Sistema de Reconocimiento de Habla en Español con adaptación al discurso

Sebastián Perez, Patricia Pelle, Claudio Estienne y Francisco Messina

Universidad de Buenos Aires, Facultad de Ingeniería, Inst. de Ingeniería Biomédica
sgperez@fi.uba.ar, ppelle@fi.uba.ar, cestien@fi.uba.ar, fmessina@fi.uba.ar,

Resumen Este trabajo presenta un sistema de reconocimiento automático de habla en idioma español de alto desempeño diseñado para cumplir con dos objetivos. En primer lugar, lograr tasas de reconocimiento que sean comparables a los sistemas que son estado del arte en su tipo. En segundo lugar, evaluar el desempeño de un nuevo método de estimación de modelos de lenguaje propuesto en un trabajo anterior por nuestro grupo. Los resultados muestran un porcentaje de reconocimiento cercano al 90 % para un vocabulario de 5000 palabras, lo cual es del mismo orden que otros resultados reportados para sistemas de similares características pero en idioma inglés. También se verificó que el modelo de lenguaje basado en el estimador propuesto por nosotros mejora significativamente el desempeño del sistema comparado con otros dos modelos de lenguaje implementados con los mejores algoritmos conocidos.

Keywords: Reconocimiento de habla, Adaptación de Modelos de Lenguaje

1. Introducción

La tarea de reconocimiento automático de habla consiste en encontrar la transcripción escrita de una frase emitida por un hablante utilizando algoritmos y modelos implementados en programas de computadora. Si bien existe gran interés en encontrar sistemas confiables que puedan realizar esta tarea, los resultados obtenidos hasta el momento son de aplicación restringida y los porcentajes de reconocimiento obtenidos en general son bastante variables e insatisfactorios. La señal de habla es una serie temporal de gran variabilidad acústica, la duración de los sonidos que componen cada fonema [1] es diferente aún dentro de una misma palabra. Las características acústicas de cada fonema varían de acuerdo a la posición que el mismo ocupa en la palabra. A esto debe sumarse la variabilidad que existe en las características acústicas entre distintos hablantes y entre un mismo hablante pronunciando la misma palabra en diferentes instancias, más

[1] Un fonema es el sonido de menor duración que una persona puede distinguir como diferente a otro en un determinado idioma. Por ejemplo, la pronunciación de la letra "a" es percibida en forma diferente a la letra "e", por lo tanto por lo tanto a ambos sonidos les asignamos el nombre de fonemas /a/ y /e/ respectivamente.

los problemas que puede traer aparejado la degradación de la señal por el ruido ambiente o el canal de transmisión o grabación.

Los métodos modernos más exitosos que atacan el problema de la transcripción de habla se basan en su gran mayoría en modelos estadísticos entrenados con locuciones de múltiples hablantes. Con el fin de reducir la enorme variabilidad de la señal acústica, la misma, luego de ser digitalizada, es convertida en una secuencia de vectores llamados vectores acústicos o evidencia acústica, que refleje sus propiedades espectrales. El resultado es una secuencia $Y = \{y_1, y_2, \cdots, y_m\}$ de vectores que contienen la información esencial de la señal acústica de habla que nos permitirá distinguir entre los diferentes fonemas pronunciados. Matemáticamente el problema del reconocimiento de una frase se puede plantear del siguiente modo ([6]): supongamos que $W = \{w_1, w_2, \cdots, w_n\}$ es una secuencia de n palabras pertenecientes a un determinado vocabulario fijo y conocido $\mathcal{V}$ (por ejemplo una frase), y que se corresponde con la emisión de una secuencia de vectores acústicos Y. Cuando decimos que se corresponde nos referimos a que la secuencia de vectores acústicos Y se origina cuando el hablante emite la frase W. Si $P(W/Y)$ corresponde a la probabilidad de que se haya pronunciado la frase W cuando se observó la evidencia acústica Y, entonces para decidir entre todas las frases posibles cual será la que mejor se ajusta a dicha secuencia Y tendríamos que encontrar la frase $\hat{W}$ que satisfaga

$$\hat{W} = arg \max_{W} P(W/Y),$$

es decir, elegir la secuencia de palabras que maximice la probabilidad de obtener la evidencia acústica Y. Utilizando el teorema de Bayes [6]. podemos escribir la ecuación anterior de la forma más operativa

$$\hat{W} = arg \max_{W} \frac{P(W)P(Y/W)}{P(Y)} = arg \max_{W} P(W)P(Y/W), \tag{1}$$

donde $P(W)$ es la probabilidad de que la frase W haya sido emitida, y $P(Y/W)$ la probabilidad de que cuando un hablante pronuncie la frase W a su vez genere los vectores acústicos Y. Dado que $P(Y)$, que es la probabilidad de esos vectores acústicos en general, no depende de W es posible escribir la segunda forma de la igualdad.

La ecuación (1) es la ecuación fundamental de nuestro problema. El diseño de un reconocedor de habla consiste entonces en estimar los modelos de $P(Y/W)$ y $P(W)$. El modelo $P(Y/W)$ se lo llama *modelo acústico* y es función de la evidencia acústica Y y de la secuencia de palabras W. El modelo $P(W)$ se lo llama *modelo de lenguaje* ya que es solamente función de la secuencia de palabras. Una vez estimados dichos modelos el proceso de reconocimiento consiste en digitalizar una frase pronunciada por un hablante, determinar el vector acústico Y correspondiente a dicha frase, y determinar para todas las secuencias de palabras posibles W cuál es la que maximiza la ecuación (1) para ese vector acústico. La determinación de la secuencia de palabras óptima es un problema de programación dinámica que se implementa mediante el algoritmo de Viterbi [6].

El presente trabajo tiene dos objetivos principales. El primero es la implementación de un sistema de reconocimiento de voz en idioma español con hablantes que pronuncian textos leídos de diarios. La base de datos con locuciones de hablantes que se dispone tiene el tamaño necesario para implementar modelos acústicos muy detallados. Sin embargo, los textos con los que se cuenta son relativamente escasos para implementar modelos de lenguaje con un nivel de detalle equivalente al de los modelos acústicos. Este es un problema frecuente en el diseño de sistemas de reconocimiento de habla ya que en general no se dispone de suficientes ejemplos de textos que contengan las palabras del vocabulario que se quiere reconocer. En nuestro caso al ser texto de lectura de diarios es simple obtener más ejemplos a través de la búsqueda de los mismos en la web. Sin embargo, dado que nuestra intención futura es implementar el mismo sistema para otros tipos de bases de datos, asumimos que el único texto de lectura de diarios que disponemos corresponde a las transcripciones de las frases leídas en la base de datos de audio usada para el entrenamiento de modelos acústicos. La mejora de los modelos de lenguaje la lograremos mediante la técnica conocida como adaptación de modelos de lenguaje.

El segundo objetivo del presente trabajo es la evaluación del desempeño de un nuevo método de estimación de modelos de lenguaje implementado por nosotros. La descripción detallada del mismo se puede ver en [5]. El mismo ha tenido un gran desempeño en una escala de datos reducida. En este trabajo se evalúa su desempeño en el sistema que se implementa y se compara con otros métodos de estimación de modelos de lenguaje implementados con algoritmos que son estado del arte.

2. Modelización estadística del habla

Esta sección describe brevemente los fundamentos del modelo estadístico utilizado en nuestro sistema de reconocimiento.

2.1. El modelo Acústico

El método estadístico más utilizado en la transcripción automática de habla son las Cadenas Ocultas de Markov [8],[6]. En este tipo de modelos se supone que la naturaleza de la señal es tal que puede ser descripta por medio de la existencia de varios estados internos diferentes subyacentes. Cada aparición de un nuevo dato supone que el sistema se encuentra en uno de esos posibles estados internos emitiendo el dato de manera probabilística. Además, con cada nueva aparición de un dato también se considera que hay una cierta probabilidad de pasar a otro estado interno para la emisión siguiente.

En el caso de la señal de habla se considera que los estados internos representan a los sonidos elementales del lenguaje, o fonemas. De este modo a medida que la señal de habla pasa por diferentes fonemas emitidos, la Cadena de Markov irá pasando por diferentes estados internos, cada uno de ellos con una función de

densidad de probabilidad determinada de emitir un vector acústico correspondiente. La función de densidad de probabilidad de cada estado será diferente para representar las diferentes características acústicas de cada fonema. Por ejemplo, una Cadena de Markov que represente la secuencia de vectores codificados que se obtuvieron durante la emisión de la palabra "nueve" podría ser representada simbólicamente del modo que se muestra en la figura siguiente:

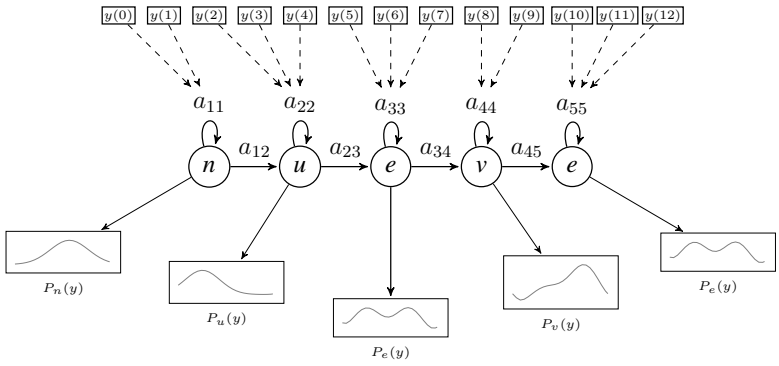

Figura 1. Representación de la emisión de la evidencia acústica para una dada palabra mediante Cadenas de Markov

Como puede verse en este gráfico esquemático, cada vector acústico es emitido por algún estado de la cadena y su valor dependerá de la función de probabilidad asociada a dicho estado.

En la práctica es usual que este modelo esquemático sea más complejo. Los estados internos suelen ser más detallados que lo que se muestra en la figura. En general se considera que cada uno de los estados internos que representan los sonidos elementales están compuestos a su vez de 3 estados, que dan cuenta de las variaciones entre inicio del sonido, parte central y final del sonido. También suele hacerse una discriminación entre fonemas que están en diferentes contextos acústicos, desdoblando la representación de cada fonema. Las unidades elementales de este tipo se denominan fonemas contexto dependiente, dando como resultado por ejemplo que la función de probabilidad de la primera "E" de la figura anterior sea diferente de la utilizada en la última "E", ya que ambas tienen diferentes contextos sonoros.

Una vez determinadas las palabras que componen nuestro vocabulario, será posible asignarles modelos acústicos a partir de la concatenación de modelos fonéticos. De igual modo, la modelización de una secuencia de palabras se implementa concatenando los modelos de cada una de ellas. Cada secuencia de palabras tendrá asociado entonces un modelo $P(Y/W)$ requerido en la ecuación (1), cuyos parámetros deberán ser estimados. El problema de estimación de este tipo de modelos fue propuesta por Baum ([6]) en el algoritmo conocido como *forward-backward*.

2.2. El modelo de lenguaje

Los modelos de lenguaje consisten en la estimación de probabilidades basadas en datos de entrenamiento recogidos a partir de texto escrito. En términos de nuestra ecuación de reconocimiento (1) se requiere encontrar el modelo $P(W)$ para cada posible frase W formada por palabras de un vocabulario definido. La estimación estadística de esta probabilidad no es trivial. Supongamos por ejemplo que queremos estimar la probabilidad de la frase: *la casa es linda*. Si pretendemos estimar dicha probabilidad contando el número de veces que dicha frase aparece en un texto de entrenamiento, obtendremos casi con certeza una probabilidad nula independientemente del tamaño del texto. El enfoque más usado para atacar este problema es la llamada modelización de *n-gramas* [8], que se basa en la suposición que la probabilidad de una palabra en una frase no es función de todas las palabras que la preceden, sino sólo de $n-1$ palabras anteriores. Si aplicamos sucesivamente el teorema de Bayes en el cálculo de $P(W)$ tendremos en forma explícita la dependencia de cada palabra con las que la preceden:

$$P(la, casa, es, linda) = P(linda/es, casa, la)P(es/casa, la)P(casa/la)P(la)$$

Si hiciéramos una aproximación de *bi-gramas*, es decir, $n = 2$, tendríamos que cada palabra depende solamente de la palabra anterior, por lo que

$$P(la, casa, es, linda) \approxeq P(linda/es)P(es/casa)P(casa/la)P(la)$$

El problema se reduce ahora a estimar solamente probabilidades de *bi-gramas* y, si la cantidad de ejemplos fuera suficiente, el estimador de la probabilidad del *bi-grama* como el cociente entre la frecuencia de ocurrencia del *bi-grama* y el número total de *bi-gramas* en un texto de entrenamiento, sería adecuado. Sin embargo, esto rara vez ocurre con todos los *bi-gramas*, por lo que se hace necesario *suavizar* los modelos con el objetivo de mejorar las estimaciones.

Adaptación de modelos de lenguaje Un problema que se encuentra habitualmente en la transcripción de habla automática es que se dispone de pocos ejemplos de textos de entrenamiento para estimar modelos de lenguaje de un vocabulario específico. Por ejemplo en una tarea de transcripción de una conversación entre médicos no es frecuente que el habla general represente adecuadamente el contenido de esa conversación. Como resultado, tendremos modelos de lenguaje pobremente estimados que se traducirán en una disminución del desempeño del reconocedor. La adaptación de modelos de lenguaje (véase [1]) se basa implementar dos tipos de modelos. El primero usando los datos de un texto no específico, por ejemplo texto de libros, pero con muchos ejemplos, que se denomina *modelo de background*. El segundo, llamado *modelo de adaptación* contiene ejemplos de palabras del vocabulario específico, pero una reducida cantidad de ejemplos. Finalmente, ambos modelos son combinados en uno nuevo que, se espera, posea las mejores características de cada uno.

3. Implementación del sistema

3.1. Implementación del modelo acústico

La primera definición que se debe hacer en nuestro sistema es el tipo de parametrización de los vectores acústicos. En nuestro caso usaremos una representación espectral conocida como coeficientes mel cepstrum (Mel Frequency Cepstral Coefficients, MFCC) basada en la respuesta en frecuencia del oído humano. Este tipo de codificación ha sido utilizada ampliamente y ha mostrado buena capacidad de representación acústica y discriminación entre diferentes porciones sonoras del habla. Para minimizar variaciones indeseadas se realiza una normalización de los coeficientes (restándoles la media de la frase entera) y además realizando las diferencias temporales primeras y segundas, que le da además la posibilidad de caracterizar variaciones temporales de las características acústicas. De este modo obtenemos un vector de 39 coeficientes cada 10 milisegundos como vector de evidencia acústica.

El siguiente paso es determinar el tipo de función de densidad de probabilidad paramétrica que se usará, que en general es una mezcla de gaussianas [6]. Cada fonema contexto dependiente es entonces representado mediante una secuencia de tres estados, y cada uno de esos estados se representará a su vez con una función de probabilidad de mezcla de gaussianas. Una vez determinada la topología de las redes de estados así como la forma paramétrica que tendrá cada función de probabilidad de cada estado comienza el proceso de estimación de los miles de parámetros que entran en juego en esas definiciones. Ya que los algoritmos de estimación son iterativos, es decir, partiendo de una configuración inicial se reentrenan las probabilidades hasta obtener la mejor estimación, el proceso de inicialización debe ser desarrollado cuidadosamente, ya que un punto de partida subóptimo dará resultados más pobres que los máximos posibles. Es posible encontrar buenos modelos iniciales si se cuenta con una serie de datos segmentados temporalmente a nivel fonético, es decir, frases en las cuales se conoce la ubicación temporal de cada fonema. En trabajos anteriores del grupo [10] obtuvimos la segmentación temporal de la base de datos Latino40, que se utilizará en este caso para inicializar nuestros modelos. El proceso sobre nuestra base de datos comienza elaborando modelos muy sencillos, unidades elementales de fonemas de una sola gaussiana (27 modelos de fonemas más 2 modelos de silencios, uno de la corta y otro de larga duración). Los parámetros de estos modelos iniciales son estimados con esos datos segmentados y luego ajustados mediante el algoritmo de Baum-Welch sobre la base de datos de este trabajo. Con los modelos iniciales de fonemas optimizados recién en ese punto se crean los primeros modelos de fonemas contexto dependiente utilizando árboles de decisión [6] para compartir parámetros entre estados que no tengan suficientes datos de entrenamiento. Inicialmente se mantiene una sola gaussiana, y son refinados nuevamente con el algoritmo de Baum-Welch hasta obtener convergencia. Luego, comienza el proceso de incrementar el número de gaussianas de los modelos con el objetivo que las funciones de probabilidad estimadas se acerquen cada vez más a las reales. Sobre el conjunto de señales de desarrollo se chequea cuál es el

número de gaussianas que da la menor probabilidad de error, resultando en 16 gaussianas por estado. Todos nuestro sistema está desarrollado en la plataforma HTK desarrollada por la Universidad de Cambridge de difusión libre [11].

3.2. Implementación del modelo de lenguaje

En este trabajo se proponen tres tipos de estimadores de modelos de lenguaje, que implementan diferentes tipos de suavizado. El primero conocido como modelo de Kneser-Ney [9] pertenece a la clase de los llamados modelos de descuento y durante mucho tiempo ha sido el modelo de mejor desempeño. El segundo modelo es el modelo propuesto por Chen [4], que parece haber superado al modelo de Kneser-Ney aunque a un costo computacional muy superior. El modelo de Chen pertenece a la clase de modelos llamados exponenciales o de máxima entropía. Este tipo de modelos permite incorporar y combinar diferentes fuentes de información, por lo que suelen tener alto desempeño a costa del aumento en la complejidad algorítmica [2]. Finalmente uno de los objetivos del presente trabajo es evaluar el desempeño de nuestro algoritmo descripto en [5], que es el tercer modelo que usaremos en nuestro sistema. El modelo, que llamamos de máxima entropía regularizada, también pertenece a la categoría de modelos exponenciales. Si bien tuvo un excelente desempeño, la evaluación se realizó sobre un conjunto de datos reducido, y no fue evaluado sobre un sistema de reconocimiento de voz. En este trabajo se pretende determinar si dicho desempeño es escalable a un sistema de reconocimiento de razonable complejidad como el descripto en este trabajo.

Implementación de los tipos de adaptación usados En este trabajo será necesario utilizar técnicas de adaptación de modelos de lenguaje para compensar la falta de datos de texto específico. Para el caso de modelos de Kneser-Ney y de máxima entropía regularizada usaremos la técnica conocida como interpolación de modelos [7] que es una de las más utilizadas. La idea consiste en realizar una combinación convexa de las probabilidades obtenidas mediante el modelo de background y el de adaptación. El factor que controla esta combinación se ajusta de modo de obtener los mejores resultados posibles sobre un conjunto de datos de desarrollo.

El modelo de Chen no utiliza interpolación de los modelos de background y de adaptación. En su lugar utiliza los parámetros de los modelos de adaptación y background y crea un nuevo modelo cuyos parámetros contienen los parámetros de los modelos separados.

4. Evaluación experimental

La medición del desempeño de los sistemas de reconocimiento de habla se realiza mediante un índice conocido como *exactitud*, que no es otra cosa que el porcentaje de palabras reconocidas correctamente sobre el total de palabras que componen el conjunto de prueba.

Para la implementación del sistema se utilizó la base de datos en español ALEC [3], que consiste principalmente de frases tomadas de diarios en español, emitidas principalmente por hablantes latinoamericanos. De todos los audios que aparecen en esta base se separó un conjunto de datos para utilizar en el entrenamiento de los modelos acústicos de aproximadamente quince horas de duración. Además se tomó un conjunto de 320 frases emitidas por ocho hablantes diferentes que se reservó como conjunto de prueba para evaluar el modelo. Finalmente se reservó un conjunto de datos similar al de prueba, de 320 frases dicha por 8 hablantes diferentes de los de prueba, que se utilizó en el desarrollo para el ajuste de parámetros.

La cantidad de palabras a modelizar, o sea el vocabulario, se definió a partir de las transcripciones de los audios de la base de datos ALEC. Se eligió un vocabulario de 5000 palabras, lo cual puede considerarse como una tarea de mediano vocabulario.

Para la implementación de los modelos de lenguaje se utilizaron tres tipos de textos. El texto de entrenamiento de la base ALEC, compuesto de un total de aproximadamente 30000 frases, que es de tamaño chico para lo que se considera habitualmente un buen cuerpo de entrenamiento, pero con la ventaja de ser un vocabulario usual y específico a la tarea que se tiene que reconocer. Luego, para ampliar el cuerpo de texto de entrenamiento se obtuvo texto tomado de libros en español de aproximadamente 7 millones de frases. De este texto además se separó un subconjunto compuesto por 30000 frases, que se usará con el propósito evaluar si es posible lograr una mejora adicionando relativamente poco texto inespecífico. Las transcripciones de los conjuntos de señales de prueba y desarrollo usados en el modelo acústico son utilizados también para el modelo de lenguaje.

Los experimentos realizados consisten en la implementación de un modelo acústico optimizado, y varias alternativas para el modelo de lenguaje. La optimización del modelo acústico se basó en la comparación del desempeño del modelo obtenido, sin considerar modelo de lenguaje, con un modelo óptimo ya conocido en idioma inglés. Este modelo es el que se aplica para la transcripción automática en inglés de la base de datos Wall Street Journal. Los resultados de esta tarea son conocidos y reportados en varias publicaciones. La evaluación del desempeño del modelo acústico puro se realizó asumiendo que las probabilidades $P(W)$ del modelo de lenguaje en la ecuación (1) son las mismas para todas las frases, dando como resultado igual desempeño para ambas tareas.

Resultados Se realizaron siete experimentos usando los tres modelos de lenguaje estimados, el modelo de Kneser-Ney, el modelo de Chen y nuestro modelo de máxima entropía regularizada. Además usamos diferentes combinaciones con los tipos de texto. Transcripciones de la base ALEC (ALEC); texto reducido de libros (l); texto de libros completo (L); concatenación de los textos (ALEC) y (l); concatenación de los textos (ALEC) y (L). Finalmente se combinaron mediante adaptación los textos (ALEC) y (l) y los textos (ALEC) y (L). El modelo de máxima entropía regularizada no pudo ser entrenado para el conjunto de libros grande (L) ni sus combinaciones con ALEC ya que el costo computacional

resultante resultó prohibitivo. Como se mencionó, el costo computacional es la principal desventaja de los modelos de máxima entropía y frecuentemente es necesario evaluar el desempeño de los mismos en conjuntos de datos reducidos [4].

Los resultados se muestran en la Tabla 1. Una primera conclusión es que frente a un cuerpo de entrenamiento de texto de tamaño similar, siempre es preferible realizar el entrenamiento con el texto específico. Esto se puede leer de la comparación de las filas 1 y 2, ya que tanto el conjunto llamado ALEC como l son de tamaño similar. Si tenemos una masa de texto inespecífico muy grande en cambio (L, fila 3 de la tabla) la conclusión puede variar, dependiendo del estimador usado. El método de Knesser y Ney (primera columna) da mejores resultados que en el caso de entrenar solamente con ALEC, mientras que el modelo de Chen mejora con respecto al conjunto chico l, pero no llega a superar al entrenamiento con ALEC solamente.

En la fila 4 y 5 de la tabla se registran resultados para el caso de combinar los textos de ALEC con los libros en cantidades pequeñas (l) o en cantidades grandes (L). Los resultados de concatenar los textos muestran que no hay diferencia en la estimación si la cantidad adicionada de texto inespecífico es similar a la del específico (fila 4). Esta estrategia sólo da mejores resultados para el caso del estimador de Knesser y Ney y sólo si la cantidad de libros es muy grande (L, fila 5), siendo además un poco mejor que estimar sólo con el conjunto L de libros.

En las filas 6 y 7 vemos los resultados para los 3 métodos en el caso de realizar adaptación en lugar de simplemente tomar todo el texto junto. Adaptando con un conjunto de texto inespecífico pequeño (fila 6) tenemos mejores resultados que simplemente concatenando los textos (fila 4) para los tres casos de estimadores. También vemos que la adaptación es mejor para el caso del total de texto inespecífico (L, fila 7). El mejor resultado se da para el modelo de Knesser y Ney adaptando con la masa de datos de libros total (L, fila 7). Es interesante notar que el modelo de Chen no es superior al modelo de Knesser-Ney como se ha reportado [4].

Por último vemos que nuestro estimador de máxima entropía regularizada es sistemáticamente superior a los estimadores de Kneser-Ney y Chen para todos los casos evaluados. Es interesante notar que adaptando ALEC con textos pequeños en el modelo de máxima entropía regularizada, sólo se obtiene una degradación del 0.5 % con respecto al mejor resultado (ALEC adaptado con L y modelo de Kneser-Ney).

5. Conclusiones

Se implementó un sistema de reconocimiento automático de habla para una tarea de reconocimiento de texto leído de diarios en idioma español. El sistema de mediano vocabulario es capaz de reconocer entre 5000 palabras diferentes y los resultados obtenidos son comparables a los sistemas en idioma inglés de similares características que son estado del arte en su tipo. Dado que se disponía de relativamente poca cantidad de texto específico de la tarea se usaron técnicas de adaptación al lenguaje que permitieron aumentar el desempeño, obteniéndose

	Modelos de Lenguaje		
Datos de entrenamiento	Kneser-Ney	Chen	Max-Ent-Reg
ALEC	87.62	87.05	**88.22**
l	83.82	83.08	**85.33**
L	88.86	86.62	-
ALEC más l	87.82	87.11	**88.12**
ALEC más L	**89.37**	87.06	-
ALEC y l adaptado	87.95	87.55	**88.49**
ALEC y L adaptado	**89.64**	89.40	-

Cuadro 1. Exactitudes obtenidas con los tres tipos de estimación de modelo de lenguaje

una exactitud de reconocimiento de 89.64 %. También se incorporó un modelo de lenguaje basado en un estimador de máxima entropía desarrollado en un trabajo anterior por nuestro grupo, el cual mejoró el desempeño respecto de los mejores modelos de lenguaje conocidos bajo varios tipos de datos evaluados. Como trabajo a futuro se desprende la importancia de una mejora algorítmica de nuestro modelo de lenguaje con el fin de incluir cantidades de texto mayores.

Referencias

1. Bellegarda, J.R.: Statistical language model adaptation: review and perspectives. Speech Communication 42, 93–108 (2004)
2. Berger, A.L., Pietra, S.A.D., Pietra, V.J.D.: A maximum entropy approach to natural language processing. Computational Linguistics 22, 39–71 (1996)
3. Bratt, H., Neumeyer, L., Shriberg, E., Franco, H.: Collection and detailed transcription of a speech database for development of language learning technologies. In: Proceedings of the International Conference on Speech and Language Processing (December 1998)
4. Chen, S.F.: Performance prediction for exponential language models. In Proceedings of NAACL HLT (2009)
5. Estienne, C., Pelle, P.A.: Smoothing of bi-grams models using regularized maximum entropy. In: Anales de la RPIC 2011, vol. 1, pp. 42–47. Entre Ríos, Argentina (2011)
6. Huang, X., Acero, A., Hon, H.W.: Spoken Language Processing: a guide to theory, algorithm, and system development. Prentice Hall PTR (2001)
7. Jelinek, F., Merialdo, B., Roukos, S., Strauss, M.: A dynamic language model for speech recognition. In: Proceedings of the workshop on Speech and Natural Language (1991)
8. Jelinek, F.: Statistical Methods for Speech Recognition. MIT Press (1997)
9. Kneser, R., Ney, H.: Improved backing-off for m-gram language modeling. In: ICASSP-95. vol. 1, pp. 181 –184 vol.1 (1995)
10. Preiti, P.P., Estienne, C., Simkin, D., Perez, S., Pelle, P.: Reconocedor de números telefónicos basado en modelos de markov ocultos. In: Actas del III MACI. pp. 611–614, Bahía Blanca, Argentina (2011)
11. Young, S.: A review of large-vocabulary continuous-speech. Signal Processing Magazine, IEEE 13(5), 45 (1996)