MODELOS DE LENGUAJE VARIANTES EN EL TIEMPO

Milone D. H. [1], Rubio A. J. [2] y López-Cózar R.[2]

[1]Universidad Nacional de Entre Ríos, Argentina, d.milone@ieee.org
[2]Univesidad de Granada, España, {rubio, rlopezc}@ugr.es

PALABRAS CLAVE

Reconocimiento automático del habla, modelos de lenguaje, modelos variantes en el tiempo.

1. INTRODUCCIÓN

En un sistema de reconocimiento automático de habla continua (RAHC) se estiman las probabilidades del modelo de lenguaje (ML) durante la fase de entrenamiento. Generalmente, durante las posteriores etapas de reconocimiento, todas estas probabilidades quedan fijas (al igual que las probabilidades del resto de los modelos). Sin embargo, puede resultar útil que durante la fase de reconocimiento, los modelos se adapten a las nuevas condiciones en que se encuentran para permitir la incorporación de más información del instante actual. Estas adaptaciones pueden ser de tipo permanente o temporal. En el caso de las adaptaciones al hablante o las adaptaciones a determinadas condiciones de ruido, suelen realizarse adaptaciones permanentes sobre los modelos acústicos (MA). Las adaptaciones que se aplican en un momento dado del proceso de reconocimiento, influirán sobre todas las instancias posteriores del mismo. En el caso de las adaptaciones temporales, se trata de realizar ciertos ajustes en los modelos para incorporar información del instante actual que beneficie el reconocimiento de la frase en evolución pero no se fijarán estos cambios, de forma que influyan para frases posteriores o para instantes posteriores de la misma frase. Éste es el tipo de adaptaciones que se proponen en el presente artículo. A partir de un ML básico, estimado con métodos estándar, palabra a palabra se realizan ajustes en las probabilidades para incorporar información de la frase que está siendo reconocida actualmente.

En la siguiente sección se presentarán los detalles y formulaciones matemáticas de esta idea. En la sección 3 se describirá el método experimental utilizado para la implementación de los modelos de lenguaje variantes en el tiempo (MLVT). La sección 4 describe los experimentos realizados y la base de datos utilizada. Finalmente, en la sección 5 se presentan los resultados y una discusión general sobre estos nuevos MLVT.

2. MODELOS DE LENGUAJE

Para comenzar, en esta sección se introducirá la bien conocida formulación de los ML. Esta idea servirá de base y distinción para la propuesta de los MLVT. En ambos casos se aceptará que la aplicación de los ML al RAHC, se realiza mediante modelos ocultos de Markov (MOM).

2.1 Modelado estadístico del lenguaje

Dada la producción de la señal X, el paradigma estocástico de RAHC busca la secuencia de palabras W para la que la probabilidad $P(W \mid X)$ es máxima. En la práctica, no es fácil de encontrar esta probabilidad y se utiliza la regla de Bayes para dividir el problema en dos partes [1]:

$$P(W \mid X) = \frac{P(X \mid W)P(W)}{P(X)}$$

Ahora, $P(X \mid W)$ es conocido como modelo acústico y es relativamente más sencilla su estimación. $P(W)$ es el modelo de lenguaje, que provee la probabilidad de una frase o secuencia de palabras dada. $P(X)$ sigue siendo difícil de determinar, pero afortunadamente no es necesaria para maximizar $P(W \mid X)$.

La determinación exacta del ML es una tarea poco práctica. Es por esto que algunas simplificaciones son necesarias y así, lo que realmente se obtienen son sólo aproximaciones de $P(W)$. Uno de los enfoques más comunes consiste en escribir $P(W)$ como:

$$P(W) = P(w_1 w_2 \ldots w_Q) = \\ = P(w_1)P(w_2 \mid w_1)P(w_3 \mid w_1 w_2)\ldots P(w_Q \mid w_1 \ldots w_{Q-1})$$

sinc(*i*) Laboratory for Signals and Computational Intelligence (http://fich.unl.edu.ar/sinc)
D. H. Milone, A. J. Rubio & R. López-Cózar; "Modelos de Lenguaje Variantes en el Tiempo"
Memorias del XXIV congreso nacional de ingeniería biomédica. Oct, 2001.

donde w_i son las palabras y la secuencia $w_1 w_2 \dots w_Q$ es la frase W. A pesar de esta simplificación, la estimación de todas las probabilidades de la productoria sigue siendo una tarea difícil. Generalmente se simplifica la historia a *n* palabras, obteniendo así las denominadas *n*-gramáticas, asumiendo la independencia estadística de la palabra y la parte más antigua de la historia. Por ejemplo, una bi-gramática tiene la forma $P(w_i \mid w_{i-1})$ y asumiendo la independencia estadística de todas las palabras en la historia el ML se reduce a:

$$P(W) = P(w_1)P(w_2 \mid w_1)P(w_3 \mid w_2)\dots P(w_Q \mid w_{Q-1}).$$

Sin embargo, en muchos casos prácticos algunas historias $(w_i\ w_{i-1})$ nunca aparecen en el corpus de entrenamiento. Es por esto que es necesario considerar el *suavizado* de las gramáticas. Por medio de estas técnicas, es posible estimar las probabilidades de las palabras cuyas historias de orden *n* nunca aparecen en el corpus de entrenamiento. Existen muchas técnicas útiles para el suavizado de gramáticas [2]. Una de las técnicas más utilizadas para la estimación y suavizado de gramáticas es la denominada *back-off* [3]. Mediante este enfoque las probabilidades de uni-gramáticas son dadas por:

$$\hat{P}(w_i) = \frac{C(w_i)}{\sum_{\forall j} C(w_j)}$$

donde $C(\cdot)$ cuenta las ocurrencias de una determinada secuencia de palabras en el corpus de entrenamiento. Las probabilidades *back-off* para una bi-gramática quedan determinadas por:

$$\hat{P}(w_i \mid w_j) = \begin{cases} \dfrac{C(w_i w_j) - \vartheta}{C(w_i)} & \text{if } C(w_i w_j) > 0 \\ \beta(w_i) P(w_j) & \text{if } C(w_i w_j) = 0 \end{cases}$$

donde $\vartheta = 0.5$ y

$$\beta(w_i) = \frac{1 - \sum_{j:C(w_i w_j)>0} \hat{P}(w_i \mid w_j)}{1 - \sum_{j:C(w_i w_j)=0} \hat{P}(w_j)}$$

2.2 Modelos de lenguaje variantes en el tiempo

En trabajos anteriores se ha incorporado información adicional a través del ML en una etapa posterior al proceso de reconocimiento. Por ejemplo, en un trabajo recientemente publicado [4], se incorporó información prosódica modificando las probabilidades de la red de hipótesis de palabras, salida de un reconocedor basado en MOM. No es usual la incorporación de información extra en etapas previas al reconocimiento. El desarrollo teórico de nuestra propuesta integra, a través del ML, información que cambia en el tiempo durante el proceso de reconocimiento de cada frase del corpus.

La principal idea de los MLVT es modificar un ML de referencia a medida que el tiempo avanza durante el proceso de reconocimiento de una frase. Con esto en mente, supongamos que el reconocedor se encuentra en medio de una búsqueda y que una de las hipótesis acústicamente plausible está dada por:

$$\boldsymbol{h}_{i_1}^n = w_{i_1-1}, w_{i_1-2}, \dots, w_{i_1-n+1}$$

con una probabilidad de transición $\hat{P}(w_{i_1} \mid \boldsymbol{h}_{i_1}^n)$ hacia las siguientes palabras. En esta expresión se ha adoptado la notación vectorial para indicar secuencias en el tiempo. Ahora, en un instante de tiempo posterior, otra hipótesis acústicamente plausible podría ser:

$$\boldsymbol{h}_{i_2}^n = w_{i_2-1}, w_{i_2-2}, \dots, w_{i_2-n+1}$$

con una probabilidad $\hat{P}(w_{i_2} \mid \boldsymbol{h}_{i_2}^n)$ para la transición hacia las próximas posibles palabras.

Para un *n* fijo (como generalmente sucede en los sistemas de reconocimiento actuales):

$$\boldsymbol{h}_{i_1}^n = \boldsymbol{h}_{i_2}^n \wedge w_{i_1} = w_{i_2} \quad \Rightarrow \quad \hat{P}(w_{i_1} \mid \boldsymbol{h}_{i_1}^n) = \hat{P}(w_{i_2} \mid \boldsymbol{h}_{i_2}^n)$$

es decir, para iguales historias en diferentes posiciones dentro de una frase, corresponden iguales probabilidades de transición entre palabras.

Sin embargo, podrían existir otras evidencias indicando que dados dos tiempos de análisis diferentes en la frase (i_1 e i_2) se observe que:

$$\hat{P}(w_{i_1} \mid \boldsymbol{h}_{i_1}^n) \neq \hat{P}(w_{i_2} \mid \boldsymbol{h}_{i_2}^n).$$

Por ejemplo, cuando *n*=2, esto sería:

$$\boldsymbol{h}_{i_1}^1 = w_{i_1-1} \text{ y } \boldsymbol{h}_{i_2}^n = w_{i_2-1}$$

con probabilidades de bi-gramática:

$$\hat{P}(w_{i_1} \mid w_{i_1-1}) \text{ y } \hat{P}(w_{i_2} \mid w_{i_2-1}).$$

sinc(*i*) Laboratory for Signals and Computational Intelligence (http://fich.unl.edu.ar/sinc)
D. H. Milone, A. J. Rubio & R. López-Cózar; "Modelos de Lenguaje Variantes en el Tiempo"
Memorias del XXIV congreso nacional de ingeniería biomédica. Oct, 2001.

Obviamente si $w_{i_1-1} = w_{i_2-1}$ y $w_{i_1} = w_{i_2}$, la probabilidad del ML es independiente de la posición de la palabra en la frase. Para los MLVT, la idea es permitir que esta probabilidad sea adaptada en diferentes momentos del proceso de reconocimiento de una frase y para las diferentes frases a reconocer. Para adaptar las probabilidades del ML durante el reconocimiento se propone la incorporación de una función de penalización:

$$\hat{P}_t(w_i \mid \boldsymbol{h}_i^n) = \varphi_t(w_i, \boldsymbol{h}_i^n, E_t)\hat{P}(w_i \mid \boldsymbol{h}_i^n)$$

donde E_t representa cualquier información *E*xtra para el tiempo t de la frase que está siendo reconocida. La función φ genera un valor numérico en el rango real [0,1]. Esta función reduce la probabilidad del ML de referencia cuando la evidencia E no sea favorable a la transición de palabra hipotética en el tiempo t.

2.3 Integración del modelo acústico y el modelo de lenguaje

Para entender la forma en que se integran el MA y el ML en el proceso de decodificación de la señal X en palabras, es conveniente considerar al modelo de lenguaje como un autómata recursivo y probabilístico de estados finitos (como lo son las *n*-gramáticas). La Figura 1 muestra un ejemplo sencillo de modelo de lenguaje con red recursiva (MLRR). Este modelo está compuesto por palabras que se conectan mediante arcos.

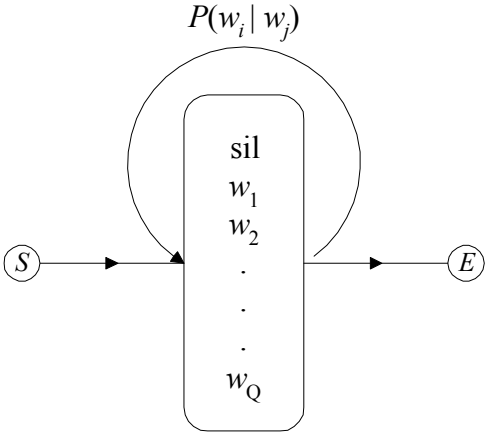

Figura 1: ML con red recursiva para una bi-gramática.

Algunos arcos y sus probabilidades asociadas se obtienen directamente desde el corpus de entrenamiento. Los arcos relacionados con silencios son comúnmente incluidos para otorgar mayor flexibilidad al modelo y poder incluir situaciones naturales en el lenguaje hablado. Con este propósito aparece en la Figura 1 la palabra etiquetada como "sil". Existen también en estos modelos muchos arcos que relacionan secuencias de palabras que no se encuentran directamente en el corpus de entrenamiento, pero sus probabilidades pueden ser estimadas mediante el suavizado de la gramática.

El algoritmo de decodificación [5] evalúa todas las posibles hipótesis partiendo del nodo con la etiqueta *S* y evalúa todas las posibles hipótesis. Generalmente, las probabilidades de transición entre palabras son acumuladas después de finalizar la decodificación acústica de cada hipótesis de palabra. Los MA de cada palabra se obtienen mediante la concatenación de un conjunto de MOM que representan sus unidades acústicas.

Como se puede ver en la Figura 1, usando un MLRR para bi-gramáticas (o n-gramáticas a través de redes recurrentes más complicadas), no es posible cambiar la probabilidad de transición de acuerdo a la posición de la palabra dentro de la frase. En estas redes, la probabilidad de transición entre dos palabras dadas no depende de la posición de las palabras en la frase. Es así como se requiere una nueva red para poder implementar los MLVT.

3. REDES EXPANDIDAS PARA MLVT

Para poder incorporar la función de penalización φ directamente en el algoritmo de decodificación de un reconocedor basado en MOM se propone una red de gramática alternativa que denominamos modelo de lenguaje con red expandida (MLRE). Basados en un ML de bi-gramática, para permitir que:

$$\hat{P}(w_{i_1} \mid w_{i_1-1}) \neq \hat{P}(w_{i_2} \mid w_{i_2-1}) \text{ para } i_1 \neq i_2$$

proponemos usar un autómata no recursivo (probabilístico y de estados finitos), en lugar del presentado en la Figura 1. Formalmente esta gramática no es una bi-gramática pero funcionalmente se pueden tomar algunas precauciones para que sea equivalente a una bi-gramática. En la Figura 2, se muestra una representación simplificada de un MLRE en la que solamente se permiten conexiones mediante arcos hacia delante. De esta forma, la red resultante deberá tener tantas capas como la cantidad máxima de palabras admitidas por frase a reconocer.

Para obtener el MLRE se estima previamente el MLRR mediante el método de *back-off*. Luego se realiza la "expansión" de gramática de forma que en la Figura 2:

sinc(*i*) Laboratory for Signals and Computational Intelligence (http://fich.unl.edu.ar/sinc)
D. H. Milone, A. J. Rubio & R. López-Cózar; "Modelos de Lenguaje Variantes en el Tiempo"
Memorias del XXIV congreso nacional de ingeniería biomédica. Oct, 2001.

$\hat{P}_l(w_i \mid w_j) = \hat{P}(w_i \mid w_j) \quad \forall l$. Dada una frase, cada transición de una palabra a la siguiente corresponde a una capa de probabilidades en un MLRE y a un bucle en el MLRR. Sin embargo, con la red expandida de la Figura 2, cada transición dentro de la frase puede ser modificada independientemente en relación con la posición de las palabras en la frase.

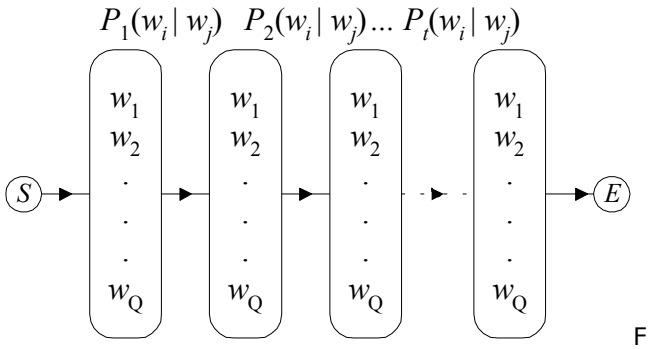

Figura 2: ML con red expandida para una bi-gramática.

Ahora, por ejemplo, la probabilidad asociada para la transición de una palabra en la primera capa (primera palabra de la frase) hacia una palabra en la segunda capa (segunda palabra en la frase) puede ser diferente a la probabilidad de transición entre las mismas palabras en las capas 3 y 4.

Hay que observar que el MLRE de la Figura 2 es más complejo que el MLRR de la Figura 1. Sin embargo, el segundo modelo no es recursivo y cada capa del MLRE se corresponde con una conexión hacia atrás en el MLRR. La propiedad recursiva del MLRR estándar ha sido sustituida por la repetición de capas idénticas. De esta forma, es una tarea sencilla obtener la versión expandida de una red de bi-gramática.

Este método reduce la complejidad de implementación de un MLVT y nos provee de la flexibilidad necesaria para realizar experimentos de laboratorio. Sin embargo, cuando el MLRE es utilizado en RAHC para simular MLVT, el reconocimiento debe hacerse en dos etapas: primero el ML se modifica para la frase que se va a reconocer (se expande hacia un MLRE y se penaliza) y luego se realiza un reconocimiento estándar con la red adaptada.

4. EXPERIMENTOS

En este trabajo se han aplicado los MLVT por medio de MLRE para introducir información acentual en un sistema de RAHC. Existen trabajos previos ([6], [7], [8], [9] y [10] entre otros) donde se ha incluido información prosódica al RAHC pero no se ha incorporado hasta el momento la información acentual en el proceso de reconocimiento.

La estructura acentual de una frase consiste en una sucesión de representaciones acentuales de las palabras que la componen. Por ejemplo, para la frase "Probando una estructura acentual", la estructura acentual puede representarse como: "LHL HL LLHL LLH", donde la H representa a las sílabas acentuadas y la L a las sílabas no acentuadas. Estas estructuras acentuales pueden ser utilizadas para restringir los caminos de búsqueda de un reconocedor. Cuando las hipótesis de reconocimiento posean una estructura acentual incorrecta se puede introducir un factor de penalización utilizando MLVT. Para cada instante de tiempo en la frase, el ML tendrá probabilidades de transición que favorecerán a la estructura acentual correcta. Aún más, para cada nueva frase a reconocer existirá un nuevo modelo de lenguaje favoreciendo la estructura acentual más adecuada. En nuestros experimentos, se utilizó un valor constante de penalización para aquellos casos en que la hipótesis de reconocimiento tenía una estructura acentual diferente a la de la frase a reconocer.

5. RESULTADOS Y DISCUSIÓN

Los resultados finales se obtuvieron con un subconjunto de 1000 frases del corpus de habla en español Albayzin [11]. Este subconjunto de frases incluye un total de 12 hablantes (6 mujeres y 6 hombres). Todas las frases son leídas en condiciones de laboratorio y el tamaño del vocabulario fue de 200 palabras.

Para realizar la validación cruzada se tomaron 5 particiones en cada una de las cuales se separaron aleatoriamente 800 frases para entrenamiento y 200 para la prueba. A partir de cada subconjunto de entrenamiento se entrenaron los MA y los MLRR. A partir de los 5 MLRR se obtuvieron los correspondientes MLRE que se utilizaron para realizar la penalización previa al reconocimiento de las frases de prueba de cada partición.

El sistema de referencia consiste en un reconocedor de habla continua basado en MOM con un modelo de lenguaje estándar (MLRR). Se utilizaron modelos de 3 estados para los distintos fonemas y el modelo del silencio [12]. Al final de cada palabra se incorpora una pausa corta modelada con un único estado. En cada estado, la señal de voz fue modelada mediante mezclas de funciones de densidad de probabilidad continuas con un pre-procesamiento basado en coeficientes cepstrales en escala de mel [13]. Para el sistema de prueba se utilizaron exactamente los mismos modelos

sinc(*i*) Laboratory for Signals and Computational Intelligence (http://fich.unl.edu.ar/sinc)
D. H. Milone, A. J. Rubio & R. López-Cózar; "Modelos de Lenguaje Variantes en el Tiempo"
Memorias del XXIV congreso nacional de ingeniería biomédica. Oct, 2001.

acústicos que se entrenaron para el sistema de referencia.

En la Tabla 1 se muestran los errores de reconocimiento para el sistema de referencia y para el sistema donde la información acentual fue incorporada mediante MLVT. En esta tabla, se puede observar una importante reducción del error de reconocimiento.

Tabla 1: tasa de error en frases (TEF) y tasa de error en palabras (TEA) para las pruebas de validación cruzada.

Partición corpus	Referencia		MLVT	
	TEF	TEA	TEF	TEA
1	44.54	7.89	24.37	4.82
2	42.86	8.06	25.21	5.10
3	32.20	5.99	18.64	3.37
4	33.61	5.96	21.85	3.16
5	31.93	8.70	19.33	5.67
Media	37.03	7.32	21.88	4.42

Se verificó previamente que utilizando un MLRE sin penalización los resultados son exactamente los mismos que para el correspondiente MLRR. En otras palabras, esto muestra que un MLRE es funcionalmente equivalente al MLRR del que fue obtenido mediante la expansión.

Para obtener estos resultados se utilizaron estructuras acentuales obtenidas de la trascripción textual de las frases en la base de datos. En español es posible obtener estas estructuras mediante la aplicación de unas simples reglas ortográficas. En otros lenguajes, como el inglés, sería necesario construir previamente un diccionario de acentuación. Esta situación es similar a la necesidad de un diccionario de pronunciaciones para los reconocedores de habla en inglés, que no son estrictamente necesarios para los reconocedores de habla en español.

Estos experimentos muestran, por un lado, la flexibilidad que proveen los MLRE para la simulación de un MLVT. Por otro lado, queda clara la utilidad de los MLVT como un nuevo modelo para la incorporación de información externa al RAHC. La forma en que esta información es incorporada constituye una nueva concepción que extiende la idea de ML.

REFERENCIAS

[1] Rabiner L. R. and Juang B. H., *Fundamentals of Speech Recognition*, Prentice-Hall, 1993.

[2] Jelinek F., Statistical Methods for Speech Recognition, MA: MIT Press, 1999.

[3] Potamianos G. and Jelinek F., "A study of n-gram and decision tree letter language modeling methods", *Speech Communication*, Vol. 24, pp. 171-192, 1998.

[4] Nöth E., Batliner A., Kieβling A., Kompe R., and Niemann H., "VERBMOBIL: The use of prosody in the Linguistic Components of a Speech Understanding System," *IEEE Trans. On Speech and Audio Processing*, Vol. 8, No. 5, 2000.

[5] Ney H. and Ortmanns S., "Dynamic Programming Search for Continuous Speech Recognition," IEEE Signal Processing Magazine, Vol. 16, No. 5, 1999.

[6] Chung G. and Seneff S., "Improvements in Speech Understanding Accuracy Through the Integration of Hierarchical Linguistic, Prosodic, and Phonological Contraints in the Jupiter Domain", Proc. of 5th International Conference on Spoken Language Processing, Spoken Language Understanding Systems 1, 1998.

[7] López E., Caminero J., Cortázar I. and Hernández L., "Improvement on Connected Numbers Recognition Using Prosodic Information", Proc. of 5th International Conference on Spoken Language Processing, Prosody and Emotion 2, 1998.

[8] Pols L. C. W., Wang X. and Bosch L. F. M., "Modeling of phone duration (using the TIMIT database) and its pottential benefit for ASR", Speech Communication, Vol. 19, pp. 161-176, 1996.

[9] Stolcke A., Shriberg E., Hakkani-Tür D. and Tür G., "Modeling the prosody of hidden events for improved word recognition," *Proc. of 7th European Conference on Speech Communication and Technology*, Vol. 1, pp. 311-314, 1999.

[10] Ross N. K. and Ostendorf M., "A Dynamical System Model dor Generating Fundamental Frequency for Speech Synthesis", IEEE Trans. On Speech and Audio Processing, Vol. 7, No. 3, 1999.

[11] Casacuberta F., García R., Llisterri J. Nadeu C., Prado J. M. and Rubio A., "Development of a Spanish Corpora for the Speech Research", Workshop on International Co-operation and Standardization of Speech Databases and Speech I/O Assessment Methods, CEC DGXIII, ESCA and ESPRIT PROJECT 2589 "SAM", Chiavari, 26-28 September 1991.

[12] Huang X. D., Ariki Y., Jack M. A., *Hidden Markov Models For Speech Recognition*, Edinburgh University Press, 1990.

[13] Deller. J. R., Proakis J. G., Hansen J. H., *Discrete-Time Processing of Speech Signals*, Prentice Hall, 1987.

Dirección de contacto:

Diego Milone (d.milone@ieee.org)
Laboratorio de Cibernética
Facultad de Ingeniería
Universidad Nacional de Entre Ríos
Casilla de Correos 47 Sucursal 3
Código Postal 3100
Paraná - Entre Ríos - Argentina

sinc(*i*) Laboratory for Signals and Computational Intelligence (http://fich.unl.edu.ar/sinc)
D. H. Milone, A. J. Rubio & R. López-Cózar; "Modelos de Lenguaje Variantes en el Tiempo"
Memorias del XXIV congreso nacional de ingeniería biomédica. Oct, 2001.