Reconocedor de habla continua independiente del contexto para el español de Argentina

Pedro Univaso
Facultad de Ingeniería, Universidad Austral. punivaso@austral.edu.ar

Jorge A. Gurlekian
Laboratorio de Investigaciones Sensoriales Facultad de Medicina, U.B.A. jag@fmed.uba.ar

Diego A. Evin
Laboratorio de Investigaciones Sensoriales Facultad de Medicina, U.B.A. devin@gmail.com

Recibido: 16-03-2009, aceptado 15-06-2009, versión final: 15-06-2009

Resumen

En este trabajo se presentan los resultados de los experimentos llevados a cabo con un sistema de reconocimiento automático de habla continua para el español de Argentina. El reconocedor implementado basado en palabras utilizó unidades independientes del contexto, denominadas en la literatura "monofonos", como unidades básicas del modelo acústico. Para la creación de dichos modelos se emplearon modelos ocultos de Markov HMM (Hidden Markov Models) de 3 estados de izquierda a derecha del tipo semi-continuo "SC-HMM" asociados a cada uno de los 31 monofonos (30 fonemas + alófonos y un modelo de silencio). La base de datos acústica estuvo conformada por 741 oraciones con 2.837 palabras distintas, que cubren el 97% de las sílabas del español, emitidas en una cámara acústica por dos locutores profesionales. Los valores óptimos de los parámetros fueron seleccionados para maximizar la tasa de reconocimiento y simultáneamente reducir el tiempo de procesamiento. La tasa de reconocimiento promedio obtenida (%Acc), empleando la metodología de "validación cruzada de 10 particiones", fue del 97.87% con una tasa de tiempo real (%RT) del 34.98%.

***Palabras claves**: reconocimiento automático de habla, español de Argentina, tasa de tiempo real, modelos ocultos de Markov.*

Abstract

This paper presents the results obtained with a continuous speech recognition system for Argentine Spanish. The word-based recognizer used context-independent units, monophones, as basic units of the acoustical model. Modeling used three left-to-right states semi-continuous Hidden Markov Models SC-HMM associated to 31 monophones (30 phonemes and a silence model). The acoustical database included 741 sentences with 2837 different words –97% of Spanish syllables– recorded at an acoustic chamber by two professional announcers. The optimal values of the recognizer parameters were selected in order to maximize the recognition rate and simultaneously to reduce the execution time. The optimal average accuracy (%Acc) obtained, using 10-fold cross validation method, was 97.87% with a Real-Time Rate (%RT) of 34.98%.

***Key words**: Automatic Speech Recognition, ASR, Spanish of Argentina, Real-Time Ratio, HMM.*

1. INTRODUCCIÓN

Una de las condiciones más favorables para el reconocimiento automático de habla es el caso de las palabras aisladas, para el cual se han logrado, para el Español de Argentina, desempeños muy elevados > 99%, como lo muestran los trabajos que utilizan tanto técnicas de comparación de patrones (Univaso *et al*., 1986a, 1986b; Rosso *et al*., 1987) como HMM (Gurlekian *et al*., 1989, 1990) y donde se estudia como incorporar el conocimiento lingüístico en las condiciones iniciales de los modelos. Como paso intermedio al reconocimiento en habla continua, se ha estudiado extensamente la influencia del contexto en el reconocimiento de fonemas oclusivos del español utilizando HMM (Franco *et al*., 1987; Franco, 1988, 1989, 1990). Actualmente y después de 30 años de investigación, el desafío sigue siendo el reconocimiento de habla continua espontánea y con un vocabulario ilimitado.

Sin embargo, no se encuentran muchos trabajos sobre reconocimiento de habla continua en Español (Villarubia *et al*., 1997; Zhan *et al*., 1996) y en Español de Argentina (de la Torre *et al*., 1996) de los que se puedan comparar resultados sobre tasas de reconocimiento y tiempos de procesamiento, como los realizados para el idioma inglés (Ravinshakar, 1996) los cuales además de comparar dichas figuras de mérito adicionan la capacidad de memoria de almacenamiento de los modelos acústicos.

El empleo de reconocedores de habla de uso libre permite, además, comparar los valores de los parámetros empleados por diversos autores para el reconocimiento de habla en el mismo o diferentes idiomas. Existe en la literatura un gran número de trabajos que emplean el HTK Toolkit por ser de libre disponibilidad para la investigación (Wiggers, 2001; Ragni, 2007). El desarrollo de HTK lo lleva adelante el grupo de habla, visión y robótica del Departamento de Ingeniería de la Universidad de Cambridge (Young *et al*., 2006), como una herramienta de implementación de modelos ocultos de Markov a ser empleada primariamente en el ámbito del reconocimiento de habla, aunque ha sido utilizado en otras aplicaciones incluyendo la investigación de síntesis de voz, reconocimiento de caracteres y secuenciamiento de ADN.

El empleo del sistema HTK y de bases de datos comparables para la lengua Española (Tapias *et al*., 1994; Casacuberta *et al*., 1991) nos permitirá evaluar el desempeño en el reconocimiento respecto de las diferentes pruebas estándar.

Como trabajo inicial presentamos aquí, pruebas de referencia para habla continua, leída, en condiciones de laboratorio, con bajo ruido, gran ancho de banda y con dependencia del hablante utilizando el sistema HTK. En próximas publicaciones analizaremos el desempeño en situaciones más críticas: el reconocimiento en condiciones de habla telefónica y con independencia del hablante en bases de datos apropiadas como la obtenida en el proyecto SALA *Speech Across Latin America* (Gurlekian *et al*., 2001b) y en grandes bases de habla espontánea.

El presente trabajo está organizado de la siguiente manera: en la Sección 2 se analizará el diseño de la base de datos acústica empleada en los experimentos; en la Sección 3 se describe la metodología empleada para el entrenamiento de los modelos y la etapa de reconocimiento, en la Sección 4 se presentan los resultados obtenidos de acuerdo a los diferentes valores de los parámetros del reconocedor y en la Sección 5 se discuten dichos resultados y se presentan las futuras líneas de trabajo.

2. BASE DE DATOS ACÚSTICA

La base de datos acústica de Español de Argentina, denominada SECYT, con la cual se realizaron los experimentos de este trabajo (Gurlekian *et al*., 2001a) fue desarrollada en el LIS con el objetivo de realizar estudios de entonación. Se seleccionaron 741 frases –compuestas por un total de 5.281 palabras, correspondientes a un vocabulario de 2.837 palabras distintas– las cuales cubren el 97% de todas las sílabas del español en las dos condiciones de acento y en todas las posiciones dentro de la palabra.

Las grabaciones fueron realizadas en una cámara acústica (Nivel de ruido de 30dB SPL y Tiempo de Reverberación de 0.2 segundos) a una frecuencia de muestreo de 16 KHz. y con 16 bits.

Dos locutores profesionales, un hombre y una mujer, nativos de Buenos Aires, grabaron en forma independiente las mismas 741 frases del corpus. Se ha de notar que el Español de Buenos Aires es el más frecuentemente empleado en los medios de comunicación masiva en Argentina. Dada la experiencia de los locutores, se les pidió que generaran variaciones de los parámetros de entonación, de manera de prevenir la monotonía en las emisiones generando movimientos tonales variados y naturales.

Posteriormente la base acústica fue etiquetada por cuatro fonoaudiólogos, empleando un software de análisis de señales de habla desarrollado en el Laboratorio de Investigaciones Sensoriales y denominado ANAGRAF (Gurlekian, 1997), el cual muestra en forma sincronizada el espectrograma, la forma de onda, el perfil de energía, el contorno de la frecuencia fundamental F0 y trece niveles de etiquetado. Estos son: fonético, grafémico, de pausas, de acentos tonales, de acentos de frase y de juntura, de parámetros acústicos, de misceláneas, de partes de habla y cuatro niveles sintácticos, de manera de permitir una caracterización del texto y del audio más precisa.

La base de datos acústica quedó finalmente conformada por las señales acústicas digitalizadas y las transcripciones fonéticas, fonémicas y prosódicas con su correlato temporal asociado.

3. METODOLOGÍA EMPLEADA

El sistema de reconocimiento se desarrolló en base al set de herramientas de modelos ocultos de Markov elaborado por la Universidad de Cambridge: HTK Toolkit ver. 3.4 (Young *et al*., 2006), de uso libre para usos académicos.

El reconocedor se implementó en una computadora personal con 512MB DDR, bajo el sistema operativo LINUX.

3.1. Procesamiento de la señal de habla

La digitalización de la señal acústica se realizo a una frecuencia de muestreo de 16 KHz y con 16 bits, con sustracción de la media temporal, de manera de eliminar cualquier offset proveniente de la etapa de grabación analógica.

Posteriormente, se empleó una ventana de análisis del tipo Hamming de 25ms a una frecuencia de ventaneo de 10 ms, empleando un filtro de preénfasis con coeficiente 0.97, habiéndose normalizado la energía de la frase y empleando una escala logarítmica para la energía.

Los parámetros empleados para la creación de los modelos fueron la energía y 12 coeficientes MFCC: *Mel-Frecuency Cepstral Coefficients,* a los cuales se les adicionó la energía delta y la aceleración, conformando un total de 39 parámetros.

3.2. Unidad de habla

La selección de la unidad de habla a emplear en este trabajo fue el monofono, para lo cual se empleo el alfabeto fonético SAMPA *Speech Assesment Methods: Phonetic Alphabet,* adaptado para el Español de Argentina (Gurlekian *et al*., 2001b). A las 30 unidades fonéticas del alfabeto SAMPA formadas con los 22 fonemas del español y 8 alófonos de uso frecuente en Argentina, se le adicionó un modelo de silencio, presente generalmente entre frases, al cual se le asoció un modelo de pausa corta entre palabras, completando así un total de 31 monofonos para representar el habla del Español de Argentina.

3.3. Etapa de entrenamiento

El entrenamiento de los modelos acústicos siguió la metodología propuesta por Young *et al*., 2006 consistente en:

1. La creación de un Modelo Oculto de Markov simple de 3 estados de izquierda a derecha para cada uno de los 31 monofonos, exceptuando la pausa corta que es asociada al estado central del modelo de silencio.

2. El incremento de la cantidad de mezclas de Gaussianas de manera que cada modelo de monofono sea representado por la suma pesada de "n" funciones de densidad de probabilidad (PDF) obteniéndose un modelo de mezclas Gaussianas continuas.

3. La transformación de los modelos anteriormente entrenados en modelos de mezclas Gaussianas asociadas, denominados en la literatura como modelos semi-continuos (SC-HMM) empleando un método de splitting con las mezclas Gaussianas anteriores.

4. El re-entrenamiento sucesivo de dichos modelos para obtener los HMM definitivos a ser empleados en la etapa de reconocimiento. Para la realización de todos los experimentos en la etapa de entrenamiento se utilizó un "ancho del haz de decodificación" de 250.

3.4. Etapa de reconocimiento

La etapa de reconocimiento utiliza la rutina HVite, un reconocedor de palabras de propósito general basado en el algoritmo de Viterbi, la cual optimiza la comparación entre una emisión de habla desconocida con una red conformada por los modelos HMM obtenidas en la etapa de entrenamiento y cuya secuencia representa una palabra del diccionario empleado, dando como resultado la transcripción de la emisión incógnita. El reconocedor realiza dicha comparación teniendo en cuenta tanto el modelo acústico (HMM) como el modelo de lenguaje.

El modelo de lenguaje estadístico del tipo *n-gram* con *n*=2 –denominado en la literatura *bigram*– estima la probabilidad de la secuencia de palabras considerando solamente 2 palabras contiguas. El mismo fue generado en base a las transcripciones de las 741 frases de la base de datos acústica, previamente etiquetadas y el diccionario empleado incluyó solamente las palabras diferentes de dicha base. En la Tabla 1 pueden verse las características de la gramática generada, de las cuales la "Perplejidad" ó *Perplexity* mide la bondad del modelo de lenguaje. Bajos valores de perplejidad representan mejores modelos de lenguaje. La perplejidad puede ser considerada como una medida promedio de la cantidad de palabras diferentes igualmente más probables que pueden seguir a cualquier palabra.

Tabla 1
Características del modelo de lenguaje empleado

Nº de Nodos = 2723 [1 nulo], Vocabulario = 2722
Entropía = 5.408873, Perplejidad = 42.484743
1000 Frases: promedio long = 9.2, min=2, max=40

Esta etapa del reconocedor posee diversos parámetros configurables, de los cuales se estudiaron la influencia de tres de ellos: el "factor de lenguaje", el "factor de penalización de palabras insertadas" y el "ancho del haz de decodificación", dejando inalterados el resto de los parámetros.

El "factor de lenguaje", cuyo valor estándar es 1.0, post-multiplica la probabilidad de verosimilitud de la red de palabras, el cual al incrementarse, amplifica la importancia del modelo de lenguaje con respecto al modelo acústico. En el caso de un factor de lenguaje nulo se considera en el reconocimiento sólo el modelo acústico.

El "factor de penalización de palabras insertadas", incrementa la probabilidad de que una nueva palabra pueda ser insertada al finalizar el reconocimiento de una; se mide en incrementos logarítmicos y con un valor default nulo. Un aumento de este factor incrementa la cantidad de palabras a ser reconocidas, llegándose a un extremo a partir de cual también se incrementa la cantidad de palabras insertadas, generando una disminución en la precisión del reconocedor.

El "ancho del haz de decodificación", restringe el crecimiento de la red de reconocimiento a aquellos modelos HMM cuyas probabilidades de verosimilitud caen dentro de un ancho de haz con respecto al modelo más probable. De esta manera, disminuyendo el ancho de haz se

procesan menos modelos y se reduce el tiempo de decodificación, aunque se reduce también el porcentaje de reconocimiento.

4. RESULTADOS

Para evaluar los resultados de reconocimiento de palabras se emplean generalmente algunas de las siguientes figuras de mérito: la Correctibilidad [*%Corr*], la Precisión [*%Acc*] y la Tasa de Error de Palabra [*%WER*] definidas como:

$$\%Corr = (N - D - S)/N \times 100\% = H/N \times 100\%$$

$$\%Acc = (N - D - S - I)/N \times 100\% = (H - I)/N \times 100\%$$

$$\%WER = 100 - \%Acc$$

Donde N es la cantidad total de palabras a reconocer, S es el numero de errores por substitución, D es el numero de errores por eliminación, I es el numero de errores de inserción y $H = (N - D - S)$.

De manera de poder comparar los diferentes resultados, en este trabajo se considerará a la Precisión [*%Acc*] como medida representativa del desempeño del reconocedor.

Para poder medir la velocidad del procesamiento del reconocedor se empleará la Tasa de Tiempo Real [*%RT*] definida como:

$$\%RT = T_{Real} / T_{Rec} \times 100$$

Donde T_{Real} es el tiempo real de duración de una frase y T_{Rec} es el tiempo de reconocimiento de dicha frase.

4.1. Parámetros del reconocedor

En esta primera parte se realizaron experimentos con las emisiones de la locutora femenina, correspondiente a la base de datos SECYT. Debido al carácter exploratorio de estos experimentos se segmentó la base de datos conformada por las 741 frases en dos particiones: una para entrenar al sistema (602 frases) y otra (139 frases) para la etapa de reconocimiento. Conociendo que esta metodología posee limitaciones, dado que introduce sesgos debidos a la evaluación del error en la partición particular elegida, para los experimentos finales se utilizó la metodología de validación cruzada de "n" particiones (ver 4.2) la cual permite realizar particiones múltiples de los datos para luego estimar el error en base al promedio sobre estas particiones.

La etapa de entrenamiento del reconocedor tiene la posibilidad de emplear como estrategias de entrenamiento inicial dos tipos de metodologías de inicialización diferentes según si las frases están etiquetadas: método de "segmentación uniforme" o no lo están: método *flat-start*. Para etiquetar una frase se debe adjuntar a la emisión acústica la transcripción ortográfica y la posición temporal de cada fono, tarea que debe realizar un experto fonetista empleando un visualizador de señales acústicas. En nuestro caso se utilizó el software ANAGRAF para el etiquetado completo de la base de datos SECYT.

El método de segmentación uniforme emplea el algoritmo de Viterbi partiendo de un modelo HMM prototipo para luego de sucesivas iteraciones converger en un modelo HMM inicializado.

Tabla 2
Comparación de estrategias de inicialización en la etapa de entrenamiento

Metodología	%Corr	%Acc
Segmentación uniforme	91,03	87,18
Flat-start	92,31	92,31

El método *flat-start* empleado en el caso de las frases no etiquetadas, las cuales no poseen marcas temporales entre fonemas, considera inicialmente un modelo HMM prototipo en el cual los fonemas se suponen equidistantes temporalmente unos de otros, siendo luego similar al método de segmentación el cual en las sucesivas iteraciones con Viterbi logra la convergencia en un modelo HMM inicializado.

Los resultados obtenidos (Tabla 2), muestran que el método *flat-start* supera en porcentaje de reconocimiento [%Acc] al del método de segmentación uniforme en un 5,1%. Es por eso que fue elegido este método de entrenamiento inicial para los futuros experimentos, lo cual permitirá reducir el trabajo de etiquetado previo de otras bases de datos a emplearse en futuros trabajos.

Otro de los parámetros que deben fijarse en el reconocedor en la etapa de entrenamiento es la cantidad de mezclas de Gaussianas continuas (PDFs) que mejor representan a cada estado de los modelos HMM. En la Fig. 1 podemos ver cómo con 6 PDFs se obtiene el máximo de %Acc (91,04%). Se debe tener en cuenta que el incremento de PDFs aumenta el tiempo de procesamiento y la memoria requerida para almacenar los modelos HMM correspondientes. Dado el alcance de este trabajo no se analizaron las causas por las cuales se produjo un mínimo en 8 PDFs, resultados que difieren de otras investigaciones similares (Ragni, A., 2007).

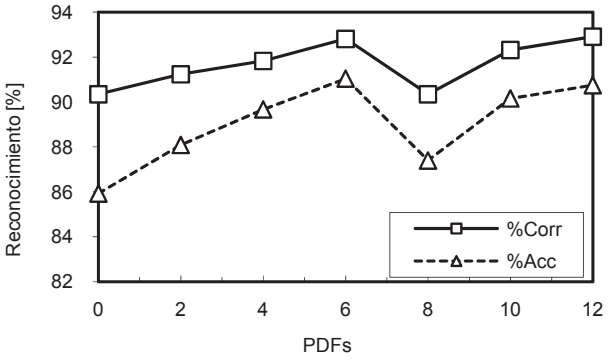

Figura 1. Porcentaje de reconocimiento para diferente cantidad de mezclas Gaussianas (PDFs)

Posteriormente (Tabla 3) se realizaron experimentos con modelos semi-continuos (SC-HMM) sin mezclas de Gaussianas previas y utilizando como mezclas asociadas (tied-mixtures) a cada uno de los monofonos, con una tasa de reconocimiento que mejoraba la anterior, obteniéndose un %Acc de 93,70%. Con el empleo de mezclas Gaussianas previas (6 PDFs) y posterior creación de modelos semi-continuos se logró el mejor valor de reconocimiento, obteniéndose un %Acc del 94,78%. Debido al incremento, en este último caso, del tiempo de procesamiento y de la memoria de almacenamiento de los HMM y dado que el incremento en la tasa de reconocimiento no es muy importante, se optó por el uso de los modelos semi-continuos sin mezclas previas para las siguientes etapas.

Tabla 3
Resultados del reconocimiento para diferentes metodologías de mezclas gaussianas

Mezclas Gaussianas	%Acc	%RT	Memoria [Kb]
6 PDFs	91,04	105,74	186,40
SC-HMM	93,70	130,24	89,60
6 PDFs + SC-HMM	94,78	289,18	105,60

La etapa de prueba se diseñó empleando los modelos HMM anteriormente generados, realizándose experimentos independientes de manera de ajustar los principales parámetros del reconocedor.

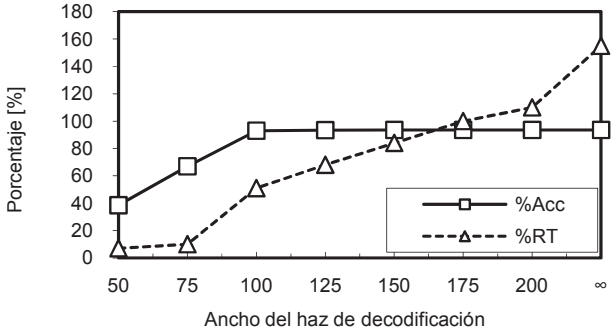

Figura 2. %Acc y %RT para diferentes valores del ancho del haz de decodificación

Los resultados obtenidos al estudiarse la influencia del "ancho del haz de decodificación" se muestran en la Fig. 2, en la cual se puede ver como se produce una estabilización en la tasa de reconocimiento %Acc en 92,93% a partir de un ancho de haz de 100, lográndose muy pequeños incrementos por sobre este valor, mientras que el tiempo de procesamiento aumenta considerablemente desde un %RT del 51,00%. En este experimento se empleó un factor de penalización de palabras insertadas nulo y un factor de lenguaje de 5.

En el caso de no utilizarse la restricción del haz de decodificación en el algoritmo de Viterbi (ancho del haz de decodificación = ∞ en Fig. 2), la tasa de reconocimiento toma su máximo valor (%Acc de 93,50%) pero también lo hace el tiempo de procesamiento, que en este caso es 1,5 veces el tiempo de emisión de la frase a reconocer.

De manera de poder capitalizar simultáneamente la mejor tasa de reconocimiento y un tiempo de procesamiento que permita desarrollar en el futuro un reconocedor de tiempo real, se seleccionó un ancho de haz de 100.

En la Fig. 3 puede verse como un "factor de lenguaje" de 10 es el que produce un máximo en la tasa de reconocimiento %Acc de 93,00%; siendo éste el valor empleado en el reconocedor final. En este experimento se empleó un factor de penalización de palabras insertadas nulo y un ancho de haz de 250.

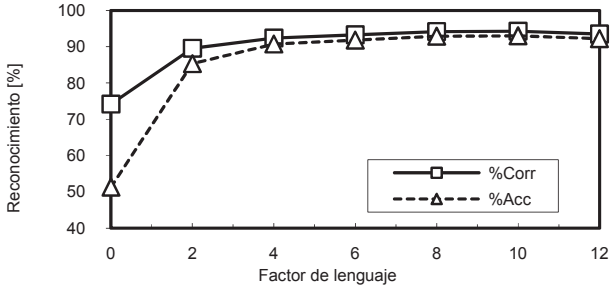

Figura 3. Porcentaje de reconocimento para diferentes valores del factor de lenguaje

Un valor nulo del factor de lenguaje elimina la influencia del modelo de lenguaje en el algoritmo de Viterbi haciendo que el reconocedor considere exclusivamente la información del modelo acústico. De los resultados obtenidos, para el reconocedor final, podemos inferir que un 55% del reconocimiento es debido al modelo acústico mientras que el modelo de lenguaje aporta el restante 45%.

En la Fig. 4 pueden verse los resultados obtenidos al utilizar diferentes valores en el "factor de penalización de palabras insertadas". El reconocedor final emplea un factor de 10 correspondiente a un máximo en %Acc de 93,50%. En este experimento se empleó un factor de penalización de palabras insertadas de 10 y un ancho de haz de 250.

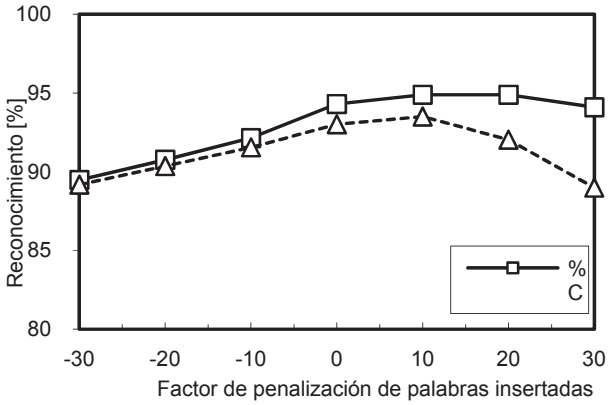

Figura 4. Porcentaje de reconocimiento para diferentes valores del factor de penalización de palabras insertadas

La mayor disminución de %Acc con respecto a %Corr, a medida que se incrementa el factor de penalización, es debida al incremento de inserciones de palabras en las frases reconocidas como puede verse en la Tabla 4.

Tabla 4
Ejemplo de inserción de palabras para diferentes valores del factor de penalización correspondiente

Frase emitida: *"Sus colocaciones son de mil pesos o aún inferiores"*		
Factor de penalización de palabras insertadas	Frase reconocida	Cantidad de inserciones (*I*)
0	*"Sus colocaciones son de mil pesos con aún inferiores"*	1
5	*"Sus colocaciones son de mil pesos con aún inferiores año"*	2
8	*"Sus colocaciones son de mil pesos con a o aún inferiores un año"*	4
10	*"Sus colocaciones son de mil pesos con a o aún inferiores es un año"*	5
11	*"Sus colocaciones son de mil pesos con a o aún inferiores es un año"*	5

4.2. Reconocedor final

En base a los resultados de la primera parte de los experimentos (Ver Sección 4.1), se implementó el reconocedor final con los siguientes parámetros:

Etapa de Entrenamiento

Ancho del haz de decodificación = 250
Método *flat-start*
Modelos semi-continuos (SC-HMM)

Etapa de Reconocimiento

Ancho del haz de decodificación = 100
Factor de lenguaje = 10
Factor de penalización de palabras insertadas = 10

Para analizar la performance del reconocedor final se utilizaron las emisiones de ambos hablantes de la base de datos SECYT, quedando el corpus comprendido por 1.482 frases con un total de 10.562 palabras (2.837 palabras distintas), siendo las mismas frases las emitidas por el hombre y la mujer. La duración total de la grabación fue de 83.59 minutos.

La base de datos fue segmentada empleando la metodología de validación cruzada de 10 particiones, considerando no incluir dentro de una partición las mismas emisiones de un hablante.

En la Tabla 5 pueden verse los resultados finales, para toda la base de datos SECYT, que arrojaron una tasa de reconocimiento %Acc de 97,87% y una tasa de tiempo real %RT de 34,98%.

Tabla 5
Resultados finales del reconocedor, empleando la metodología de validación cruzada de 10 particiones

Nº de partición	%Corr	%Acc	%RT
1	96,70	96,51	43,45
2	99,71	99,43	31,54
3	96,06	95,89	40,36
4	92,77	92,57	52,94
5	99,54	99,42	29,78
6	99,64	99,64	31,11
7	99,39	98,78	31,82
8	99,55	99,10	27,89
9	99,41	98,83	30,66
10	99,01	98,51	30,29
Promedio	98,18	97,87	34,98
Desviación Estándar	2,31	2,25	8,01

5. CONCLUSIONES

En este trabajo se presentaron los resultados de un sistema de reconocimiento de habla continua de palabras empleando la base de datos SECYT para el Español de Argentina.

Se comprobó que el empleo del método *flat-start* en la etapa de entrenamiento inicial permite incrementar la tasa de reconocimiento en un 5.1%, reduciendo a la vez las tareas de etiquetado previo de la base de datos.

El uso de unidades básicas del modelo acústico independientes del contexto permitió lograr porcentajes de reconocimiento (97,87%) similares a los que se publican en la literatura (98,50%) con unidades dependientes del contexto (Villarrubia *et al*., 1996). El uso de unidades independientes del contexto posibilitó el empleo de modelos semi-continuos (SC-HMM) asociados a cada uno de los monofonos. Esta solución produjo un incremento en la tasa de reconocimiento del 2.7% y una disminución del 51.9% en la memoria a pesar de incrementarse en un 24.5% el tiempo de procesamiento con respecto al uso de modelos de mezclas Gaussianas continuas simples.

Aunque se emplearon emisiones acústicas con un alto grado de calidad (bajo ruido y emitidas por locutores profesionales), los resultados mostraron que el modelo de lenguaje (45%) cumple un rol comparable con el modelo acústico (55%).

Si se requiere implementar un sistema de reconocimiento en tiempo real es necesario imponer restricciones al haz de decodificación en el algoritmo de Viterbi para lograr tasas de tiempo real (%RT) menores a la unidad, no observándose, en ese caso, disminución de la tasa de reconocimiento.

Con el empleo de los parámetros de ancho del haz de decodificación de 100, factor de lenguaje de 10 y factor de penalización de palabras insertadas de 10, se logró una tasa de reconocimiento del 97,87% y una tasa de tiempo real de 34,98%, la cual nos permitirá implementar un reconocedor en tiempo real en futuros desarrollos.

En el futuro también será necesario investigar el uso de unidades dependientes del contexto, como los trifonos, como unidades básicas acústicas y la extensión de los experimentos a multihablantes en canal telefónico con habla espontánea no leída.

RECONOCIMIENTOS

Los autores agradecen la colaboración del Dr. Juan M. Ale (Director del Laboratorio de Data Mining, Facultad de Ingeniería, Universidad Austral) en la corrección del presente documento.

REFERENCIAS

Casacuberta, F., García, R., Llisterri, J., Nadeu, C., Pardo, J.M., Rubio, A. (1991). *Development of Spanish Corpora for Speech Research (ALBAYZIN)*. In: Workshop on International Cooperation and Standardization of Speech Databases and Speech I/O Assesment Methods, Chiavari, Italy, 26-28 September.

de la Torre, C., Caminero-Gil, F.J., Álvarez, J., Martín del Alámo & C., Hernández-Gómez, L. (1996). *Evaluation of the Telefónica I+D Natural Numbers Recognizer over different Dialects of Spanish from Spain and America*. 4th International Conference on Spoken Language Processing, Philadelphia, PA, USA, October 3-6, 1996.

Franco, H.E., Gurlekian, J. A. (1987). *Context dependent recognition of Spanish Stops*, Academy of Sciences of the Estonian S.S.R. Institute of Language and Literature, Vol. 2, pp. 384-387, 1987

Gurlekian, J. A., Franco, H.E. and Santagada, M. (1990). *Speaker independent recognition of isolated Spanish digits.* Proceedings of the ICSLP´90. Kobe. Japan, Vol. 1, 529-532.

Gurlekian, J. A., Franco, H.E., Santagada; M. y Rosso, E. (1991). *Reconocimiento automático de dígitos con desempeño mayor a 99%, con independencia del hablante masculino*, Revista Telegráfica Electrónica. 1ra. y 2da. parte, Nro. 924: 93-96 y 925: 172-175 y 187.

Gurlekian, J. A., Rodriguez, H., Colantoni, L. & Torres, H. (2001a). *Development of a Prosodic Database for an Argentine Spanish Text to Speech System*. Proc. of the IRCS Workshop on Linguistic Databases (B.Bird and Liberman eds.) University of Pennsylvania, Philadelphia, USA, pp. 99-104.

Gurlekian, J. A., Colantoni, L., Torres, H., Rincón, A. Moreno A. y Mariño (2001b). *Database for an Automatic Speech Recognition System for Argentine Spanish*. Proc. of the IRCS Workshop on Linguistic Databases (B.Bird and Liberman eds.) Univ. of Pennsylvania, Philadelphia, USA, pp. 92-98.

Gurlekian, J.A., Colantoni, N. & Torres, H. (2001b). *El alfabeto fonético SAMPA y el diseño de córpora fonéticamente balanceados*. Fonoaudiológica. Editorial: ASALFA. Tomo: 47, Número: 3, pp. 58-69.

Gurlekian, J.A, (1997). *El Laboratorio de Audición y Habla del LIS, en Procesos Sensoriales y Cognitivos*. Editorial Dunken. Buenos Aires. Guirao M. (ed).

Ragni, A. (2007). *Initial Experiments with Estonian Speech Recognition*. In Proceedings of 16th Nordic Conference of Computational Linguistics, Tartu, Estonia.

Ravinshakar, M. (1996). *Efficient Algorithms for Speech Recognition*. Doctoral Thesis, School of Computer Science, Computer Science Division, Carnegie Mellon University, Pittsburgh.

Rosso E., Univaso P. y Franco H. (1987). *Reconocimiento Automático de Dígitos, Programación Dinámica*. Revista Telegráfica Electrónica.

Tapias, D., Acero, A., Estevez, J. & Torrecilla, J.C. (1994). *The VESTEL, Telephone Speech Database*. ICSLP-94, Japan, pp. 1811-1814.

Univaso, P., Rosso E., and Franco, H. E. (1986a). *Automatic recognition of isolated Spanish CV syllables*. Journal of the Acoustical Society of America. Volume 79, Issue S1, pp. S96-S96.

Univaso, P. y Rosso, E. (1986b). *Reconocimiento automático de dígitos*. Revista Telegráfica Electrónica Nro. 875, pp. 997-1009.

Villarrubia, L., Gómez, L.H., Elvira & J.M., Torrecilla, J.C. (1996). *Context-dependent Units for Vocabulary-independent Spanish Speech Recognition*. Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings, 1996 IEEE International Conference on Volume 1, Issue, 7-10 May 1996, Page(s): 451-454, Vol. 1.

Wiggers, P. (2001). *Hidden Markov Models for Automatic Speech Recognition and their Multimodal Applications*. Master Thesis, Delft University of Technology, The Netherlands.

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., Valtech, V. & Wooland, P. (2006). *The HTK Book*. Cambridge University Press.

Zhan, P., Ries, K., Gavalda, M., Gates, D., Lavie, A., Waibel, A. (1996). *JANUS-II: Towards Spontaneous Spanish Speech Recognition*. Spoken Language, 1996. ICSLP 96. Proceedings, Fourth International Conference on Volume 4, Issue, 3-6 Oct 1996, Page(s): 2285-2288, Vol. 4.