución/ Tiempo real en función del número de procesos paralelos para la tarea de reconocimiento de dígitos.*

El segundo experimento llevado a cabo fue realizar medidas de retardos, a partir de una simulación de diálogo llevada a cabo lanzando hasta 17 clientes simultáneos desde 3 ordenadores personales, que se conectaban todos con el mismo PC servidor presentado anteriormente. La conexión se realizó por una red LAN de 100 Mbps. de manera que el retardo de la comunicación tuvo una influencia mínima en la medida de las prestaciones. Las tareas que se trataron fueron tanto el reconocimiento de dígitos como el reconocimiento de calles para el mismo experimento. Cada uno de los clientes que se iba lanzando realizaba medidas de retardo, sabiendo con cuántos clientes se estaba accediendo simultáneamente al servidor. A partir de estas medidas de retardo, se obtuvieron histogramas de retardo en función del número de clientes conectados al servidor de forma simultánea, obteniendo los resultados presentados en la Figura 3.

Se puede observar cómo a partir de unos 25 procesos paralelos, la distribución estadística de los retardos comienza a tener mayor varianza, a pesar de subir ligeramente en media (la media está representada en color blanco sobre la gráfica). Es importante anotar que el valor de retardo mínimo que aparece (algo menor de 1 segundo) se corresponde con la ventana temporal de decisión del detector de actividad de voz para decidir si un comando ha sido reconocido o no.

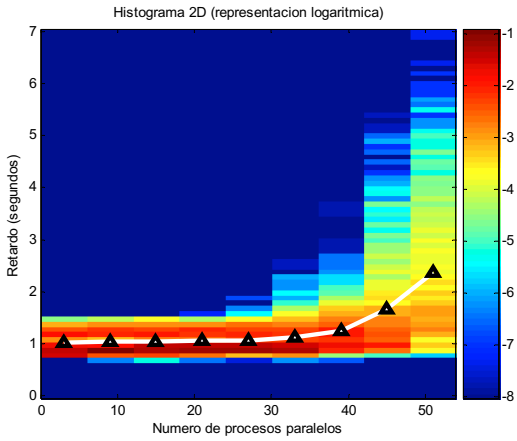

Figura 3. *Log-histograma de retardos de respuesta en función del número de clientes conectados simultáneamente al sistema. Superpuesto aparece representada la curva con el valor medio de dicho retardo.*

5. CONCLUSIONES

En este artículo se ha presentado un sistema de control por voz para entornos logísticos mediante una arquitectura distribuida cliente-servidor que puede llegar a ser muy útil para liberar las manos de los operarios en almacenes de tal forma que así puedan realizar varias acciones simultáneamente.

Es importante señalar que los recursos de ancho de banda consumidos por el sistema distribuido son muy pequeños (4.4 kbps de subida y 6.2 kbps de bajada) con unas buenas prestaciones de reconocimiento y una muy aceptable calidad de reproducción de la respuesta oral. Con este ancho de banda de bajada y subida, haciendo uso de una red inalámbrica WIFI, se pueden dar cabida a un número muy elevado de operarios simultáneamente.

También se ha realizado un estudio experimental de dimensionado, en el que se ha demostrado que con un ordenador personal convencional con un procesador Intel Pentium IV a 3.4 GHz. bajo el sistema operativo Microsoft Windows XP, se puede dar cabida con una calidad de servicio óptima y retardo mínimo a unos 25 usuarios. Si el servidor elegido para hacer los experimentos hubiese sido una máquina más potente de las que se dispone hoy en día normalmente en los servidores de elevada capacidad de cómputo este número de usuarios soportados hubiese sido mucho mayor.

Los autores desean agradecer a la empresa Alerce Informática S.A. su colaboración y aportaciones para la realización de este trabajo.

6. BIBLIOGRAFÍA

[1] ETSI standard doc. “Speech Processing, Transmission and Quality Aspects (STQ); Distributed speech recognition; Front-end Feature extraction algorithm; Compression algorithms”, ETSI ES 201 108 Ver.1.1.2 (2000-04).

[2] Valin, J.M. “The Speex Codec Manual”, Disponible en línea: http:// www.speex.org (Visitado: 27/07/2008).