ASERRADERO: Un Sistema IVR basado en Transceptores

Laura Docío-Fernández *Carmen García-Mateo* *Antonio Cardenal-López*

Departamento de Tecnoloxías das Comunicacións
Universidade de Vigo

ldocio@gts.tsc.uvigo.es carmen@gts.tsc.uvigo.es cardenal@gts.tsc.uvigo.es

ABSTRACT

In this paper, we describe the design procedure for a wireless communication interactive voice response (IVR) system. We will address the sensible aspects of three components of the application: the voice activity detector (VAD), the automatic speech recognition (ASR) system, and the confidence measure (CM) determination. The application must work in a very noisy environment which has imposed many desing constraints. In order to get a satisfactory product, it has been necessary to reduce the important mismatch between available linguistic and acoustic resources and the operational environment. Adaptation techniques for the acoustic models of the speech recognition system have proven to be effective to speed up the application deployment time.

1. INTRODUCCIÓN

En los últimos años se han desarrollado muchos servicios y productos en el ámbito de las telecomunicaciones que utilizan la denominada Tecnología del Habla. Para obtener un producto de calidad se necesita personal altamente especializado y bien entrenado. Aunque se haya realizado un diseño cuidadoso de las especificaciones del proyecto y una selección adecuada de sus componentes, se debe realizar una fase de prueba y error para ajustar todo el sistema.

La aplicación que se describe en este artículo es un sistema de comando-control muy simple, pero el entorno acústico (un aserradero muy ruidoso), el canal de transmisión (un canal radio *half-duplex*), y los dos posibles lenguajes (gallego y castellano) dificultan la transferencia de la tecnología disponible a un producto completo y funcional. Por ello se van a describir algunas técnicas que se deben aplicar para mejorar las prestaciones y ergonomía de la aplicación.

2. DESCRIPCIÓN DE LA APLICACIÓN

"Aserradero" es una aplicación de comando-control que opera en un entorno muy adverso: un aserradero con un alto nivel de ruido debido a maquinaria del propio aserradero y a camiones y elevadores-transportadores.

La tarea a realizar es etiquetar una pila de tablas, siendo el objetivo final calcular el volumen en m3 de madera y rellenar una plantilla con las características de cada tabla.

La aplicación en sí impone limitaciones que nos han llevado a tomar las siguientes decisiones:

- Empleo de *walkie-talkies* como medio de comunicación entre el ordenador y el operario.
- Sistema de reconocimiento automático independiente del locutor basado en unidades de tipo fonético.
- Confirmación explícita de la salida del reconocedor.
- Cuidadoso diseño del sistema de respuestas al ser un canal *half-duplex*.

La figura 1 muestra un diagrama de bloques de la aplicación.

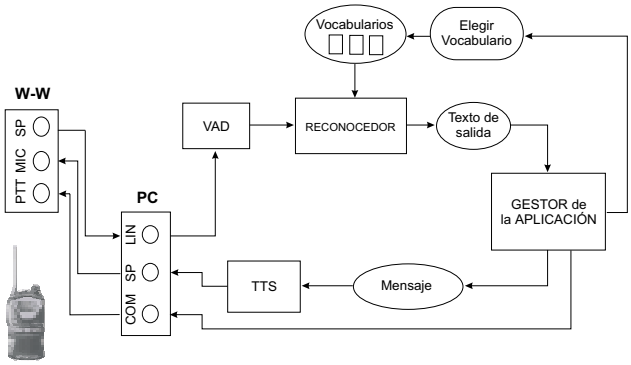

Figura1: *Diagrama de bloques de "Aserradero".*

3. DESCRIPCIÓN DEL VAD

El diseño de un detector de actividad vocal (Voice Activity Detector – VAD) es crucial en cualquier sistema interactivo. En nuestro sistema, se utilizan dos VAD: uno para detectar cuando el operario abre el canal y comienza a hablar, y otro para detectar cuando termina de hablar. El primero se basa en HMMs para realizar la decisión entre "canal" y "silencio/voz", y el segundo se basa en medidas de energía. Éste VAD es muy sensible a cambios en el ruido de fondo y a variaciones en el volumen del transmisor, por lo que se ha implementado un algoritmo que permite hacer un reajuste de los umbrales utilizados en las decisiones cada vez que el operario lo solicite.

4. DESCRIPCIÓN DEL SISTEMA ASR

La señal de voz de entrada se muestrea a 8 KHz y se preemfatiza utilizando un filtro de primer orden con un factor de 0.97. Las tramas son de 25ms con un desplazamiento entre ellas de 10ms. Como características acústicas se han utilizado 12 coeficientes

mel-cepstrum y el coeficiente c0 más las derivadas de primer y segundo orden. Se aplica también una normalización cepstral. Todos los modelos tienen tres estados y una topología de derecha-a-izquierda. Los modelos acústicos son independientes del contexto y se han entrenado utilizando un subconjunto de la base de datos VOGATEL[2].
Como un primer intento de reducción del desajuste entre VOGATEL y el entorno de operación hemos filtrado la base de datos de entrenamiento a través de una estimación del canal radio, y entrenado un nuevo conjunto de modelos de fonemas. Esto resultó efectivo pero no suficiente para considerar el problema resuelto.
Se han aplicado también técnicas de adaptación para construir modelos acústicos mejorados. Hemos utilizado adaptación basada en regresión lineal de máxima verosimilitud (MLLR) en modo estático utilizando el software HTK[5]. El objetivo es realizar una adaptación al entorno y no a un locutor particular.

4.1. Bases de datos de test

Se han recogido tres diferentes córpora para llevar a cabo los experimentos y ajustar los diferentes componentes del sistema. Las principales características de cada corpus son:

- Corpus1: Grabado en un laboratorio utilizando los transceptores. Se dispone de 5 locutores y 1800 frases.
- Corpus2: Grabado en el aserradero. Se dispone de 6 locutores y 310 frases.
- Corpus3: Grabado en el aserradero. Se dispone de 7 locutores y 363 frases.

Para la adaptación utilizando el algoritmo MLLR se ha utilizado una parte del corpus2.

4.2. Resultados

La tabla 1 muestra las prestaciones para tres conjuntos diferentes de HMMs independientes del contexto. HMM1 son modelos entrenados a partir de los datos telefónicos. HMM2 son modelos entrenados a partir de la base de datos telefónica filtrada por una estimación de la respuesta al impulso del canal. HMMs adaptados son modelos obtenidos a partir de HMM2, los cuales se han adaptado a través de MLLR con 1, 5 y 8 clases de regresión.

Tabla 1

Corpus	HMM1	HMM2	HMMs Adaptados		
			1CR	5CR	8CR
C1	87.27%	89.84%	91.42%	**92.03%**	89.76%
C2	85.48%	92.26%	96.45%	**97.74%**	96.45%
C3	81.54%	88.98%	91.44%	**92.84%**	91.44%

5. MEDIDA DE CONFIANZA

Incluso en una aplicación como ésta, donde se asume la cooperación del usuario, se necesita una medida de confianza para reducir la tasa de falsa alarma y rechazar así palabras reconocidas erróneamente y eventos de fuera del vocabulario. Nuestro decodificador proporciona una medida de confianza basada en la relación entre la verosimilitud del comando reconocido y la verosimilitud de un reconocedor fonético para el mismo intervalo de tiempo.
La figura 2 muestra los histogramas de esta medida de confianza.

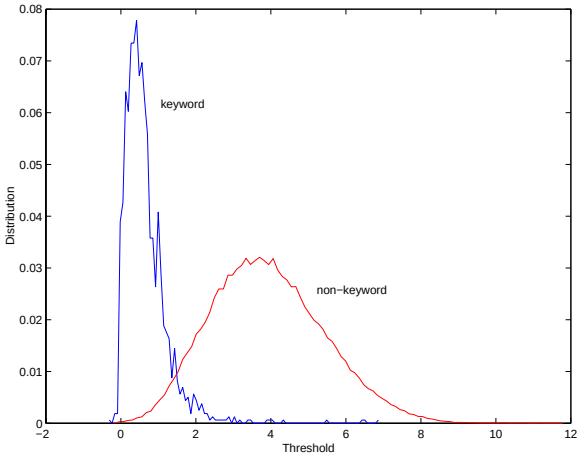

Figura 2: *Histogramas de la medida de confianza.*

6. CONCLUSIONES

En este artículo se han descrito tres componentes de un sistema IVR real que requiere especial atención en su etapa de diseño. Hemos probado la eficacia de técnicas de adaptación tanto en la mejora de prestaciones como en el ahorro de tiempo de desarrollo.

7. AGRADECIMIENTOS

Este trabajo ha sido parcialmente financiado por la CICYT bajo el proyecto 1FD97-0077-C02-01. Los autores agradecen a Francisco Méndez Pazó su trabajo en la programación de la aplicación.

8. REFERENCIAS

[1] J.L. Gauvain and L. Lamel., *"Large-Vocabulary Continuous Speech Recognition: Advances and Applications"*, Proceedings of the IEEE, 88(8):1181-1200, 2000.
[2] L. Villarrubia et al., *"VOCATEL AND VOGATEL: Two Telephone Speech Databases of Spanish Minority Languages (Catalan and Galician)"*, Workshop on Language Resources for European Minority Languages. Granada (Spain). 1998.
[3] C.H. Lee and Q. Huo., *"On Adaptive Decision Rules and Decision Parameter Adaptation for Automatic Speech Recognition"*, Proc. Of the IEEE, 88(8):1241-1269, 2000.
[4] A. Acero et al. *"Robust HMM-Based EndPoint Detector"*, Eurospeech93, pp. 1151-1554, 1993.
[5] S. Young et al. *"The HTK Book"*. January 1999.