Capítulo 6

6 Identificación de Palabras Clave usando Sílabas como Fillers

Este capítulo es similar al anterior, con la diferencia de que en vez de modelar habla fuera del vocabulario con fonemas, se usan sílabas y clases generales de sílabas. En este capítulo se hacen experimentos con fillers de sílabas y con sílabas comunes agrupadas en clases generales. En ambos casos se experimenta con y sin modelos de lenguaje estocásticos. En un experimento final se usan sílabas comunes agrupadas en clases generales con múltiples pronunciaciones.

En este capítulo se hace un estudio sobre las sílabas en el español hablado en México, con el fin de generar y clasificar las sílabas que serán usadas como fillers. Al final de este capítulo, los experimentos son comparados con los resultados del experimento base que usa una gramática predefinida.

6.1 Modelado de Sílabas como Unidades Inferiores a la Palabra

Uno de los principales inconvenientes en reconocimiento de habla continua es el efecto de co-articulación provocado al pronunciar sonidos fluidamente. De esta forma, los fonemas como sonidos básicos, claros, precisos y bien definidos, sufren grandes variaciones al pronunciarse con otros fonemas en habla continua, debido a que un fonema se ve influenciado fuertemente por sus fonemas colindantes. Por consiguiente, el reconocimiento a nivel fonema es muy difícil. Al agrupar unidades progresivamente superiores, la primera unidad superior al fonema es la sílaba, que puede abarcar uno o varios fonemas,. Las sílabas sirven como interfaz importante de representación de lenguaje entre el nivel bajo (fonético y fonológico) y el nivel alto (morfológico y léxico).

Evidencias fonológicas y psicoacústicas sugieren que la sílaba como unidad de representación juega un importante papel en el procesamiento del lenguaje hablado, particularmente bajo situaciones acústicas adversas [Lau, 1998; Zhilog et al, 1994]. Para efectos de esta tesis, esta evidencia se experimentará modelando sílabas como fillers.

6.1.1 Generación de Sílabas

El proceso de generación de las sílabas que serán usadas como fillers se obtienen de acuerdo al proceso de división silábica explicado en el apéndice B, sobre el conjunto de entrenamiento descrito en la sección 5.3.2. El número de sílabas encontradas fue de 860 sílabas. Aunque el total de las sílabas encontradas forma escasamente una quinta parte del total de las sílabas que se pueden encontrar en español, creemos que las sílabas aquí encontradas pueden formar la mayoría de las palabras en habla continua en español.

6.2 Modelado de Fillers de Sílabas

Esta técnica consiste en usar sílabas como fillers y modelar palabras clave completas. En la figura 6-1 se muestra un esquema de esta técnica, que consiste de una palabra clave y M sílabas como fillers. La salida del sistema es una transcripción completa de la frase reconocida y el resultado es la palabra clave (los fillers son ignorados). La gramática que describe esta técnica es mostrada en el apéndice A3.

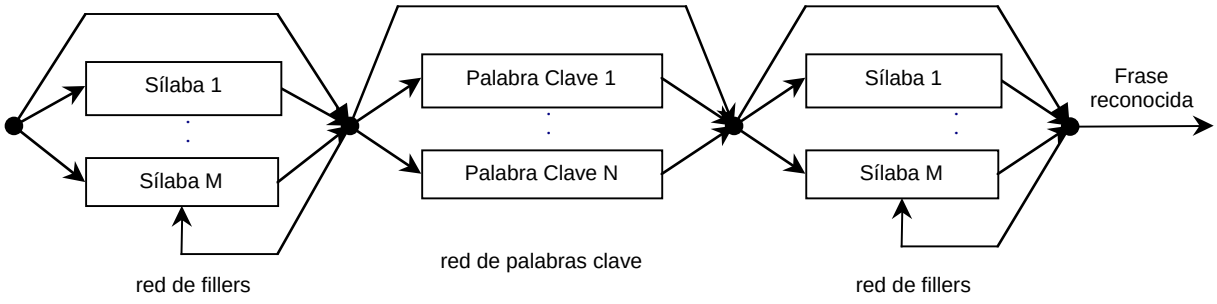

Figura 6-1 Esquema del sistema de identificación de palabras clave usando sílabas como fillers.

6.3 Modelado de Fillers de Sílabas Comunes usando Modelos de Lenguaje

En la sección anterior, se presentó el primer experimento de la técnica investigada en este capitulo, utilizando 860 sílabas como fillers. En los sistemas de identificación de palabras clave, usar muchos fillers es una desventaja, debido a que entre más grande el número de fillers, más alto es el costo computacional.

En un esfuerzo por obtener un buen desempeño al modelar sílabas como fillers sin tener un alto costo computacional. Como segundo experimento se propone agrupar las sílabas en clases generales de los fonemas que las componen. Para esto se propone agrupar los fonemas de las sílabas de acuerdo a su lugar de articulación, permitiendo definir clases de sílabas con sonidos acústicamente similares. La figura 6-2 ilustra la agrupación de fonemas, en donde las letras mayúsculas son las representantes de cada grupo de fonemas.

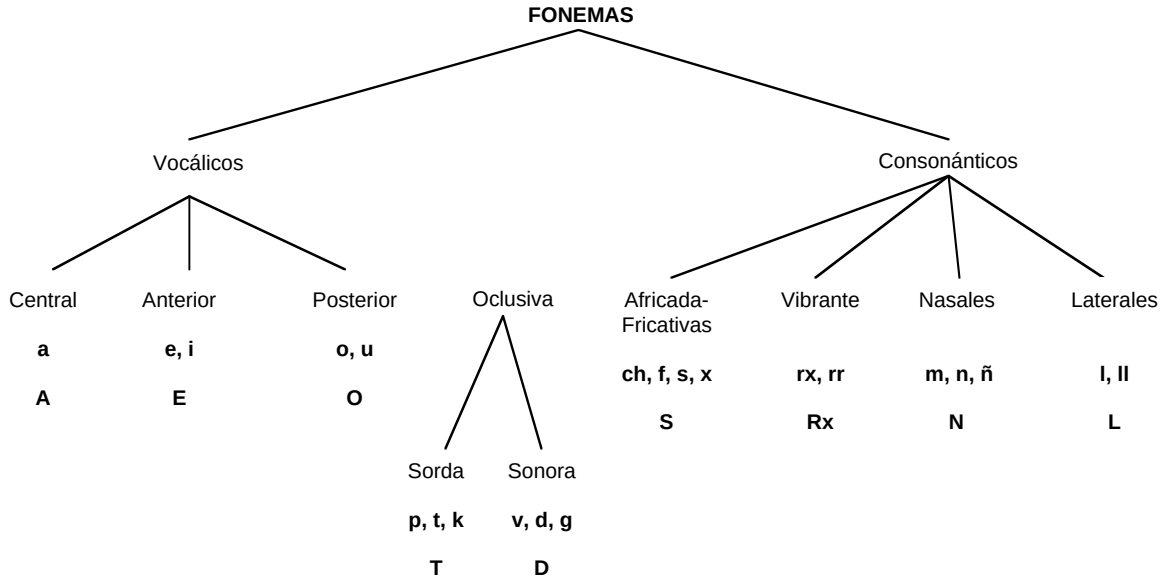

Figura 6-2 Clasificación de fonemas en clases generales de acuerdo a su lugar de articulación.

La agrupación de sílabas se realizó siguiendo el siguiente procedimiento: Se realizó el proceso de silabificación (división de sílabas) en el conjunto de datos de entrenamiento (ver sección 5.3.2) y se reemplazaron los fonemas por sus fonemas representantes. De acuerdo a esta agrupación, se obtuvieron 344 sílabas. Desafortunadamente, aún con esta clasificación el costo computacional sigue siendo alto. Debido a esta situación, se decidió clasificar las sílabas en *comunes* y *no comunes*. Una sílaba es considerada *común* cuando su ocurrencia es mayor o igual a la media aritmética de la ocurrencia de todas las sílabas. La sílaba *no común* es aquella que tiene una ocurrencia debajo de la media aritmética. Las sílabas comunes agrupadas en clases generales se listan en la tabla 6-1.

El experimento de esta sección consiste en modelar sílabas comunes usando bigrams. El número de sílabas comunes fue de 49, con el cual se entrenaron modelos bigram y se adaptaron al contexto CONMAT usando el procedimiento explicado en la sección 5.3.3 (las sílabas no comunes no fueron consideradas en el entrenamiento de los bigrams). Los resultados de este experimento se muestran en la siguiente sección.

6.4 Modelado de Fillers de Sílabas Comunes con Múltiples Pronunciaciones

Como variante al experimento anterior, se propone otro experimento con las sílabas comunes presentadas en la tabla 6-1, en un esfuerzo por proporcionar al reconocedor más herramientas para modelar los fillers en el habla fuera del vocabulario. El experimento consiste en modelar sílabas comunes con múltiples pronunciaciones. El esquema que usa este experimento es igual al que se propone en la sección 6.1.

Las sílabas con sus diferentes pronunciaciones se muestran en la tabla 6-2, las pronunciaciones que no aparecen en esta tabla y que pueden ser generadas a partir de la figura 6-2 (por ejemplo: /a ll/, /i l, /e ll/) no fueron consideradas debido a que no forman sílabas. En la siguiente sección se muestran los resultados generados al evaluar la técnica de fillers de sílabas con los experimentos mencionados en este capítulo.

SÍLABAS COMUNES	FRECUENCIA RELATIVA
de	7.94%
te	7.46%
se	5.49%
to	5.11%
a	4.73%
ta	4.72%
en	3.57%
do	3.50%
e	3.48%
ne	3.29%
la	3.10%
no	2.78%
na	2.76%
da	2.64%
so	2.36%
o	2.07%
es	1.99%
rxa	1.93%
ton	1.84%
rxe	1.80%
sa	1.57%
lo	1.55%
el	1.47%
le	1.42%
seon	1.32%
trxe	1.27%
los	1.27%
torx	1.16%
trxo	1.04%
des	0.99%
nen	0.91%
nas	0.90%
las	0.88%
seo	0.85%
tan	0.85%
dos	0.84%
nes	0.83%
sen	0.83%
tos	0.81%
rxo	0.72%
on	0.71%
serx	0.69%
terx	0.68%
nos	0.67%
trxa	0.67%
sea	0.65%
sos	0.64%
al	0.62%
ten	0.62%

Tabla 6-1 Sílabas comunes agrupadas en clases generales de fonemas.

SÍLABA COMUN	PRONUNCIACIONES
a	/a/
al	/a l/
da	/d a/, /v a/, /g a/
de	/d e/, /v e/, /g e/, /d i/, /v i/, /g i/
des	/d e s/, /v e s/, /g e s/, /d i s/, /v i s/, /g i s/
do	/d o/, /v o/, /g o/, /d u/, /v u/, /g u/
dos	/d o s/, /v o s/, /g o s/, /d u s/, /v u s/, /g u s/
e	/e/, /i/
el	/e l/
en	/e n/, /i n/, /e m/, /i m/
es	/e s/, /i s/
la	/l a/, /ll a/
las	/l a s/, /ll a s/
le	/l e/, /ll e/
lo	/l o/, /ll o/
los	/l o s/, /l u s/, /ll o s/
na	/n a/, /m a/, /nn a/
nas	/n a s/, /m a s/
ne	/n e/, /n i/, /m e/, /m i/
nen	/n e n/, /n i n/, /m e n/, /m i n/
nes	/n e s/, /n i s/, /m e s/, /m i s/
no	/n o/, /n u/, /m o/, /m u/
nos	/n o s/, /m o s/, /m u s/
o	/o/, /u/
on	/o n/, /o m/
rxa	/rx a/, /rx a /
rxe	/rx e/, /rr i/, /rx e/, /rx i/
rxo	/rx o/, /rr u/, /rx o/, /rx u/
sa	/s a/, /ch a/, /f a/, /x a/
se	/s e/, /s i/, /ch e/, /f e/, /x e/, /ch i/, /f i/, /x i/
sea	/s e a/, /s i a/, /ch i a/, /f i a/, /x i a/
sen	/s e n/, /s i n/, /ch e n/, /f e n/, /f i n/, /x e m/
seo	/s e o/, /s e u/, /s i u/, /f i u/, /x i o/
seon	/s e o n/, /s i o n/, /x i o n/
serx	/s e rx/, /s i rx/, /f e rx/, /f i rx/, /x e rx/, /x i rx/
so	/s o /, /s u/, /ch o/, /ch u/, /f o/, /f u/, /x o/, /x u/
sos	/s o s/, /s u s/, /ch o s/, /f u s/, /x o s/, /x u s/
ta	/t a/, /p a/, /k a/
tan	/t a n/, /t a m/, /p a n/, /k a n/
te	/t e/, /t i/, /p e/, /p i/, /k e/, /k i/
ten	/t e n/, /t i n/, /p e n/, /p i n/, /k i n/
terx	/t e rx/, /t i rx/, /p e rx/
to	/t o /, /t u/, /p o/, /k o/, /k u/
ton	/t o n/, /t u n/, /p o n/, /p u n/, /k o n/, /k u n/, /k u m/, /t u m/, /k o m/
torx	/t o rx/, /t u rx/, /p o rx/, /k o rx/, /k u rx/
tos	/t o s/, /t u s/, /p o s/, /k o s /
trxa	/t rx a/, /p rx a/, /k rx a/
trxe	/t rx e/, /t rx i/, /p rx e/, /p rx i/, /k rx e/, /k rx i/
trxo	/t rx o/, /t rx u/, /p rx o/, /p rx u/, /k rx o/, /k rx u/

Tabla 6-2 Sílabas comunes agrupadas en clases generales con múltiples pronunciaciones.

6.5 Resultados de experimentos

Los resultados obtenidos de la técnica propuesta en este capítulo, se pueden observar en las figuras 6-3 y 6-4. Los resultados se muestran por figura de mérito (FOM) y costo computacional.

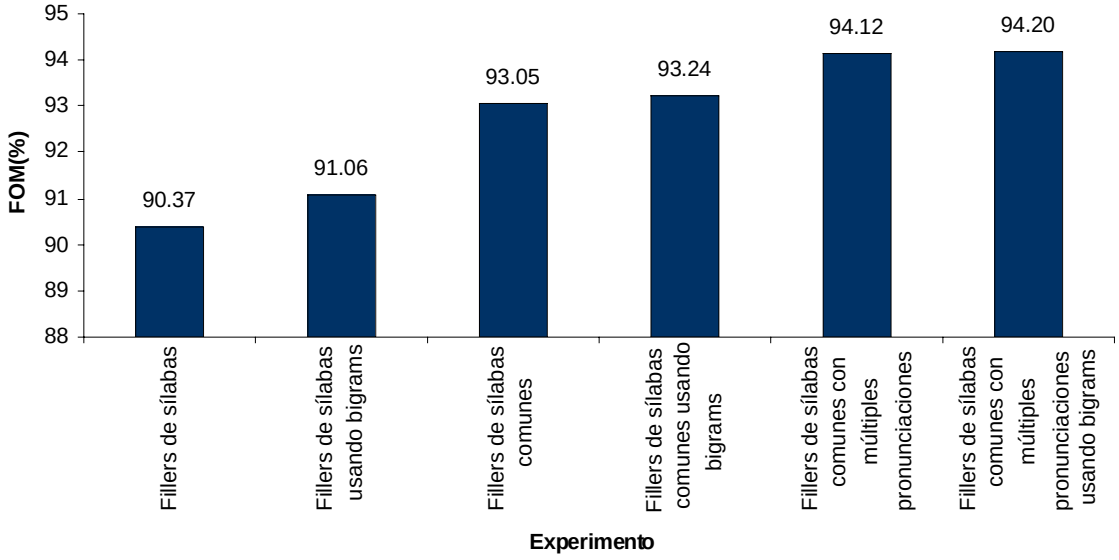

Figura 6-3 Figura de mérito de los experimentos con la técnica de sílabas como fillers.

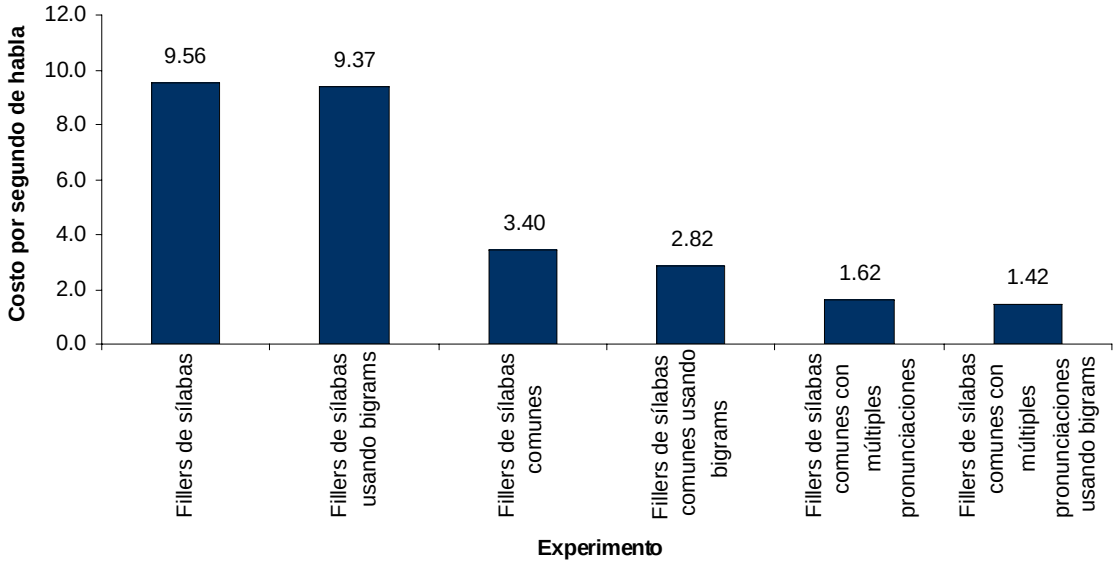

Figura 6-4 Costo computacional de los experimentos de la técnica de sílabas como fillers.

Todos los experimentos ilustrados en las figuras 6-3 y 6-4 usan los mismos umbrales de confianza que los experimentos base antes de ser optimizados (550 para el umbral bajo y 900 para el umbral alto). En la tabla 6-3, se muestra un resumen de los resultados de las medidas de desempeño de cada experimento.

EXPERIMENTO	rr_in	ca_in	cc_in	fa_in	fr_in	fc_in	crout	faout	fcout	FOM
Gramática predeterminada	0.857	0.579	0.258	0.009	0.067	0.087	0.365	0.115	0.519	91.59%
Fillers de sílabas	0.617	0.278	0.327	0.018	0.345	0.031	0.875	0.038	0.087	90.37%
Fillers de sílabas usando bigrams	0.590	0.216	0.359	0.007	0.412	0.007	0.981	0.000	0.019	91.06%
Fillers de sílabas comunes	0.757	0.465	0.278	0.004	0.174	0.078	0.750	0.019	0.231	93.05%
Fillers de sílabas comunes usando bigrams	0.764	0.481	0.263	0.013	0.160	0.082	0.760	0.010	0.231	93.24%
Fillers de sílabas comunes con múltiples pronunciaciones	0.817	0.575	0.223	0.007	0.096	0.100	0.644	0.010	0.346	94.12%
Fillers de sílabas comunes con múltiples pronunciaciones usando bigrams	0.817	0.581	0.218	0.007	0.091	0.102	0.654	0.010	0.337	94.20%

Tabla 6-3 Resultados de las medidas de desempeño de los experimentos con la técnica sílabas como fillers.

De acuerdo a las figuras 6-3 y 6-4, se puede observar que el experimento de *fillers de sílabas comunes usando múltiples pronunciaciones usando bigrams* es el que obtiene el mejor desempeño, equilibrando desempeño de reconocimiento y costo computacional. Al optimizar los umbrales de confianza sobre el mejor experimento se obtuvieron los resultados que se muestran en la tabla 6-4. Los umbrales de confianza optimizados fueron **620** para el umbral bajo y **880** para el umbral alto.

EXPERIMENTO	rr_in	ca_in	cc_in	fa_in	fr_in	fc_in	Crout	faout	fcout	FOM
Fillers de sílabas comunes con múltiples pronunciaciones usando bigrams	0.817	0.604	0.171	0.007	0.149	0.069	0.808	0.010	0.183	94.48%

Tabla 6-4 Medidas de desempeño del experimento *fillers de sílabas comunes con múltiples pronunciaciones usando bigrams* (después de optimizar los umbrales de confianza).

6.5.1 Análisis de Resultados

Los resultados del mejor experimento de fillers de sílabas clasificados por nodos terminales se muestran en la tabla 6-5.

NODO TERMINAL	rr_in	ca_in	cc_in	fa_in	fr_in	fc_in	crout	faout	fcout	# FRASES
T1	0.000	0.000	0.000	0.000	0.500	0.500	NaN	NaN	NaN	2
T2	0.182	0.000	0.182	0.000	0.545	0.273	NaN	NaN	NaN	11
T3	NaN	NaN	NaN	NaN	NaN	NaN	0.892	0.000	0.108	65
T4	NaN	NaN	NaN	NaN	NaN	NaN	0.667	0.000	0.333	6
T5	NaN	NaN	NaN	NaN	NaN	NaN	0.667	0.030	0.303	33
T6	1.000	0.600	0.400	0.000	0.000	0.000	NaN	NaN	NaN	10
T7	0.880	0.662	0.170	0.005	0.113	0.050	NaN	NaN	NaN	399
T8	0.000	0.000	0.000	0.000	0.000	1.000	NaN	NaN	NaN	1
T9	0.154	0.038	0.115	0.038	0.577	0.231	NaN	NaN	NaN	26

Tabla 6-5 Resultados de reconocimiento del experimento *fillers de sílabas comunes con múltiples pronunciaciones usando bigrams,* clasificados por nodos terminales usando los niveles de confianza optimizados para este experimento (620 y 880).

A partir de los resultados de reconocimiento de la tabla 6-5 y comparados con el experimento base que usa una gramática predefinida (ver tabla 4-5), se pueden hacer las siguientes afirmaciones:

- Buen desempeño de rechazo en frases con habla fuera del vocabulario (T3, T4 y T5).
- El desempeño con frases que contienen solo palabras clave, casi permanece sin cambios (T6 y T7).
- Permanece bajo desempeño con frases que contienen palabras no clave combinadas con palabras clave (T1, T2, T8 y T9). Aunque el desempeño es un poco mejor en el experimento que modela fonemas usando bigrams.

6.6 Resumen

En este capítulo, se hacen varios experimentos con la técnica de sílabas como fillers. Entre los cuales se destacan los experimentos de sílabas, sílabas comunes y sílabas comunes con múltiples pronunciaciones. Otros experimentos son similares a los antes descritos, con la diferencia de que agregan modelos de lenguaje bigram con la finalidad de disminuir el costo computacional que representan los modelos filler. Además, se hace una clasificación de las sílabas que son usadas como fillers. Por ultimo, se hace una evaluación de todos los experimentos en donde el mejor con respecto a desempeño de reconocimiento y costo computacional es el de *fillers de sílabas comunes con múltiples pronunciaciones usando bigrams*.

El siguiente capítulo explica la técnica de identificación de palabras clave usando palabras completas como fillers. Se hacen varios experimentos combinando palabras no clave con fonemas y sílabas, en un intento por modelar las palabras no clave desconocidas.