# Reconocimiento Automático del Habla en vehículos, resultados con SpeechDat-Car

*Eduardo Lleida, Enrique Masgrau, Alfonso Ortega, Antonio Miguel, Luis Buera*

Instituto de Investigación en Ingeniería de Aragón
Universidad de Zaragoza
lleida@posta.unizar.es

## Resumen

En este artículo se presentan resultados de reconocimiento automático del habla en vehículos obtenidos utilizado la base de datos SpeechDat-Car en castellano. En concreto se presentan resultados sobre 5 tareas que incluyen secuencias de dígitos, números de teléfono, deletreo y comandos de control de dispositivos del vehículo. En estos experimentos, utilizando una parametrización estándar (Mel-cepstrum), se han estudiado las prestaciones del sistema para diferentes modelos acústicos, incorporando modelos acústicos genéricos para representar las palabras no contenidas en el vocabulario de reconocimiento.

## 1. Introducción

Cada día es mayor el número de aplicaciones *Telemáticas* que son incorporadas en los automóviles. El término "*Telematics"* se utiliza para indicar el conjunto de aplicaciones informáticas que tienen como base las telecomunicaciones y que aportan un valor añadido al automóvil. De este modo, el vehículo no es más que otro punto de acceso móvil a los sistemas telemáticos; productos, servicios y sistemas que implican acceso a la información, comunicaciones y entretenimiento. Sin embargo, a diferencia de otras formas de acceso a la información, en el caso de vehículos, prima la seguridad en la conducción lo que introduce serias restricciones en la forma de presentar la información al usuario. La investigación en "Interfaces Inteligentes" es de vital importancia para la introducción segura de las tecnologías de la información y las comunicaciones en un sector con amplia demanda como es la automoción.

Desde el punto de vista de la seguridad, el reconocimiento automático del habla juega un papel fundamental como sistema de entrada. Sin embargo,por razones de seguridad y confort, es necesario utilizar micrófonos situados en campo lejano, la señal de entrada al reconocedor adolece de una relación señal a ruido muy baja y está afectada por diversos tipos de ruido.

Los objetivos de este trabajo han sido varios. Por un lado estudiar como se ven afectadas las tasas de error para las distintas posiciones de micrófonos utilizadas en SpeechDat-Car. Por otro lado se ha experimentado con modelos acústicos dependientes e independientes del contexto. Un último objetivo ha sido estudiar la utilización de modelos acústicos de grupos silábicos para el modelado de palabras no contenidas en el vocabulario de reconocimiento.

Este trabajo está financiado por el MCyT TIC2001-2812-C05-04 y TIC2002-04103-C03-01(parcialmente financiado con fondos FEDER)

## 2. La base de datos SpeechDat-Car

### 2.1. Características genéricas

La base de datos SpeechDat-Car [1] contiene grabaciones de banda ancha (60-7000 Hz) para sistemas que están instalados y operan en automóviles. Las grabaciones son multicanal, conteniendo 4 canales correspondientes a

- 1 micrófono de cercanía, utilizado como referencia ( de aquí en adelante lo referenciaremos como MIC-0)
- 3 micrófonos de campo lejano con las siguientes posiciones
  - techo del vehículo cerca de la ventanilla (MIC-1)
  - techo central enfrente del conductor (MIC-2)
  - techo cerca del retrovisor (MIC-3)

La base de datos contiene grabaciones de 306 locutores grabados en 600 sesiones. En todas las grabaciones se han utilizado los siguientes tipos de micrófonos

MIC-0: Shure SM-10 A
MIC-1: AKG Q400 Mk3 T
MIC-2: Peiker ME15/V520-1
MIC-3: AKG Q400 Mk3 T

La distribución de la relación señal a ruido para los cuatro canales sobre las 600 sesiones es la mostrada en la tabla I.

| SNR | MIC-0 | MIC-1 | MIC-2 | MIC-3 |
|---|---|---|---|---|
| 5-10 | 0 | 59 | 176 | 122 |
| 10-15 | 0 | 292 | 260 | 282 |
| 15-20 | 3 | 208 | 138 | 149 |
| 20-25 | 65 | 41 | 26 | 47 |
| 25-30 | 491 | 0 | 0 | 0 |
| 30-35 | 41 | 0 | 0 | 0 |

Tabla I. Distribución SNR con el número de sesiones para los cuatro micrófonos.

### 2.2. Tareas definidas

Se han definido 5 tareas distintas de reconocimiento sobre la base de datos SpeechDat-Car. Las tareas son las siguientes:

Tarea T1. Cadenas de dígitos.

Esta tarea está formada por secuencias de dígitos extraídas de: códigos PIN formados por secuencias de 6 dígitos, números de tarjeta de crédito formados por secuencias de 16 dígitos, dígitos aislados, secuencias de 10 dígitos con pausas entre ellos y secuencias de 6 dígitos conectados.

Siguiendo la recomendación de SpeechDat-Car se ha definido un corpus de entrenamiento y otro de test independiente del locutor. El corpus de entrenamiento está compuesto por 1065 ficheros y el corpus de test está formado por 654 ficheros

Tarea T2. Números telefónicos.

Esta tarea está formada por números de teléfono leídos. Cada locutor a pronunciado 3 números de teléfono leídos de forma natural.
Siguiendo la recomendación de SpeechDat-Car se ha definido un corpus de entrenamiento y otro de test independiente del locutor. El corpus de entrenamiento está compuesto por 398 ficheros y el corpus de test está formado por 246 ficheros

Tarea T3. Secuencias de letras.

Esta tarea está formada por secuencias de letras deletreadas.
El corpus de entrenamiento está compuesto por 795 ficheros y el corpus de test está formado por 492 ficheros

Tarea T4. Palabras de Aplicación Telefonía móvil

Esta tarea contiene palabras usuales en aplicaciones de telefonía móvil.
El corpus de entrenamiento está compuesto por 928 ficheros y el corpus de test está formado por 574 ficheros

Tarea T5. Control de servicios instalados en vehículos

Esta tarea contiene una lista de comandos típicos utilizados para controlar sistemas instalados en vehículos.
El corpus de entrenamiento está compuesto por 4522 ficheros y el corpus de test está formado por 2622 ficheros

## 3. El sistema de reconocimiento

Los experimentos se han realizado utilizando un sistema de reconocimiento basado en modelos ocultos de Markov con una parametrización estándar 12 coeficientes melcepstrum, sus primeras y segundas derivas y la derivada de la energía.

### 3.1. Tareas y gramáticas

El sistema de reconocimiento está dirigido por una gramática de estados finitos. Esta gramática es distinta según la tarea.
Para la tarea T1 (secuencia de dígitos) se ha definido una gramática que permite cualquier longitud de secuencia de dígitos.
Para la tarea T2 (números de teléfono) se ha definido una gramática que acepta la pronunciación de un número de teléfono de forma natural y sin utilizar información del número de dígitos contenido en el número de teléfono.
Para la tarea T3 (deletreo) la gramática permite cualquier concatenación de letras.
Para las tareas T4 (comandos teléfono) y T5 (comandos de aplicaciones) la gramática solo acepta los comandos definidos en las especificaciones de la base de datos.
Para todas las tareas se ha utilizado la misma penalización por inserciones.

### 3.2. Modelos Acústicos, entrenamiento

Se han realizado experimentos con dos tipos de modelos acústicos. Por un lado se han utilizado modelos acústicos independientes del contexto. En este caso se han utilizado 25 modelos acústicos que modelizan 25 fonemas del castellano mas un modelo para el silencio. Se han utilizado modelos ocultos de Markov con densidades de probabilidad continua y una estructura de 3 estados por fonema a excepción del modelo de silencio que es de 1 estado. Las funciones de densidad de probabilidad se ha modelizado con una mezcla de gaussianas que por estado puede llegar como máximo a 16 mezclas.

Las unidades contextuales se han definido dividiendo cada unidad fonética en su contexto izquierdo, una unidad sin contexto y su contexto derecho [2]. Cada nueva unidad se modeliza con un modelo oculto de Markov de un estado y un máximo de 16 gaussianas para la función de densidad de probabilidad. Con los 25 fonemas sin contexto mas el silencio, se obtienen 1326 unidades contextuales monoestado.
Por ejemplo, la palabra casa se transcribiría en unidades sin contexto como /k/ /a/ /s/ /a/ y en unidades contextuales como /#<k/ /k/ /k>a/ /k<a/ /a/ /a>s/ /a<s/ /s/ /s>a/ /s<a/ /a/ /a>#/ donde # indica cualquier contexto, < contexto izquierdo y > contexto derecho.
Para el entrenamiento de los modelos acústicos sin contexto se ha realizado el siguiente procedimiento. Se han utilizado los corpus de entrenamiento definidos para cada tarea de forma conjunta, es decir, no se han entrenado modelos dependientes de la tarea. Partiendo de modelos acústicos sin contexto entrenados con una base de datos genérica (frases fonéticamente balanceadas), se ha realizado una segmentación inicial de todo el corpus de entrenamiento correspondiente al micrófono de cercanía (MIC-0). Se ha utilizado el algoritmo EM con 4 iteraciones sobre toda la base de datos. Para el resto de micrófonos se ha tomado como modelos iniciales los obtenidos en el proceso anterior para el micrófono MIC-0. Para estos micrófonos se han realizado 4 iteraciones del algoritmo EM.
Para el entrenamiento de los modelos acústicos contextuales, se han tomado como modelos iniciales los modelos sin contexto, asignando el primer estado del modelo sin contexto al modelo contextual izquierdo, el segundo estado al modelo sin contexto y el tercer estado al modelo contextual derecho. A partir de esta asignación se ha realizado 4 iteraciones del algoritmo EM para cada micrófono. El número final de gaussianas para los modelos acústicos del MIC-0 son 1174 para el modelado sin contexto y 8656 para el contextual. Para el MIC-2 son 1125 gaussianas para el modelado sin contexto y 9088 para el contextual.

### 3.3. Grupos silábicos

Para modelar palabras no contenidas en el vocabulario de reconocimiento, en este trabajo utilizamos modelos acústicos de grupos silábicos [3]. Si hacemos una clasificación de los sonidos del castellano en 4 grupos: ‘v’ vocálicos, ‘s’ consonantes sonoras, ‘c’ consonantes sordas y ‘n’ nasales y líquidas, podemos cubrir mas del 96 % de las estructuras silábicas del castellano definiendo 16 grupos silábicos. Si **b** significa consonante y **a** vocal, el 96 % de las estructuras silábicas se corresponden con **a**, **ab**, **ba**, **bab** [4]. Los 16 grupos silábico definidos se muestran en la tabla II.

Para la inclusión de los grupos silábicos en el reconocedor se modifican las gramáticas de forma que se permita cualquier concatenación de grupos silábicos después de cada palabra del vocabulario.

| Estructura silábica | Grupo silábico |
|---|---|
| a | v |
| ab | vs, vn, vc |
| ba | sv, nv, cv |
| bab | svs, svn, svc, nvs, nvn, nvc, cvs, cvn, cvc |

Tabla II. grupos silábicos

En los experimentos que se presentan es este artículo, se han utilizados modelos ocultos de Markov con un número de estados igual a la longitud de la estructura silábica a que corresponde el grupo silábico. Para modelos de más de 1 estado se permite cualquier transición de izquierda a derecha entre estados. Se han utilizado 16 gaussianas por estado.

## 4. Resultados experimentales

Todos los experimentos, si no se indica lo contrario, están realizados utilizando las sesiones correspondientes a las siguientes condiciones acústicas del vehículo:
CLIMCONTROL=OFF,
AUDIO=OFF,
WINDOW_L_FRONT=CLOSE,
WINDOW_R_FRONT=CLOSE,
WINDOW_REAR=CLOSE,
ROOF=CLOSE,
WIPERS=OFF,
CROSS_TALK=NO
En cuanto a los escenarios, no se ha realizado distinción, se han tomado sesiones de los cuatro escenarios. Estos escenarios son “High Speed Good Road”, “Stop, motor running”, “Low speed rough road” y “town traffic”.

### 4.1. Dependencia con el micrófono

Un primer conjunto de experimentos sobre la tarea T1 (secuencia de dígitos) se ha realizado para tener una estimación de las tasas de errores para los tres micrófonos de campo lejano. De este experimento hemos seleccionado el mejor y el resto de resultados se dan comparando el micrófono cercano con el seleccionado en estos experimentos. En la tabla III se muestran los resultados en términos de tasa de aciertos. Claramente los resultados del MIC-2 son mejores que para el resto de micrófonos de campo lejano. En otras pruebas realizadas sobre otras tareas se obtiene la misma conclusión. Una de las razones de este mejor comportamiento puede ser debido a que el micrófono MIC-2 es de distinta marca, y posiblemente mejor calidad, que los MIC-1 y MIC-3 .
Para verificarlo, se ha estimado la densidad espectral de potencia de los segmentos de ruido de 5 ficheros tomados aleatoriamente para cada uno de los 4 escenarios definidos en la base de datos. En la figura 1 y figura 2 se muestran las densidades espectrales de potencia de ruido estimadas con un estimador de Welch para las condiciones “Stop, motor running” y “High speed good road”.
En ambos casos, en el margen de 1000 a 6000 Hz, la densidad espectral de potencia de ruido es menor en el micrófono MIC-2 que en los otros dos. Para los otros escenarios se vuelve a repetir el mismo resultado. Por esta razón, los resultados que vamos a presentar únicamente utilizan los micrófonos MIC-0 (cercano) y MIC-2. Cabe destacar como en baja frecuencia, por debajo de 500 Hz la densidad espectral de potencia de ruido es mayor en MIC-2 que en el resto.

| | **Entrenamiento** | | | |
|---|---|---|---|---|
| **Test** | **MIC-0** | **MIC-1** | **MIC-2** | **MIC-3** |
| **MIC-0** | 99,71 | | | |
| **MIC-1** | 90,48 | 96,52 | 96,24 | 96,10 |
| **MIC-2** | 92,83 | 97,72 | 98,23 | 97,51 |
| **MIC-3** | 89,12 | 96,63 | 95,52 | 95,62 |

Tabla III. Tasas de acierto

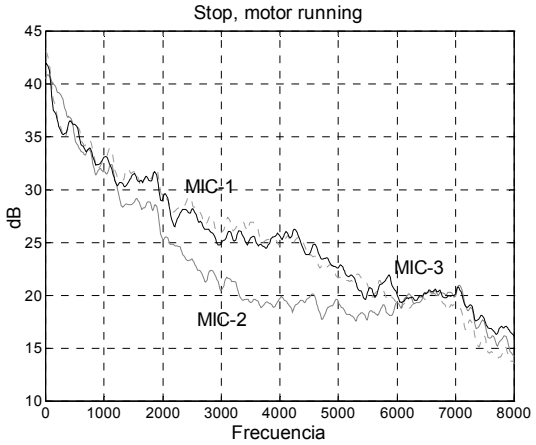


Figura 1. Densidad espectral de potencia de ruido para el escenario “Stop, motor running”.

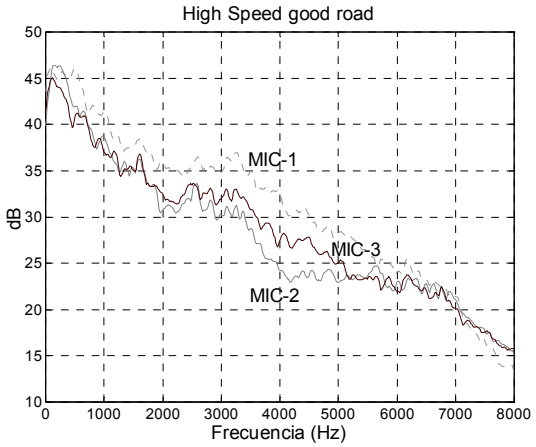


Figura 2. Densidad espectral de potencia de ruido para el escenario “High speed good road”.

### 4.2. Resultados tarea T1, secuencias de dígitos

En primer lugar mostramos los resultados obtenidos utilizando modelos sin contexto. En la tabla IV se muestran los resultados obtenidos en términos de la tasa de error, inserciones, borrados y sustituciones para la señal “limpia” (MIC-0) y para el micrófono lejano MIC-2. También se ha realizado reconocimientos con modelos desadaptados (entrenamiento con MIC-0 y reconocimiento con MIC-2).

| Test | Entrenamiento | Errores | INS | BOR | SUS |
|---|---|---|---|---|---|
| MIC-0 | MIC-0 | 1,75 | 0,79 | 0,55 | 0,41 |
| MIC-2 | MIC-2 | 5,34 | 1,49 | 2,42 | 1,43 |
| MIC-2 | MIC-0 | 19,64 | 6,42 | 3,97 | 9,25 |

Tabla IV. Resultados tarea T1 modelos sin contexto

En la tabla V se muestran los resultados obtenidos cuando se permite la inclusión de grupos silábicos entre dígitos

| Test | Entrenamiento | Errores | INS | BOR | SUS |
|---|---|---|---|---|---|
| MIC-0 | MIC-0 | 9,05 | 0,15 | 8,58 | 0,32 |
| MIC-2 | MIC-2 | 11,59 | 0,50 | 10,10 | 0,99 |

Tabla V. Resultados tarea T1 con grupos silábicos y modelos sin contexto

Las tablas VI y VII muestran los resultados con modelos contextuales. Utilizando los modelos monoestado contextuales, la tasa de error se reduce significativamente, consiguiéndose un mejor comportamiento cuando se utilizan grupos silábicos para modelar palabras fuera del vocabulario. Este resultado es coherente con el hecho de que los modelos contextuales al estar mas especializados dan unas ratios de probabilidades entre modelos contextuales y grupos silábicos más altos que entre modelos sin contexto y grupos silábicos. El coste computacional, por el hecho de tener mas gaussianas, se incrementa en un 36 % al utilizar modelos contextuales frente a sin contexto. Por otro lado se reduce a la mitad la tasa de error, tanto en señal limpia como en la contaminada. Cuando se reconoce con modelos acústicos del MIC-0 la señal captada por el MIC-2 la tasa de error aumenta muy significativamente, tal y como era de esperar. Este resultado lo hemos incorporado para poder estudiar en experimentos posteriores las mejoras que se pueden obtener con la adaptación de modelos acústicos o con las técnicas cancelación de ruido.

| Test | Entrenamiento | Errores | INS | BOR | SUS |
|---|---|---|---|---|---|
| MIC-0 | MIC-0 | 1,02 | 0,73 | 0,14 | 0,15 |
| MIC-2 | MIC-2 | 2,51 | 0,85 | 0,87 | 0,79 |
| MIC-2 | MIC-0 | 10,62 | 4,49 | 1,46 | 4,67 |

Tabla VI. Resultados tarea T1 con modelos contextuales.

| Test | Entrenamiento | Errores | INS | BOR | SUS |
|---|---|---|---|---|---|
| MIC-0 | MIC-0 | 1,28 | 0,12 | 1,08 | 0,09 |
| MIC-2 | MIC-2 | 3,26 | 0,14 | 2,62 | 0,50 |

Tabla VII. Resultados tarea T1 con modelos contextuales y grupos silábicos

### 4.3. Resultados tarea T2, números telefónicos

Las tablas VIII, IX y X muestran los resultados sobre la tarea T2 utilizando modelos sin contexto, modelos con contexto y modelos con contexto más grupos silábicos respectivamente. Destaca el aumento significativo en la tasa de error, con un gran desajuste en la tasa de inserciones y borrados, sobre todo con modelos sin contexto. Se trata de una tarea con un vocabulario de 48 palabras, donde aparecen muchos errores del tipo “seis ciento” en lugar de “seiscientos” o “diez siete” por “diecisiete”, etc.

Destaca de nuevo la reducción a la mitad en la tasa de errores al pasar de modelos sin contexto a modelos contextuales monoestado con un incremento en el tiempo de cálculo del 40 %.

| Test | Entrenamiento | Errores | INS | BOR | SUS |
|---|---|---|---|---|---|
| MIC-0 | MIC-0 | 22,10 | 1,62 | 10,04 | 10,44 |
| MIC-2 | MIC-2 | 33,21 | 1,30 | 17,56 | 14,35 |
| MIC-2 | MIC-0 | 49,43 | 1,94 | 24,55 | 22,93 |

Tabla VIII. Resultados tarea T2 con modelos sin contexto

| Test | Entrenamiento | Errores | INS | BOR | SUS |
|---|---|---|---|---|---|
| MIC-0 | MIC-0 | 10,87 | 2,53 | 3,36 | 4,98 |
| MIC-2 | MIC-2 | 17,75 | 1,98 | 7,79 | 7,99 |
| MIC-2 | MIC-0 | 33,81 | 3,08 | 12,81 | 17,91 |

Tabla IX. Resultados tarea T2 con modelos con contexto

| Test | Entrenamiento | Errores | INS | BOR | SUS |
|---|---|---|---|---|---|
| MIC-0 | MIC-0 | 12,97 | 2,02 | 6,17 | 4,78 |
| MIC-2 | MIC-2 | 20,88 | 1,42 | 12,69 | 6,76 |

Tabla X. Resultados tarea T2 con modelos con contexto y grupos silábicos

### 4.4. Resultados tarea T3, deletreo

La tarea de deletreo tiene un vocabulario de 30 palabras y en principio es la tarea más difícil de todas por la similitud acústica entre palabras. Las tablas XI, XII y XIII muestran los resultados obtenidos con modelos sin contexto, contextuales y con grupos silábicos respectivamente. De nuevo destaca la mejora significativa obtenida con los modelos contextuales monoestado y la mínima influencia de los grupos silábicos sobre la tasa de error. El incremento de cálculo por utilizar modelos contextuales es del 36 %.

| Test | Entrenamiento | Errores | INS | BOR | SUS |
|---|---|---|---|---|---|
| MIC-0 | MIC-0 | 23,73 | 1,55 | 7,30 | 14,87 |
| MIC-2 | MIC-2 | 37,39 | 1,41 | 17,51 | 18,47 |
| MIC-2 | MIC-0 | 48,84 | 3,06 | 13,24 | 32,54 |

Tabla XI. Resultados tarea T3 con modelos sin contexto

| Test | Entrenamiento | Errores | INS | BOR | SUS |
|---|---|---|---|---|---|
| MIC-0 | MIC-0 | 11,71 | 2,05 | 1,62 | 8,03 |
| MIC-2 | MIC-2 | 16,66 | 2,45 | 3,20 | 11,0 |
| MIC-2 | MIC-0 | 34,80 | 3,30 | 5,63 | 25,87 |

Tabla XII. Resultados tarea T3 con modelos contextuales

| Test | Entrenamiento | Errores | INS | BOR | SUS |
|---|---|---|---|---|---|
| MIC-0 | MIC-0 | 12,86 | 1,32 | 4,08 | 7,47 |
| MIC-2 | MIC-2 | 17,58 | 1,79 | 5,54 | 10,25 |

Tabla XIII. Resultados tarea T3 con modelos contextuales y grupos silábicos

### 4.5. Resultados tarea T4, comandos teléfono

La tarea T4 está formada por un vocabulario de 70 palabras que definen los comandos de control de un teléfono móvil en un vehículo. Las tablas XIV, XV y XVI muestran los resultados para modelos sin contexto, contextuales y grupos silábicos respectivamente. El aumento de la carga computacional al pasar de modelos sin contexto a contextuales es del orden del 80 %, sin embargo, la reducción en la tasa de error para el MIC-2 es un factor mayor de 5. La introducción de grupos silábicos no modifica las tasas de error.

| Test | Entrenamiento | Errores | INS | BOR | SUS |
|---|---|---|---|---|---|
| MIC-0 | MIC-0 | 1,32 | 0,36 | 0,12 | 0,84 |
| MIC-2 | MIC-2 | 3,84 | 0,12 | 1,44 | 2,28 |
| MIC-2 | MIC-0 | 14,76 | 0 | 5,28 | 9,48 |

Tabla XIV. Resultados tarea T4 con modelos sin contexto

| Test | Entrenamiento | Errores | INS | BOR | SUS |
|---|---|---|---|---|---|
| MIC-0 | MIC-0 | 0,12 | 0 | 0 | 0,12 |
| MIC-2 | MIC-2 | 0,72 | 0,12 | 0,24 | 0,36 |
| MIC-2 | MIC-0 | 8,28 | 0,48 | 2,88 | 4,92 |

Tabla XV. Resultados tarea T4 con modelos con contexto

| Test | Entrenamiento | Errores | INS | BOR | SUS |
|---|---|---|---|---|---|
| MIC-0 | MIC-0 | 0,12 | 0 | 0 | 0,12 |
| MIC-2 | MIC-2 | 0,72 | 0,12 | 0,24 | 0,36 |

Tabla XVI. Resultados tarea T4 con modelos con contexto y grupos silábicos.

### 4.6. Resultados tarea T5, comandos aplicaciones

La tarea T5 está formada por un vocabulario de 152 palabras que definen los comandos de control ciertas aplicaciones de a bordo de un vehículo (p.e. el sistema de navegación, el equipo de música, etc.). Las tablas XVII, XVIII y XIX muestran los resultados para modelos sin contexto, contextuales y grupos silábicos respectivamente. El aumento de la carga computacional al pasar de modelos sin contexto a contextuales es del orden del 70 %. La reducción en la tasa de error para el MIC-2 es un factor mayor de 2. La introducción de grupos silábicos modifica levemente las tasas de error. En esta tarea existen 12 apariciones de palabras no contenidas en el vocabulario de reconocimiento.

Hasta ahora todos los experimentos se han realizado con el climatizado apagado. Para ver como afecta el climatizador, se han realizado reconocimientos para esta tarea T5 con las sesiones donde el climatizado estaba conectado. El corpus de test está formado por 1528 ficheros. La tabla XX y XXI muestra los resultados obtenidos utilizando modelos sin contexto y contextuales. Se aprecia un mayor aumento de la tasa de error en el caso de los modelos contextuales frente a los sin contexto.

| Test | Entrenamiento | Errores | INS | BOR | SUS |
|---|---|---|---|---|---|
| MIC-0 | MIC-0 | 3,79 | 0,07 | 1,55 | 2,16 |
| MIC-2 | MIC-2 | 8,27 | 0,07 | 3,69 | 4,50 |
| MIC-2 | MIC-0 | 26,31 | 0,02 | 10,93 | 15,36 |

Tabla XVII. Resultados tarea T5 con modelos sin contexto

| Test | Entrenamiento | Errores | INS | BOR | SUS |
|---|---|---|---|---|---|
| MIC-0 | MIC-0 | 1,62 | 0,07 | 0,42 | 1,13 |
| MIC-2 | MIC-2 | 2,83 | 0,05 | 1,08 | 1,69 |
| MIC-2 | MIC-0 | 16,61 | 0,10 | 6,00 | 10,50 |

Tabla XVIII. Resultados tarea T5 con modelos con contexto

| Test | Entrenamiento | Errores | INS | BOR | SUS |
|---|---|---|---|---|---|
| MIC-0 | MIC-0 | 1,91 | 0,05 | 1,05 | 0,81 |
| MIC-2 | MIC-2 | 3,30 | 0,05 | 1,94 | 1,30 |

Tabla XIX. Resultados tarea T5 con modelos con contexto y grupos silábicos

| Test | Entrenamiento | Errores | INS | BOR | SUS |
|---|---|---|---|---|---|
| MIC-0 | MIC-0 | 3,95 | 0,75 | 1,62 | 1,58 |
| MIC-2 | MIC-2 | 11,36 | 0,75 | 5,20 | 5,41 |

Tabla XX. Resultados tarea T5 con el climatizador funcionando y modelos sin contexto

| Test | Entrenamiento | Errores | INS | BOR | SUS |
|---|---|---|---|---|---|
| MIC-0 | MIC-0 | 2,71 | 0,83 | 0,79 | 1.08 |
| MIC-2 | MIC-2 | 4,58 | 0,83 | 1,37 | 2,37 |

Tabla XXI. Resultados tarea T5 con el climatizador funcionando y modelos contextuales.

### 4.7. Frases de activación y grupos silábicos

Una de las funciones de los grupos silábicos es ayudar a rechazar palabras no contenidas en el vocabulario de reconocimiento. En la base de datos SpeechDat-Car se han definido un conjunto de 5 frases de activación:

*Llamar por teléfono*
*Finalizar la llamada*
*Seleccionar un número*
*Seleccionar una persona*
*Contestar la llamada*

Tomando como señal de test la captada por el MIC-2, la tasa de error utilizando modelos con contexto entrenados con MIC-2 es del 0,81 % (0,61 % de borrados) sobre 164 ficheros. Si introducimos los grupos silábicos y aumentamos los ficheros de test con los de las tareas anteriores, la tasa de error aumenta al 2,64 % (2,44 % de borrados) pero no se detecta ninguna frase de activación en el resto de ficheros. Este resultado nos muestra las buenas perspectivas que tiene la utilización de grupos silábicos para la activación oral del sistema de reconocimiento de un vehículo.

## 5. Conclusiones

En este artículo se han presentado los resultados básicos que se obtienen en distintas tareas definidas sobre la base de datos SpeechDat-Car para el reconocimiento automático del habla en el interior de vehículos. Se han presentado resultados utilizando modelos acústicos sin contexto, modelos acústicos monoestado con contexto y la utilización de grupos silábicos para representar palabras no contenidas en el vocabulario de reconocimiento. De los resultados se pueden extraer diversas conclusiones. En primer lugar, para SpeechDat-Car, aunque las relaciones señal a ruido de los tres micrófonos lejanos son similares, la señal captada por el

micrófono central (MIC-2) da unas tasas de reconocimiento significativamente superiores a los otros dos micrófonos de campo lejano. En cuanto a los modelos acústicos, claramente la utilización de los modelos monoestado contextuales reducen en factores superiores a 2 las tasas de error para todas las tareas. La introducción de grupos silábicos permite dar mas robustez al sistema frente locuciones que no contengan palabras del vocabulario de reconocimiento, sin degradar significativamente las tasas de error.

Estos resultados pretenden ser una base para tener una referencia de las tasas de error que un reconocedor estándar obtiene reconociendo diversos vocabularios en el interior de un vehículo. El paso siguiente es estudiar métodos de adaptación on-line y de análisis robusto que permitan reducir las tasas de error obtenidas con el micrófono de campo lejano MIC-2, acercándolas o mejorando a las del micrófono cercano MIC-0.

## 6. Referencias


[1] Moreno, Asunción, Noguiera, Albino, Sesma, Alberto., "SpeechDat-Car: Spanish", *Technical Report SpeechDat.*

[2] Batlle I Mont, Eloi, *Modelatge acústic adaptatiu per al reconeixement de la parla,* Tesis doctoral,UPC, Diciembre 1999.

[3] Lleida E., Mariño J.B., Slavedra J., Bonafonte A., "Syllabic Filler for Spanich HMM Keyword Spotting", Proc. ICSLP 92, pp 1-4, 1992

[4] Martínez Celdrán, E., *Fonética*, Ed. Teide, Barcelona, 1984.