Capítulo 5

5 Identificación de Palabras Clave usando Fonemas como Fillers

Aquí se investiga y experimenta la técnica de identificación de palabras clave usando fonemas como fillers. También se introduce el uso de modelos de lenguaje estocásticos para mejorar el desempeño de los fillers. En este capítulo se hace un estudio sobre los fonemas en el español hablado en México, con el fin de hacer una clasificación de los fonemas que serán usados como fillers. Al final de este capítulo, los experimentos son comparados con los resultados de los experimentos base.

5.1 Modelado de fillers con unidades inferiores a la palabra

Una de las soluciones al problema de identificación de palabras clave en habla continua se basa en el reconocimiento de un conjunto de unidades lingüísticas fundamentales, escogidas de forma que todas las palabras de un lenguaje puedan ser representadas con sus correspondientes sonidos acústicos. El modelado con unidades inferiores a la palabra de fillers tiene dos importantes ventajas:

- Permite frases con vocabularios flexibles en los sistemas de identificación de palabras clave, ya que todas las palabras no clave pueden ser modeladas mediante un conjunto de unidades fundamentales.
- El modelado con unidades inferiores a la palabra facilita la difícil tarea de predecir las palabras no clave para cada nueva aplicación cuando se modela habla fuera del vocabulario (por ejemplo, en CONMAT se tuvo que diseñar una gramática especial, ver Apéndice A1).

Los fonemas representan la unidad mínima del habla. En el español hablado en México, se pueden definir aproximadamente 22 fonemas. La siguiente sección explica los fonemas que son usados como fillers por esta técnica.

5.1.1 Clasificación de los Fonemas

El lenguaje hablado puede representarse como una serie de unidades básicas de sonido llamadas fonemas. Existen diferentes formas para clasificar a los fonemas. Los fonemas se separan principalmente en dos grupos: vocales y consonantes. A partir de esta separación, se hace una clasificación mas detallada considerando el modo de articulación, lugar de articulación y la presencia de vibraciones de las cuerdas bucales (ver figura 5-1). Una descripción mas detallada puede encontrarse en [Quilis y Fernández, 1964]

Brevemente, el modo de articulación separa las consonantes en cinco grupos: *oclusivas, fricativas, africadas, nasales y líquidas (laterales y vibrantes).*

En cuanto al lugar de articulación, a las vocales se les asignan los términos de anterior, central y posterior, y de abierta, media y cerrada. El lugar de articulación de las consonantes se define según los órganos que actúen, así como la zona donde incidan. Se pueden clasificar en seis grupos: *bilabiales, labiodentales, dentales, alveolares, palatales y velares.*

En cuanto a las vibraciones de las cuerdas bucales, las consonantes pueden ser clasificadas como *sordas y sonoras.*

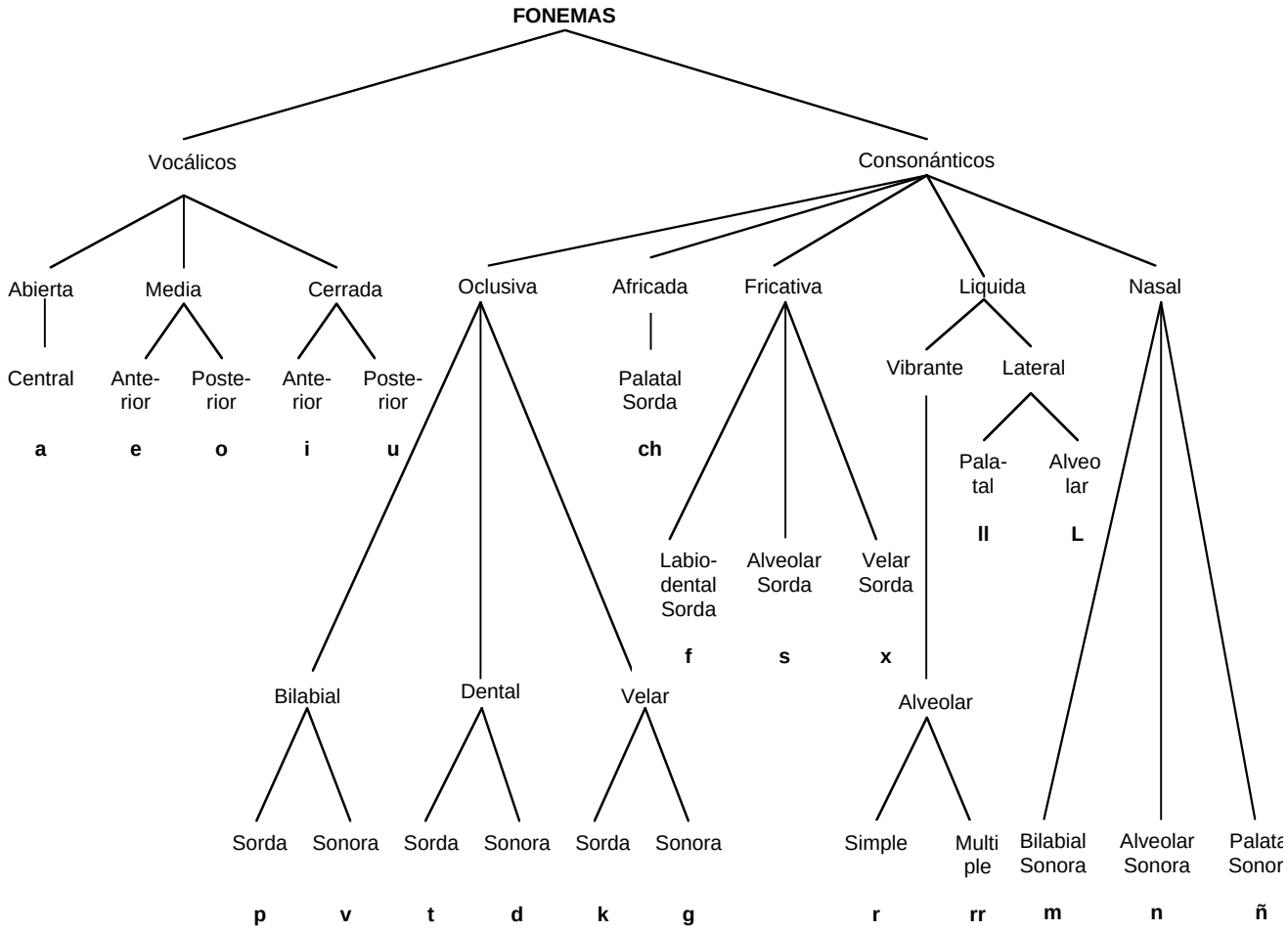

Figura 5-1 Clasificación de los fonemas del español hablado en México.

5.2 Modelado de Fillers de Fonemas

Aquí se describe la primera variante de la técnica de identificación de palabras clave usando fonemas como fillers. Consiste en usar los fonemas descritos en la sección anterior como fillers y modelar palabras clave completas, ya que las palabras clave son exactamente lo que se desea reconocer. Un esquema genérico de esta técnica se ilustra en la figura 5-2. Cualquier transición entre fonemas y palabras clave es permitida, también se permiten transiciones entre fonema y fonema y entre palabra clave y palabra clave. La salida del sistema es una transcripción completa de la frase reconocida y el resultado es la palabra o palabras clave, mientras que los fillers son ignorados.

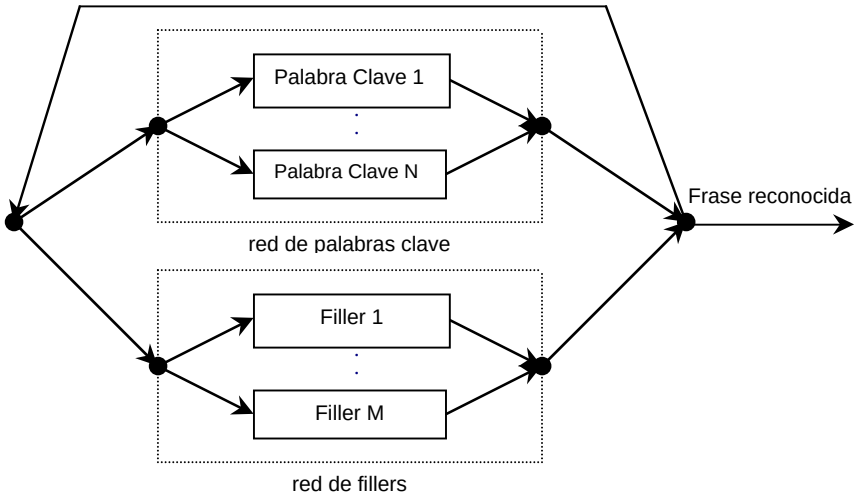

Figura 5-2 Esquema genérico de un sistema de identificación de palabras clave.

Para efectos de la estructura experimental, CONMAT solo se permite una palabra clave. Otras aplicaciones pueden tener varias palabras clave. Por lo tanto las gramáticas que se describen en esta tesis consideran solo una palabra clave. A continuación se ilustra el esquema usado por la técnica a experimentar en este capítulo.

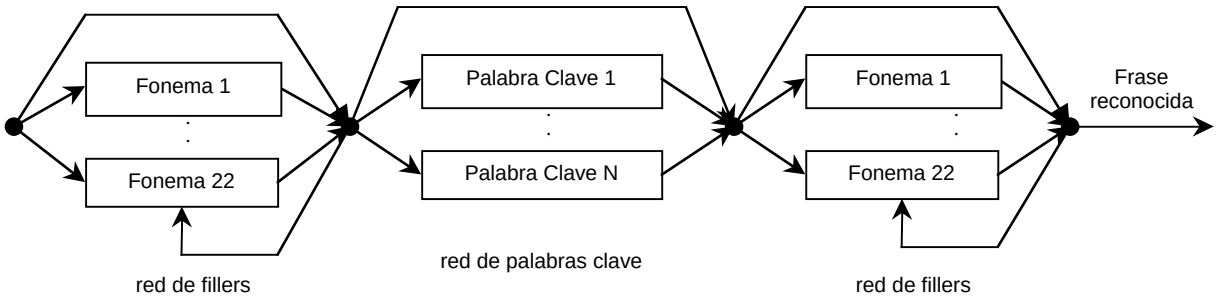

Figura 5-3 Esquema del sistema de identificación de palabras clave usando fonemas como fillers.

La gramática que describe esta técnica es mostrada en el apéndice A2, donde los fillers y la palabra clave son opcionales. Al final del capítulo anterior se explicó que uno de los problemas al que se enfrentan los sistemas de identificación de palabras clave son las frases con solo palabras no clave. El esquema de la figura 5-3 permite la omisión de la palabra clave. En el caso cuando la mejor hipótesis contiene solamente fillers, la frase es rechazada sin considerar el nivel de confianza. Por lo tanto con este mecanismo se mejora el desempeño de rechazos correctos en frases con habla fuera del vocabulario.

5.3 Modelado de Fillers de Fonemas usando Modelos de Lenguaje

La precisión de reconocimiento puede ser mejorada tomando ventaja de la posible información *a priori* en las secuencias a ser reconocidas. En reconocimiento de habla, conocimiento *a priori* puede significar lo siguiente:

- Las palabras permitidas suelen corresponder a un vocabulario predeterminado.
- Algunas secuencias de palabras son más probables que otras.

El objetivo de un modelo de lenguaje consiste en asignar probabilidades a las posibles secuencias de palabras. Estas probabilidades pueden amplificar la diferencia entre alternativas que son acústicamente similares, así que combinado con la técnica de pruning, disminuye el número de caminos vigentes en el espacio de búsqueda. Existen varios tipos de modelos de lenguaje, entre ellos se encuentran los *modelos uniformes*, *lenguajes de estado finito*, *gramáticas libres de contexto*, y *modelos estocásticos*.

Entre los diferentes tipos de modelos de lenguaje, los modelos estocásticos N-gram han probado ser efectivos en sistemas de reconocimiento de voz. Los modelos estocásticos N-gram son una buena solución para reducir significativamente el espacio de búsqueda del reconocedor, con lo que se reduce el costo computacional. Esta técnica es la más usada en aplicaciones de reconocimiento de grandes vocabularios en habla continua (LVCSR) [Becchetti y Prina, 1999].

5.3.1 Modelos Bigram

Existen varias razones para entrenar modelos 2-gram (*bigram*). Por un lado, las secuencias de fonemas en español son muy regulares, por ejemplo: /pe/, /ta/, /da/, son secuencias comunes de fonemas y /xp/, /tp/, /cb/ son secuencias poco común. Por otro lado el reconocimiento al nivel de fonemas es muy difícil debido a que un fonema esta

influenciado por sus fonemas colindantes. Además, resultados experimentales [Becchetti y Prina, 1999], muestran que el modelo bigram es una buena opción que produce resultados satisfactorios y que no presenta problemas en el procedimiento de estimación de probabilidades. Debido a esto, la segunda variante de identificación de palabras clave que se describe en esta sección agrega modelos bigram a la gramática de fonemas que modela el habla fuera del vocabulario.

El modelo bigram esta basado en la aproximación de que una palabra es estáticamente dependiente de la palabra anterior (en este capítulo una palabra corresponde a un fonema). El modelo bigram asigna una probabilidad a una secuencia de palabras *W* de acuerdo a la siguiente ecuación:

$$P(W)..P(w_1)\prod_{i=2}^{N} P(w_i / w_{i-1}) \tag{5.1}$$

donde la probabilidad de tener la palabra w_i cuando la palabra anterior es w_{i-1} esta dada por:

$$P(w_i / w_{i-1})...\frac{N(w_{i-1}, w_i)}{N(w_{i-1})} \tag{5.2}$$

En la ecuación anterior $N(w_i / w_{i-1})$ es el número de ocurrencias de la secuencia $\langle w_{i-1}, w_i \rangle$ en el conjunto de entrenamiento y $N(w_{i-1})$ es el número de ocurrencias de la palabra w_{i-1} en el mismo conjunto de entrenamiento.

5.3.2 Conjunto de Entrenamiento

Para construir un modelo N-gram, se requiere de un conjunto de entrenamiento. El conjunto de entrenamiento definido para experimentos en esta tesis consiste en un archivo de texto con 60,033 palabras en español conformando 306,724 fonemas. Dicho conjunto de entrenamiento contiene tópicos de salud, educación, política y literatura.

La tabla 5-1 lista los fonemas con su correspondiente definición fonológica, ejemplo y frecuencia relativa en este corpus.

FONEMA	DEFINICION FONOLÓGICA	EJEMPLO	FRECUENCIA RELATIVA
E	Vocal media anterior	mesa	13.99%
A	Vocal abierta central	caso	12.21%
S	Fricativa alveolar sorda	casa	9.87%
O	Vocal media posterior	modo	9.16%
I	Vocal cerrada anterior	piso	8.52%
N	Nasal alveolar sonora	nada	7.12%
Rx	Vibrante alveolar simple	pero	6.10%
D	Oclusiva dental sonora	donde	4.86%
t	Oclusiva dental sorda	tino	4.61%
l	Lateral alveolar sonora	lado	4.40%
k	Oclusiva velar sorda	casa	3.64%
m	Nasal bilabial sonora	mano	2.95%
u	Vocal cerrada posterior	cura	2.89%
p	Oclusiva bilabial sorda	punto	2.70%
b	Oclusiva bilabial sonora	baños	2.08%
g	Fricativa labiodental sorda	ganga	1.79%
ll	Lateral palatal sonora	pollo	0.93%
f	Fricativa labiodental sorda	falda	0.68%
x	Fricativa velar sorda	jamás	0.63%
rr	Vibrante alveolar compuesta	perro	0.56%
ch	Africada palatal sorda	chato	0.17%
ñ	Nasal palatal sonora	baño	0.13%

Tabla 5-1 Lista de fonemas del español hablado en México.

5.3.3 Modelos Bigram en el Contexto CONMAT

Además de las probabilidades fonema a fonema estimadas en la sección anterior, el reconocimiento puede mejorar su desempeño estimando las probabilidades de acuerdo al contexto en el que las frases de una aplicación pueden ocurrir (*silencio a fonema, fonema a palabra clave, etc.*). Estas probabilidades se entrenaron con datos de un corpus de 5835 frases, dentro de las cuales se encuentra el corpus de desarrollo.

Para entrenar los bigrams en el contexto CONMAT, se ajustan las probabilidades de los bigrams de acuerdo las posibles transiciones entre frases que pueden ocurrir (*pausa a fonema, fonema a palabra clave, etc.*) y se calculan las probabilidades de las transiciones usando la ecuación 5.2. A continuación se lista una clasificación de los tipos de frases que pueden ocurrir en sistemas con solo una palabra clave. La tabla 5-2 lista los componentes y el número de ocurrencias de estos tipos de frases:

- Habla (palabras y sonidos) fuera del vocabulario antes de una palabra clave
- Habla fuera del vocabulario después de una palabra clave
- Habla fuera del vocabulario antes y después de una palabra clave
- Sólo Palabra(s) clave
- Sólo habla fuera del vocabulario

COMPONENTES DE LA FRASE	NUMERO DE OCURRENCIAS
pausa_inicial + fonemas + palabra_clave + pausa_final	269
pausa_inicial + palabra_clave + fonemas + pausa_final	86
pausa_inicial + fonemas + palabra_clave + fonemas + pausa_final	78
pausa_inicial + palabra_clave + pausa_final	1663
pausa_inicial + fonemas + pausa_final	3739

Tabla 5-2 Clasificación de frases en el corpus de CONMAT.

Las ecuaciones 5.3 y 5.4 son utilizadas para calcular las probabilidades de las transiciones que inician con fonema. A continuación se ilustran las probabilidades de las transiciones de pausa a pausa y de fonema a palabra clave. Las demás probabilidades son calculadas de manera similar:

$$P(fonema / pausa) = \frac{N(pausa, fonema)}{N(pausa)} \tag{5.3}$$

$$P(fonema / palabra\_clave) = \frac{N(palabra\_clave, fonema)}{N(palabra\_clave)} \qquad (5.4)$$

La tabla 5-3 muestra de manera general los bigrams de fonemas en el contexto CONMAT, el bigram completo agrega las transiciones de fonema a fonema.

	PAUSA	FONEMA	PALABRA CLAVE
PAUSA	0	0.51	0.49
FONEMA	0.91	P(f1/f2)	0.09
PALABRA CLAVE	0.95	0.05	0

Tabla 5-3 Bigrams de fonemas en el contexto CONMAT.

5.4 Resultados de experimentos

Los resultados obtenidos con la técnica propuesta en este capítulo se pueden observar en las figuras 5-4 y 5-5. Los resultados se muestran por figura de mérito (FOM) y costo computacional. Todos los experimentos usan los niveles de confianza de los experimentos base antes de ser optimizados (550 en el umbral bajo y 900 en el umbral alto).

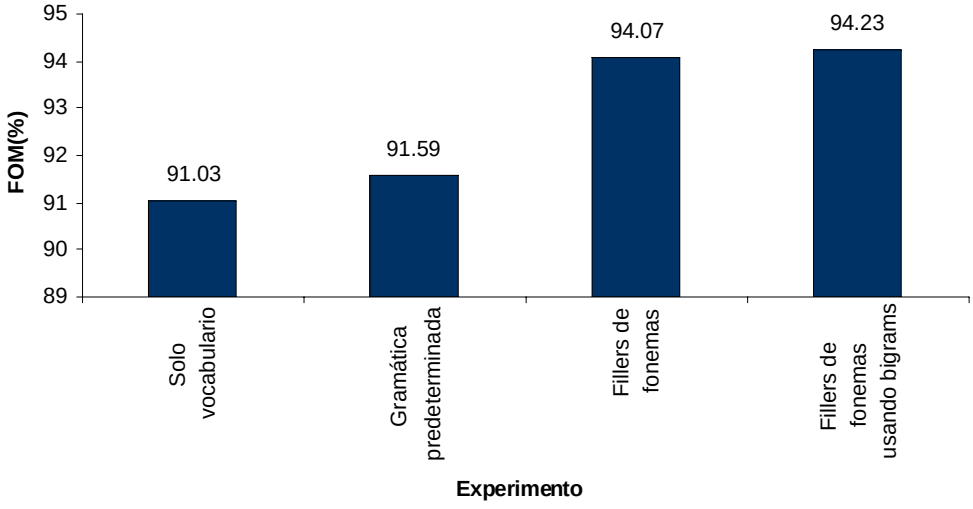

Figura 5-4 Figura de mérito de los experimentos con la técnica de fonemas como fillers.

En las figuras 5-4 y 5-5 se hace una comparación de los resultados obtenidos con los resultados de los experimentos base (solo vocabulario y gramática predefinida). En la tabla 5-4, se muestra un resumen de los resultados de las medidas de desempeño de cada experimento.

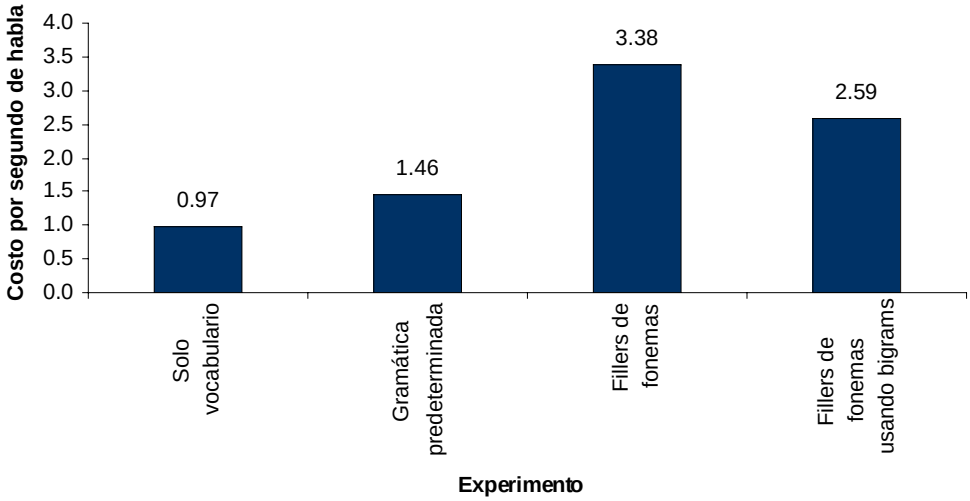

Figura 5-5 Costo computacional de los experimentos de la técnica de fonemas como fillers.

EXPERIMENTO	rr_in	ca_in	cc_in	fa_in	fr_in	fc_in	crout	faout	fcout	FOM
Solo vocabulario	0.811	0.581	0.214	0.018	0.087	0.100	0.385	0.115	0.500	91.03%
Gramática predeterminada	0.857	0.579	0.258	0.009	0.067	0.087	0.365	0.115	0.519	91.59%
Fillers de fonemas	0.817	0.566	0.229	0.009	0.102	0.094	0.654	0.010	0.337	94.07%
Fillers de fonemas usando bigrams	0.822	0.568	0.232	0.004	0.102	0.094	0.673	0.010	0.317	94.23%

Tabla 5-4 Medidas de desempeño de los experimentos con la técnica fonemas como fillers.

De acuerdo a las figuras 5-4 y 5-5, se puede observar que el experimento de *fillers de fonemas usando bigrams* es el que obtiene el mejor desempeño y a la vez el que presenta menos costo computacional. Al optimizar los umbrales de confianza en este experimento, se obtuvieron los resultados que se muestran en la tabla 5-5. Los umbrales de confianza optimizados fueron: **620** para el umbral bajo y **860** para el umbral alto.

EXPERIMENTO	rr_in	ca_in	cc_in	fa_in	fr_in	fc_in	Crout	faout	fcout	FOM
Fillers de fonemas usando bigrams	0.822	0.608	0.167	0.009	0.151	0.065	0.808	0.010	0.183	94.5%

Tabla 5-5 Medidas de desempeño del experimento *fillers de fonemas usando bigrams,* después de optimizar los umbrales de confianza.

5.4.1 Análisis de Resultados

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

NODO TERMINAL	rr_in	ca_in	cc_in	fa_in	fr_in	Fc_in	crout	faout	fcout	# FRASES
T1	0.500	0.500	0.000	0.000	0.000	0.500	NaN	NaN	NaN	2
T2	0.273	0.091	0.182	0.091	0.455	0.182	NaN	NaN	NaN	11
T3	NaN	NaN	NaN	NaN	NaN	NaN	0.954	0.000	0.046	65
T4	NaN	NaN	NaN	NaN	NaN	NaN	0.500	0.000	0.500	6
T5	NaN	NaN	NaN	NaN	NaN	NaN	0.576	0.030	0.394	33
T6	1.000	0.800	0.200	0.000	0.000	0.000	NaN	NaN	NaN	10
T7	0.870	0.649	0.170	0.005	0.118	0.058	NaN	NaN	NaN	399
T8	0.000	0.000	0.000	0.000	1.000	0.000	NaN	NaN	NaN	1
T9	0.308	0.154	0.115	0.038	0.577	0.115	NaN	NaN	NaN	26

Tabla 5-6 Resultados de reconocimiento del experimento *Fillers de fonemas usando bigrams*, clasificados por nodos terminales con niveles de confianza optimizados (620 y 860).

Comparando estos resultados con los del experimento base que usa una gramática predefinida (tabla 4-5), se pueden hacer las siguientes afirmaciones:

- La gramática definida para el experimento de fonemas como fillers muestra un desempeño significativo para rechazar frases que contienen únicamente habla fuera del vocabulario (T3, T4 y T5).
- El desempeño con frases que contienen solo palabras clave, casi permanece sin cambios (T6 y T7).

- Bajo desempeño con frases que contienen palabras no clave combinadas con palabras clave (T1, T2, T8 y T9). Debido a que la búsqueda prefiere a los fonemas en vez de la palabra clave.

5.5 Resumen

En este capítulo, se investiga la técnica de identificación de palabras clave usando fonemas como fillers, en donde se hacen experimentos con dos variantes. El primer experimento modela fillers de fonemas y en el segundo se agrega el uso de modelos de lenguaje estocásticos bigram, en un esfuerzo por disminuir el costo computacional. En este capítulo, se hace una clasificación de los fonemas del español hablado en México. Por ultimo, se reporta que el experimento de *fillers de fonemas usando bigrams* presenta mejores resultados tanto en desempeño como en costo computacional.

El siguiente capítulo explica y experimenta la técnica de identificación de palabras clave usando sílabas, en un esfuerzo por mejorar el desempeño presentado en este capitulo.