SELECCIÓN DE UNIDADES PARA EL RECONOCIMIENTO CONTINUO DEL HABLA

Eugenio Vives

Los avances tecnológicos de las últimas décadas han permitido el desarrollo de nuestros sistemas informáticos, electrodomésticos y máquinas en general hasta límites insospechados. Habitualmente nos comunicamos con ellos a través de órdenes tecleadas, pulsaciones de botones al fin y al cabo; sin embargo, el habla representa el medio de comunicación más fácil, rápido y corriente entre los seres humanos. Surge entonces el reto de intentar dotar a nuestros sistemas de la capacidad de "reconocer" órdenes y mensajes transmitidos de viva voz.

Como su propio nombre indica, un sistema de reconocimiento debe "volver a conocer". Para ello es necesario que previamente hallamos suministrado información al sistema sobre los mensajes que previsiblemente va a recibir. Esta tarea se realiza en la denominada fase de entrenamiento o aprendizaje creando un **número manejable de modelos de unidades básicas.** En la fase de reconocimiento propiamente dicha el sistema debe comparar de alguna forma el mensaje recibido con los modelos creados en la fase de entrenamiento.

El éxito de un sistema de reconocimiento del habla se basa principalmente en los factores que a continuación se resumen:

Codificación de la señal de voz.

Se debe extraer de la señal de voz aquellas características que son imprescindibles para lograr el reconocimiento. Es decir, no nos interesan aquellos parámetros que distinguen una voz masculina de una femenina pero sí es importante tener bien caracterizadas las diferencias entre una "p" y una "t" o entre un sonido sordo y uno sonoro por ejemplo. Las codificaciones más habituales se suelen hacer en el dominio frecuencial, siendo las más frecuentes los coeficientes predictivos lineales (LPC) y la codificación por banco de filtros [1].

Creación de referencias.

Se deben **modelar las unidades** elegidas para el reconocimiento a partir de la codificación de un conjunto representativo de pronunciaciones de estas unidades. Las referencias más utilizadas hoy en día por la mayoría de los sistemas de reconocimiento son los Modelos Ocultos de Markov [2].

Surge [...] el reto de intentar dotar a nuestros sistemas de la capacidad de "reconocer" órdenes y mensajes transmitidos de viva voz.

Algoritmos de comparación.

Se debe disponer de algoritmos eficientes para comparar la señal recibida con las referencias creadas en la fase anterior. (Uno de los más habituales es una adaptación del algoritmo de Viterbi [3]).

En el presente escrito nos vamos a ocupar de explicar los diferentes tipos de unidades básicas que utilizan algunos de los sistemas desarrollados hasta la fecha y de los criterios que rigen la elección de las mismas.

1. Unidades propuestas para el reconocimiento del habla.

Las unidades que se nos ocurriría proponer en una primera aproximación al problema serían seguramente las palabras y los fonemas. En la discusión siguiente veremos algunas ventajas e inconvenientes de éstas y otras unidades como fonemas, trifonemas, sílabas, semisílabas etc. que han sido propuestas en diferentes trabajos como alternativas a las dos primeras.

1.1- Palabras.

Las palabras son la unidad más natural que se nos puede ocurrir puesto que son exactamente lo que queremos reconocer. Presentan la ventaja de capturar bien las pronunciaciones diferentes de un mismo fonema que se pueden hacer en el seno de una palabra. Cuando es posible entrenar adecuadamente los modelos

EUGENIO VIVES LAMARCA está realizando actualmente el Proyecto Fin de Carrera en el Grupo de Procesado de Voz del Departamento de Teoría de Señal y Comunicación.

de palabras, se suelen alcanzar los mejores resultados, sin embargo cuando queremos implementar un sistema de reconocimiento **para amplios vocabularios no es factible el uso de modelos de palabras** puesto que necesitaríamos una enorme cantidad de datos de entrenamiento para que cada palabra apareciese un número de veces considerable.

Por otra parte, en algunas aplicaciones es conveniente contar con la opción de que el usuario añada nuevas palabras al vocabulario. Utilizando modelos de palabras se necesitarían muchas repeticiones de la palabra nueva para que pueda ser entrenada adecuadamente, con lo que el coste de adaptación puede ser inaceptable.

Cuando queremos implementar un sistema de reconocimiento para amplios vocabularios no es factible el uso de modelos de palabras.

1.2- Fonemas.

La unidad subléxica con la que estamos más familiarizados es el fonema. Los fonemas presentan la ventaja, debido a su reducido número, de que pueden ser entrenados fácilmente con unos pocos cientos de frases. Los fonemas, sin embargo, no suelen ser una unidad adecuada por sí mismos, ya que se asume que un fonema en cualquier contexto es equivalente al mismo fonema en cualquier otro contexto. Este hecho no es cierto, ya que cuando pronunciamos un fonema, no lo hacemos independientemente del contexto en que éste está inmerso, pues nuestras articulaciones no se pueden mover instantáneamente de una posición a otra.

Hemos visto que mientras a los modelos de palabras les falta generalidad, los modelos de fonemas generalizan en exceso. En los siguientes apartados intentaremos buscar unas unidades que no tengan los defectos anteriores.

1.3- Unidades multifónicas.

Una manera de tener en cuenta los efectos coarticulatorios que se dan dentro de una palabra es usando unidades más largas. Algunos sistemas de reconocimiento en castellano proponen el uso de modelos de sílabas o semisílabas para aprovechar el carácter marcadamente silábico de esta lengua [4]. El problema que se nos presenta es el elevado número de unidades que tendríamos que modelar (del orden de 20.000 sílabas y 1.000 semisílabas en inglés).

1.4- Modelado de las transiciones.

Un intento de solución al problema que representa coarticulación sería modelar explícitamente las transiciones de un fonema a otro, eliminando la parte estacionaria de los mismos. De esta idea surge la definición de difonema.

1.5- Fonemas dependientes de la palabra.

El modelado de fonemas dependientes de la palabra pretende alcanzar un compromiso entre los modelos de palabras y los de fonemas. En estos modelos, un fonema que aparece en una palabra se representa de forma diferente que el mismo fonema en otra palabra [5].

El modelado con fonemas dependientes de la palabra es más eficiente que el de modelos de palabras desde dos puntos de vista. En primer lugar, si una palabra aparece muy pocas veces, sus parámetros pueden ser interpolados o promediados con los de los modelos de fonemas independientes de la palabra. En segundo lugar, si se desea añadir una nueva palabra al vocabulario, no es necesario repetirla muchas veces, ya que con el conjunto básico de fonemas independientes podremos conseguir una tasa de reconocimiento aceptable.

1.6- Fonemas dependientes del contexto.

Por contexto normalmente se entiende el fonema vecino inmediatamente anterior, el posterior o ambos. Así, tendremos modelos de fonemas dependientes del contexto izquierdo, del derecho o de ambos (trifonemas) [5].

Los modelos de trifonemas suelen estar pobremente entrenados porque hay muchos trifonemas. Como en el caso anterior, pueden ser interpolados con modelos más robustos. El modelado con trifonemas es muy potente porque tiene en cuenta los efectos coarticulatorios más importantes y es mucho más sensible que el modelo de fonemas.

En general, una de las ventajas de los fonemas dependientes del contexto con respecto a los fonemas dependientes de la palabra es que son más independientes de la aplicación que tratemos. Podemos seguir utilizando los mismos modelos aunque cambie el vocabulario a reconocer, en cambio, con los modelos dependientes de la palabra sería necesario realizar un nuevo entrenamiento si el nuevo vocabulario no es un subconjunto del viejo.

2. Un ejemplo práctico de elección de unidades.

Como muestra del problema que representa para el diseñador de un sistema de reconocimiento la elección adecuada de las unidades a modelar en función de la cantidad de entrenamiento disponible, presentamos el siguiente caso en donde nos plantearemos la viabilidad de usar un determinado modelo de unidades.

Se pretende diseñar dos tipos de aplicaciones: una, de carácter totalmente general en el que se permite una amplia gama de vocabulario y otra, más específica en la que las cuestiones a reconocer se limitan a

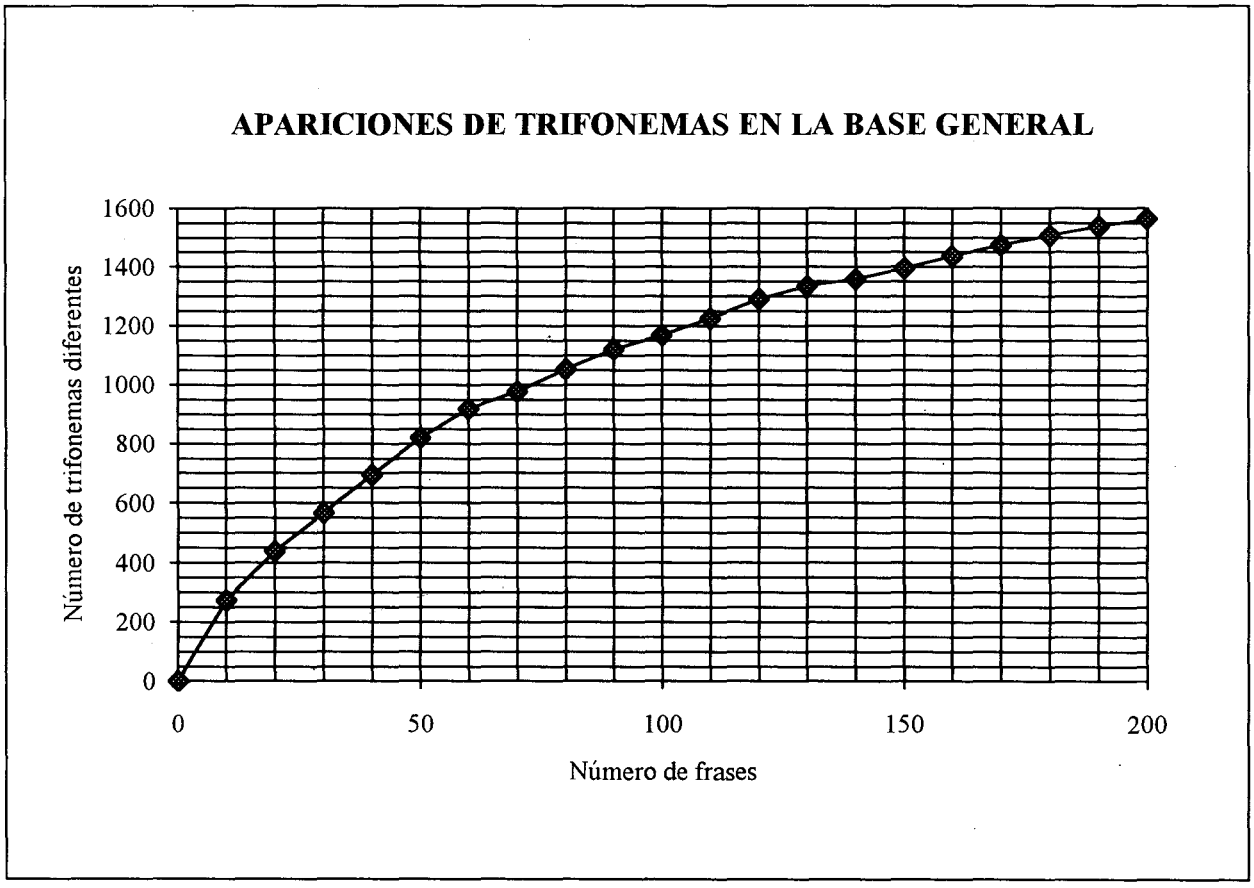
una temática concreta (en este caso, preguntas a una pequeña base de datos sobre la geografía española). Para ello se hace un estudio de algunas unidades (palabras, sílabas, trifonemas y fonemas) que aparecen en 200 frases utilizadas para el entrenamiento de ambas aplicaciones.

A continuación se presentan dos gráficas correspondientes al estudio de los trifonemas como unidad básica (recordemos que, a grandes rasgos, un trifonema es un fonema en donde se ha tenido en cuenta la influencia de su vecino derecho e izquierdo).

En las gráficas podemos observar como el número de trifonemas diferentes de la aplicación general tiende a crecer por encima de 1600 una vez analizadas las 200 frases, mientras que en la aplicación geográfica el número de éstos tiende a estabilizarse asintóticamente en torno a los 800 trifonemas diferentes (Para 500 frases se obtienen únicamente 774). Si consideramos que el número de unidades aparecidas en los dos corpus o conjuntos de frases es aproximadamente el mismo (en torno a los 8000 trifonemas) parece obvio deducir que en el segundo cada unidad aparece repetida más veces que en el primero. Si fijamos un umbral de 30 apariciones, en la aplicación geográfica nos aparecen 64 trifonemas que superan o igualan este número, mientras que en la general tan sólo lo hacen 38. Con estos trifonemas conseguimos modelar el 54% de la aplicación geográfica y tan sólo el 24.8% de la general. Situaciones análogas se repiten en el estudio de las palabras y de las sílabas como unidades básicas.

Podríamos pensar que para un tipo de aplicaciones muy específicas [...] las unidades modeladas pueden ser más largas y especializadas.

Podríamos considerar a partir de los datos anteriores, que los trifonemas, completados con modelos de fonemas, son una unidad adecuada para modelar la totalidad de la aplicación geográfica puesto que van a poder ser entrenados adecuadamente con unas decenas de repeticiones de las 200 frases. Por el contrario, parecen ser los trifonemas una unidad demasiado específica para la primera aplicación, en donde habría que limitarse quizás al uso de modelos más generales como los de fonemas. Extrapolando estos resultados podríamos pensar que para un tipo de aplicaciones muy específicas (como puedan ser una base de datos geográfica para consulta, un informador de vuelos y reservas en un aeropuerto o un gestor de compraventa de acciones en una Bolsa), las unidades mo-



deladas pueden ser más largas y especializadas, llegando incluso a poder ser conveniente el modelado de grupos de palabras de forma conjunta.

3. Resumen y conclusiones.

En esta breve exposición se han pretendido resaltar dos importantes propiedades de las unidades: el grado de sensibilidad con el que tienen en cuenta los efectos de coarticulación y la facilidad o dificultad de las mismas para ser entrenadas adecuadamente.

La elección de un conjunto de unidades adecuado para el reconocimiento no es en absoluto un problema resuelto y matematizado.

La elección de un conjunto de unidades adecuado para el reconocimiento no es en absoluto un problema resuelto y matematizado. La elección puede depender de causas tan dispares como: el volumen de frases del que dispongamos para el entrenamiento, la aplicación a la que está destinada el sistema de reconocimiento o el idioma en el que se desarrolle el mismo.

La mayoría de los trabajos actuales se nos presentan en un esquema similar: en primer lugar se definen un conjunto básico de unidades de tipo fonético ("baseline"). Posteriormente, y a partir de él, producen una ampliación del mismo, creándose unidades más especializadas que tengan en cuenta los efectos contextuales.

4. Referencias.

[1] L. R. RABINER : *Fundamentals of Speech Recognition*, Prentice Hall 1993, capítulo 3.

[2] L. R. RABINER: "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", Proceedings IEEE, vol 77, No. 2, Febrero 1989, págs. 257-284.

[3] A. J. VITERBI: "Error Bounds for Convolutional Codes and an Asymptotically Optium Decoding Algorithm", IEEE Transactions on Information Theory, Abril 1967, págs. 260-269.

[4] J. B. MARIÑO, A. BONAFONTE, A. MORENO, E. LLEIDA, J. HERRERO Y J. CASANOVAS: *RAMSES: Spanish Continuous Speech Recognition System.*

[5] LEE KAI-FU: *Automatic Speech Recognition, the Development of the SPHINX System*, Kluwer Academic Publishers, 1989, capítulo 6.