ISSN 2007-9737

Etiquetado fonético automático al nivel palabra usando la dinámica de cambio de los vectores del libro código

Sergio Suárez Guerra, José Luis Oropeza Rodríguez

Instituto Politécnico Nacional,
Centro de Investigación en Computación
México

{ssuarez, joropeza} @cic.ipn.mx

Resumen. Se describe una solución alternativa referente al etiquetado fonético que componen un conjunto de palabras de pronunciadas por un locutor, susceptible de utilizarse en cualquier idioma, según sean las necesidades y características asociadas a la propuesta. El procedimiento se basa en el seguimiento de la dinámica de cambio de los vectores cepstrales asociados a la frecuencia de Mel (MFCCs) que conforman el Libro Código (LC), extraído de la palabra a etiquetar. Esta dinámica de cambio analiza dónde ocurre una transición de un vector (MFCC) del LC a otro, así como las perturbaciones que ocurren en la zona de cambio debido a la concatenación fonética. Se establecen métricas para considerar el ruido de coarticulación y definir la ubicación de la frontera de separación fonética. Se usan dos métodos para evaluar la dinámica de cambio de los vectores y entregar el etiquetado más acertado. El porciento de reconocimiento y etiquetado correcto obtenido con esta aplicación es del 97.9%, inferior en un 1.06%, con respecto al porcentaje de reconocimiento obtenido sobre el mismo corpus de palabras, pero haciendo uso de un etiquetado manual. Lo más impórtate es que, el tiempo utilizado en el etiquetado del corpus de voz de forma automática, es significativamente menor que el estimado de hacerse manualmente, además de eliminar la subjetividad personal en el trabajo de etiquetado.

Palabras clave. Etiquetado fonético, reconocimiento de voz.

Automatic Phonetic Labeling at Word Level Using the Dynamics of Changing Codebook Vectors

Abstract: An alternative solution is described regarding the phonetic labeling that compose a set of pronounced by an announcer, susceptible of being used in any language, according to the needs and characteristics associated with the proposal. The procedure is based on the monitoring of the dynamics of change of the cepstral vectors associated with the frequency of Mel (MFCCs) that make up the Book Code (LC), extracted from the word to be labeled. This dynamics of change analyzes where a transition from one vector (MFCC) of the LC occurs to another, as well as the disturbances that occur in the zone of change due to the phonetic concatenation. Metrics are established to consider coarticulation noise and define the location of the phonetic separation boundary. Two methods are used to evaluate the dynamics of vector change and deliver the most accurate labeling. The percentage of recognition and correct labeling obtained with this application is 97.9% lower by 1.06%, with respect to the percentage of recognition obtained on the same corpus of words, but using manual labeling. The more important are that, the time used in the labeling of the voice corpus automatically is significantly less than the estimate of being done manually, in addition to eliminating personal subjectivity in the labeling work.

Keywords. Phonetic labeling, voice recognition.

1. Introducción

Una de las tareas básicas al diseñar un sistema de reconocimiento de voz (SRV), la constituye la generación los modelos de palabras que se reconocerán en el idioma predeterminado. Como una de las tareas iniciales, se crea el diccionario de palabras iniciales (corpus de palabras) y a partir de ahí comienza el trabajo de modelado de las mismas [1-2]. Para crear los modelos de palabras se parte de crear los modelos de fonemas que las constituyen, por lo que es deseable que el conjunto de palabras iniciales, contengan todos los

fonemas básicos que constituyen el idioma del lenguaje para el que se está realizando el SRV.

Las palabras habladas están constituidas por la concatenación de fonemas. La cantidad de fonemas por palabras es variada. Crear los modelos de fonemas es una tarea ardua y requiere de mucha atención, así como de gasto temporal humano. No todos los fonemas de un lenguaje pueden ser expresados oralmente por separados y si así se hiciera, se requiere de especialistas en foniatría que garanticen su auténtica pronunciación.

Por ello, se parte de que locutores voluntarios leen texto donde están las palabras que constituyen el conjunto de las palabras iniciales, y de las grabaciones de estas lecturas, se separan las palabras y se procesan de forma independiente, para extraer los segmentos fonéticos las constituyen.

Con la extracción de los segmentos fonéticos de las palabras, se forma el corpus fonético y a partir del corpus fonético se crean los modelos fonéticos del lenguaje. Con la concatenación de estos modelos de fonemas, se forman los modelos de las palabras a reconocer en una aplicación determinada. La creación de un corpus fonético se realiza separando los segmentos fonéticos que constituyen las palabras. Este procedimiento se nombra 'etiquetado fonético' (EF). El EF normalmente se realiza de forma manual, en aquellos casos en donde no se cuente con él o se trate de un corpus nuevo.

Lo anterior se hace utilizando, principalmente, herramientas de procesamiento digital de señales que posibilitan la visualización de señales, la marcación de los segmentos, su audición y archivos de transcripción fonética; en éste último se expresan explícitamente, las fronteras y el nombre de los fonemas entre las fronteras que contiene cada palabra.

El EF es una tarea de suma atención y considerable tiempo de trabajo de especialistas de alto nivel, ya que conlleva determinar dónde están las fronteras entre fonemas (zonas de coarticulación), lo cual no siempre es claro pues existe la concatenación de fonemas que son muy difíciles de separar. Una vez realizada la segmentación fonética y conociendo la transcripción fonética de cada palabra, se extraen los segmentos fonéticos correspondientes a cada fonema en cada palabra y se agrupan como los datos comunes a cada fonema. Para cada agrupación se obtiene el modelo del fonema que le corresponde. El EF automático se ha realizado mediante dos métodos principales: teniendo en cuenta los cambios acústicos con usando transformada wavelet [3-4] y teniendo en cuenta los cambios fonéticos mediante su reconocimiento, etiquetado forzado [5-7].

Los métodos que tienen en cuenta los cambios acústicos se basan en determinar los cambios de energía de las señales en diferentes bandas de frecuencia, principalmente en nuestros días, usando segmentación wavelets. La segmentación mediante los cambios fonéticos, requieren de una clasificación y entrenamiento del corpus fonético y representan la base sobre las que se fundamenta el presente trabajo.

Si se tiene en cuenta que para cada idioma (lenguaje) existente se tiene que realizar esta tarea, hay muchas horas-especialista en juego. Aquí formulamos algunas preguntas que fundamentan el objetivo de esta investigación: ¿Cómo ocurre en nuestro sistema nervioso central, el cerebro, la separación de los fonemas que componen una palabra y la modelación de los mismos, para luego crear los modelos de cada palabra? ¿Podemos crear un algoritmo que nos lleve a encontrar las fronteras fonéticas al interior de una concatenación de fonemas, independientemente de la palabra fuente y entregar la segmentación fonética que le corresponde a su transcripción fonética? Al aplicar el algoritmo antes ideado, es posible obtener un alto porcentaje del corpus fonético del idioma analizado.

Una vez que se tiene el corpus fonético del idioma a usar, es posible crear los modelos de palabras nuevas, diferentes a las que se utilizaron para crear el corpus fonético. En este momento es importante dejar claro la diferencia que existe entre esta propuesta y lo que se realiza en trabajos previos para la obtención del EF. En primera instancia, se trata de evitar el uso de modelos fonéticos pre-existentes para llevar a cabo el EF. Lo anterior, debido a que, como se comentó con antelación, ello requiere de la participación de especialistas en el área correspondiente.

En contraparte, si se considera, como es el caso en el presente trabajo, de considerar la variación

dinámica fonética de manera directa de la señal de voz, es posible encontrar las fronteras entre tales fonemas, de manera directa de la señal de voz hablada. Dado el razonamiento anterior, resulta adecuado y útil a considerarlo como una posible forma o manera en la que aprendemos a hablar.

Luego entonces, con la descripción anterior se pretende disponer de una aplicación que de forma automática, realice la segmentación fonética de las palabras habladas, a partir de la dinámica de cambio de las características de la señal de voz, cuya transcripción fonética es conocida, independiente del idioma de que se trate, lo cual es una aportación importante. Este es el objetivo central de este trabajo de investigación. Una vez obtenida la segmentación fonética es posible construir el corpus fonético del idioma.

2. Descripción de la tarea

Los fonemas articulados que constituyen las palabras en un idioma, se diferencian por las características de su contenido armónico, frecuencia y propiedades naturales, que los componen. Las características básicas de frecuencias de un fonema son únicas, pero no idénticas. Cada persona emite los fonemas de una forma muy aproximada a cualquier otra, pero no necesariamente los fonemas emitidos son exactamente iguales.

Una técnica para clasificar diferentes contenidos armónicos de una señal, es extraer los patrones característicos principales que la componen. Para ello se utilizan cotidianamente los MFCC [8] entre los más utilizados. Así el contenido armónico del segmento fonético de la señal de la vocal ‘a’ se diferencia sustancialmente de un segmento fonético de la señal de la silbante ‘s’. Igualmente hay segmentos fonéticos muy similares: el contenido armónico de la señal de la consonante sonora ‘l’ se parece mucho al de la señal de la vocal ‘u’. La dinámica de cambio de las características de los fonemas a lo largo de una palabra vocalizada, es una información principal en la detección de las fronteras fonéticas. Para el idioma español la cantidad de fonemas a considerar en un corpus fonético, dependen de la región y país que se trate.

Así para el español de México las más recientes propuestas contemplan los niveles de: T22, T29, T44 o T54 fonemas [Alfabeto MEXBET, Carlos UNAM, CIEMPIESS Corpus] [9], dependiendo si se consideran los fonemas básicos, básicos más las semivocales y lo anterior más las vocales tónicas más fonemas especiales derivados del Náhuatl. Una buena medida es considerar el caso de 22 fonemas más 5 alófonos y el fonema silencio ‘sil’ [10-11]. Sobre esta base se realizó este trabajo. En la tabla 1 se muestra la propuesta de fonemas que se utiliza en este trabajo de investigación.

2.1. Etiquetado automático de los fonemas en una palabra

Para realizar el etiquetado automático de los fonemas de una palabra se siguieron las siguientes etapas y técnicas:

- Elección de un conjunto de palabras cortas que contengan en conjunto todos los fonemas que componen el idioma utilizado, los cuales se mostraron en la tabla 1.
- Extracción de características MFCC de las palabras por segmentos de 20 mseg. Traslapados 2 mseg. Vector característico: energía, 15 coeficientes MFCC, deltas simples y dobles (diferencias). Vector de 45 características.
- Extracción del libro código mediante cuantificación vectorial VQ (VQ *Vector Quantization*, por sus siglas en inglés) para analizar la dinámica de cambio de los mismos. Uso de la transcripción fonética de la palabra a etiquetar para nombrar los segmentos fonéticos en la señal.
- Diseño de dos modelos de seguimiento de la dinámica de cambio de los vectores VQ al tiempo de producirse la vocalización de la palabra. Determinación del Inicio – Fin de los segmentos fonéticos.
- El tratamiento para solucionar el etiquetado automático en palabras que contienen fonemas explosivos: b, d, g, k, p, t.

Tabla 1. Fonemas utilizados en este trabajo

Descripción	Mexbet T22 + 6	Etiquetador
Vocal abierta	/a/	/a/
Vocal media palatal	/e/	/e/
Vocal cerrada palatal	/i/	/i/
Vocal media velar	/o/	/o/
Vocal cerrada velar	/u/	/u/
Alveolar lateral	/l/	/l/
Alveolar vibrante simple	/r(/	/r(/
Alveolar vibrante múltiple	/r/	/r/
Palatal fricativa sonora [j] -> propuesta usar j	/Z/	/j/
Labial nasal	/m/	/m/
Alveolar nasal	/n/	/n/
Palatal nasal -> propuesta usar ñ	/n~/	/ñ/
Palatal africada sorda [ch] -> propuesta usar H	/tS/	/H/
Palatal africada sonora (alófono) [ll, y] -> propuesta usar y	/tZ/	/y/
Bilabial oclusiva sorda	/p/	/p/
Dental oclusiva sorda	/t/	/t/
Velar oclusiva sorda	/k/	/k/
Bilabial oclusiva sonora	/b/	/b/
Dental oclusiva sonora	/d/	/d/
Velar oclusiva sonora	/g/	/g/
Bilabial aproximante sonora (alófono)	/V/	/V/
Dental aproximante sonora (alófono)	/D/	/D_/
Velar aproximante sonora (alófono)	/G/	/G_/
Labiodental fricativa sorda	/f/	/f/
Alveolar fricativa sorda	/s/	/s/
Velar fricativa sorda	/x/	/x/
Paravocal velar (alófono)	/w/	/w/
Silencio	sil	sil

- El tratamiento para solucionar el etiquetado automático en palabras que contienen el fonema ‘ch’ (H).
- Creación del archivo etiquetado palabra.lab correspondiente.

En la figura 1, se presenta el diagrama de flujo del algoritmo general para obtener el etiquetado automático de los fonemas de una palabra, contenidos dentro del corpus de palabras correspondiente.

2.1.1. Elección de un conjunto de palabras cortas que contengan en conjunto todos los fonemas que componen el idioma utilizado

Para que la propuesta antes indicada genere resultados con un alto nivel de rendimiento, se

recomienda que cada palabra del corpus a considerar contenga a lo sumo 4 fonemas, incluyendo los silencios 'sil' al inicio y fin de cada una de ellas. Es posible usar palabras hasta con cinco fonemas diferentes y algunos de ellos repetidos, casos de vocales preferiblemente. De acuerdo a los experimentos realizados, es recomendable evitar, en lo posible, que las palabras contengan dos fonemas explosivos. Una condición fuertemente necesaria, para crear un buen corpus fonético, es que las pronunciaciones de las palabras se produzcan con intensidad media, sin gritar y articulando correctamente los contenidos fonéticos de las mismas.

2.1.2. Extracción de características MFCC

Cada palabra se grabó en un archivo .wav con una frecuencia de muestreo fs = 11025 Hz. Se segmentó en unidades de 20 mseg. Traslapadas 2 mseg.

Se aplicó el algoritmo de cálculo de características MFCC] para cada segmento de 20 mseg. Se aplicó el cálculo de energía y se extrajo un vector MFCC de 14 coeficientes, sus deltas y doble deltas para tener al final un vector de tamaño de 45 elementos por cada segmento.

2.1.3. Extracción del libro código VQ, para analizar la dinámica de cambio de los mismos

Se aplicó el algoritmo LBG, propuesto por [Linde, Buzo and Gray, 1980] [12], éste agrupa un conjunto de vectores de entrenamiento M en un conjunto de vectores del libro de códigos (LC), $M=4=2^2$, en este trabajo.

El método de cálculo del LC entrega la cantidad de centroides optimizados $M=2^n$, siendo en este caso n=2. Los resultados finales son 4 centroides en este caso. Cada centroide optimizado final está asociado con un conjunto del total de los vectores M de la palabra, tal que: M=c1+c2+c3+c4, siendo c1, c2, c3 y c4 la cantidad de elementos de M que pertenecen a cada uno de los 4 centroides finales. Ahora bien, la cantidad de elementos asociados a cada centroide difiere y está relacionada con las características fonéticas que existen en los segmentos del audio de cada una de las palabras y duración de cada segmento fonético.

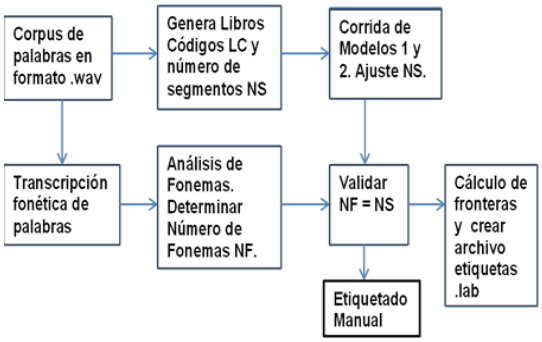

Fig. 1. Diagrama del algoritmo de etiquetado automático de una palabra

En la figura 2, se observa la distribución de centroides que se obtienen para n = 2. Para el caso de n = 0, primer nivel, el centroide está constituido por el vector medio de todos los vectores que integran la palabra.

En el nivel n = 1, segundo nivel, el vector medio de bipartición y se buscan los vectores en la vecindad de los nuevos vectores, se optimizan y se entrega el resultado de dos centroides. Para el nivel n = 2, tercer nivel, cada vector del nivel 1 se aplica el algoritmo de bipartición y se agrupan con sus vectores vecinos y optimizan, entregándose finalmente cuatro centroides. Es necesario observar que los vectores del nivel n = 1: 1 y 2; respectivamente, tienen como herederos en el nivel n = 2, los vectores c1 y c3 para el caso de 1 y c2 y c4 para el caso de 2. Esto se utilizará como un elemento de decisión importante en el modelo 1 para realizar la segmentación automática.

Los valores que representan a los centroides cambian en el transcurso de la producción de una señal de voz. De esta forma estamos considerando que se puede describir y encontrar las fronteras fonéticas, analizando la dinámica de cambio de los centroides (vectores o símbolos), a lo largo del tiempo de duración de cada palabra.

2.1.4. Uso de la transcripción fonética de la palabra a etiquetar para nombrar los segmentos fonéticos en la señal

La transcripción fonética de cada palabra consiste en disponer de forma escrita de la secuencia de caracteres que describen o contiene

cada palabra, a manera de ejemplo para la palabra ‘usa’ su descripción fonética es: ‘sil’, ‘u’, ‘s’, ‘a’, ‘sil’; o bien: ‘u’, ‘s’, ‘a’; si se considera que no se tendrán en cuenta los silencios anteriores y posteriores a la palabra.

El número de fonemas de la transcripción fonética es NF (número de fonemas) y se obtiene del archivo de la transcripción fonética de la palabra en análisis. El número de segmentos (NS) fonéticos de la señal de voz articulada, se determina usando la dinámica de cambio de las características cepstrales señaladas en 2.1.3. Una vez hallado NS, se procede a comparar la cantidad de segmentos fonéticos encontrados NS vs el número de fonemas NF de la transcripción fonética de la palabra en proceso de etiquetado. Los resultados NF = NS son los esperados. Si hay fonemas explosivo intermedio, éste está precedido por un ‘sil’, el cual no se señala en la transcripción fonética, pero el sistema se encarga de añadirlo. Lo mismo se realiza para el fonema ‘ch’.

2.1.5. Diseño de dos modelos de seguimiento de la dinámica de cambio de los vectores del VQ al tiempo de producirse la vocalización de la palabra. Determinación del Inicio – Fin de los segmentos fonéticos

Siguiendo la dinámica de cambio de los valores de los vectores VQ del libro código, se pueden encontrar los cambios de fronteras de los segmentos fonéticos en la palabra. Las fronteras son zonas de coarticulación fonética. Hay que tener presente, que también ocurren cambios de los valores de los vectores VQ producto del ‘ruido fonético’, por variaciones de las características MFCC en un mismo segmento fonético y debido a incorrecta pronunciación por parte de los locutores. Estas variaciones son generalmente de muy corta duración y se pueden ‘filtrar’ eliminar en buena parte. Al observar la dinámica de cambio de los valores, nos encontramos con palabras donde la diferencia entre las características fonéticas contiguas es notable y donde no lo es. Se utilizan dos modelos para encontrar las fronteras de cambio.

Tomando la información presentada en la Figura 2, se asigna a cada centroide heredero un número simbólico o índice que lo identifica: c1 es 1, c2 es 2, c3 es 3 y c4 es 4. Las señales de voz utilizadas, a manera de ejemplo para este trabajo, son: aji1.wav y mas2.wav.

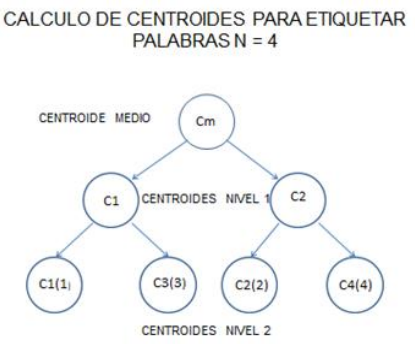

Fig. 2. Distribución de centroides del Libro Código (LC)

2.1.5.1. Modelos utilizados para el etiquetado automático

Se proponen y diseñan dos modelos para etiquetar segmentos fonéticos, sobre la base del análisis de la variación dinámica de los vectores símbolo de los centroides del LC de la VQ. La representación de las señales: aji1.wav y mas2.wav y sus espectrogramas, así como los resultados obtenidos para los dos Modelos propuestos, se observan en las figuras 4 y 7.

Descripción del Modelo 1

En este método se aprovecha la propiedad de formación de los centroides herederos c1 - c3 (1-3) y c2 - c4 (2-4) en el nivel 2 de la VQ, los cuales provienen de los centroides 1 y 2 del nivel 1.

Tomando el ejemplo aji1.wav,sil /a/ /j/ /i/ sil, NF=5; los resultados de la extracción de centroides y su distribución dinámica en el tiempo, figura 3, gráfica superior, son: 24(4), 1(2), 6(3), 1(2), 20(3), 56(1), 17(3), 64(2), 6(3), 38(1), 42(3), 28(2) y 13(4); segmentos con símbolos iguales (NS = 13). Total de centroides 316. El número antes del paréntesis es la cantidad de centroides iguales en cada segmento. El primer segmento tiene 24 centroides del símbolo 4, luego un segmento con un centroide de símbolo 2, luego 6 con 3, un 2, 20 de 3, 56 de 1 (hay combinaciones de: 4-2 y 3-1), etc.

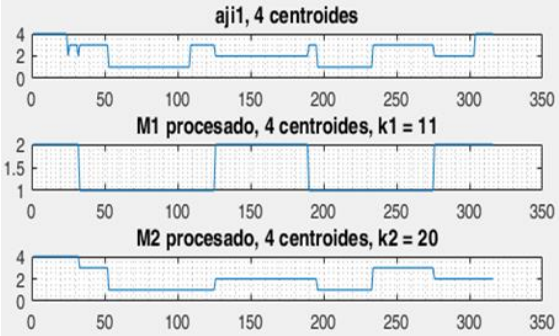

Fig. 3. Dinámica de centroides y agrupación Método 1 para aji1.wav

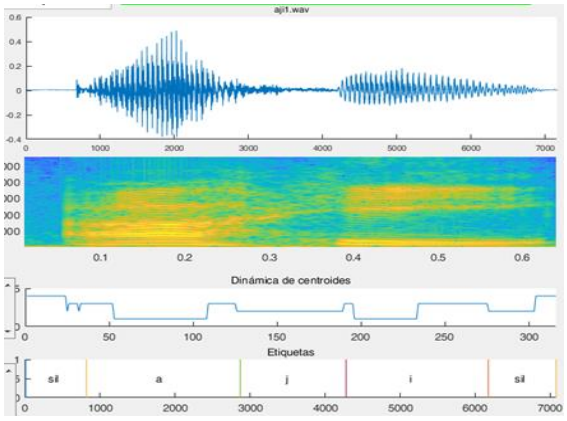

Fig. 4. Etiquetado automático de segmentos fonéticos, palabra aji1

Cada centroide es de un segmento de 20 mseg., de señal, traslapado 2 mseg. El algoritmo para M1 agrupa los segmentos con centroides herederos consecutivos 1-3 y 2-4. Agrupando y tomando como valor el símbolo resultante 1 para los centroides 1-3 y 2 para los centroides 2-4, por lo que la cadena resultante es: 25(2), 6(1), 1(2), 93(1), 64(2), 86(1) y 41(2); segmentos con símbolos iguales NS = 7.

Podemos observar que hay al menos cinco segmentos bien definidos con agrupación de segmentos, con una cantidad total mayor a 20 segmentos c/u, pero hay una región que tiene ruido fonético: 6(1), 1(2) y quedan sin agrupar.

Para los que quedan sin agrupar se analiza si constituyen ruido fonético en la secuencia. VER NOTA 1. Se considera que hay un nuevo segmento fonético, si la cantidad de símbolos en el segmento analizado es mayor o igual a once, k1=11, lo que representa un tiempo menor a 22 mseg., para una frecuencia de muestreo igual a 11025 Hz.

Si la cantidad de símbolos en un segmento es inferior a 11, entonces se consideran esos valores como ruido fonético y se sustituyen por valores idénticos al del segmento anterior, 6(1) pasa a 6(2), así podemos ver que al segmento 25(2) se le asignan las nuevas cantidades del segmento 6(1) y 1(2) respectivamente.

Se ejecuta agrupación nuevamente y se obtiene el resultado: 32(2), 93(1), 64(2), 86(3) y 41(2). Tenemos 5 segmentos fonéticos (NS = 5), en secuencia dinámica bien definidos. Ver Figura 3, gráfica intermedia. Luego NS = NF = 5, M1 da el resultado esperado. En la gráfica inferior de la figura 3 se muestra el resultado de la corrida del Modelo 2, el cual explicaremos a posteriori y que para la palabra procesada aji.wav no da buen resultado. Se corre el Modelo 2 y el resultado con Modelo 2 da NS = 7 segmentos fonéticos: 32, 20, 73, 70, 38, 42 y 41. Luego M2 no da solución. NS = 7 no es igual a NF = 5. Ver gráfica inferior de la figura 3.

Concluyendo, fonéticamente tenemos NS = 5 que coincide con NF = 5, lo que se corresponde con la transcripción fonética esperada: sil /a/ /j/ /i/ sil. Ver Figura 4. Ultima gráfica 'Etiquetas'. Se observa que las fronteras de los segmentos fonéticos se encuentran en la zona media de la coarticulación con una distancia media de 10 mseg. El algoritmo del Modelo 1 se presenta en la figura 5.

NOTA 1. En varios experimentos, se ve que la duración del ruido fonético es muy variable, dada la variabilidad que se presenta en la producción de diferentes palabras e incluso sobre una misma palabra articulada repetidamente por un mismo locutor. El umbral de k1 puede funcionar bien para: 6, 11 o 15 (12, 22 o 30 mseg.), según el caso. Siempre que el Modelo 1 no da solución correcta se puede cambiar el valor de k1 y el usuario decide que hacer (ej. Repetir en automático con nuevo k1 o No). Al recibirse un resultado de etiquetado

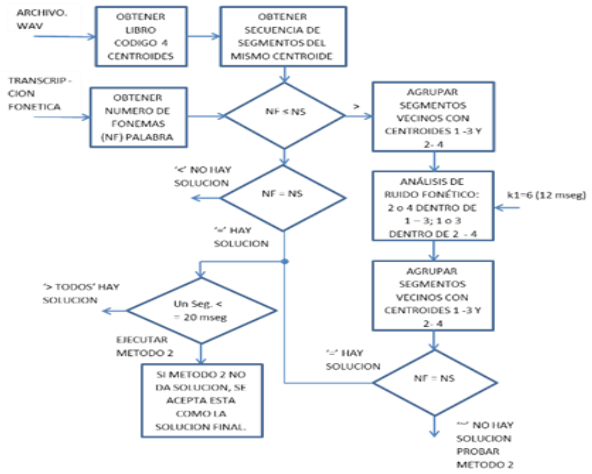

Fig. 5. Modelo 1. Diagrama del algoritmo

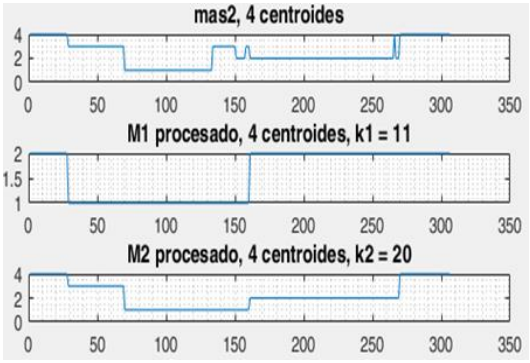

Fig. 6. Dinámica de centroides y agrupación Modelo 1 y Modelo 2 para mas2.wav

correcto donde NS = NF, las fronteras fonéticas pueden no ser las esperadas (lo cual se observa por el usuario de forma visual). El usuario puede cambiar el umbral de k1 y repetir el proceso de etiquetado automático o aplicar etiquetado manual. Por defecto K1 = 11.

Descripción del Modelo 2

Hay casos en que el Modelo 1 no resuelve la segmentación fonética (etiquetado fonético) debido a que existen fonemas que producen centroides herederos del mismo tipo o con el mismo símbolo: vocales coarticuladas, silencios coarticulados con silbantes, vocales coarticuladas con semivocales (ñ, l m) o consonantes sonoras (V, D, G), etc.

Tomando a manera de ejemplo el archivo de señal mas2.wav, se obtiene la secuencia de segmentos con centroides idénticos, ver figura 4 gráfica superior: 28(4), 41(3), 64(1), 17(3), 7(2), 3(3), 105(2), 1(4), 3(2) y 37(4); segmentos con centroides iguales (NS = 10). Total, de centroides 306. Donde la unión de 'm' y 'a' produce segmentos fonéticos con centroides hijos 1 – 3 y el Modelo 1 los une. Los resultados por el Modelo 1 para este ejemplo son: mas2.wav: 28, 122 y 156; tres segmentos fonéticos en NS = 3, gráfica central, lo cual no es solución del etiquetado fonético, según la transcripción fonética esperada NF = 5. Ejecutando el Modelo 2 para mas2.wav tenemos, se analiza solamente la cantidad de centroides que contienen cada uno de los segmentos. Este análisis se realiza por simple inspección, considerando que un segmento fonético lo es, si la cantidad de centroides que posee es mayor a k2=20 (40 mseg).

De ser menor a 21, se considera ruido fonético o concatenación y se añade y modifica su valor al del segmento anterior. En este caso queda: 28, 41, 64+17+7+3, 105+1+3 y 37. Se analiza si la cantidad de segmentos hallados NS coincide con la cantidad de fonemas de la transcripción fonética NF. De ser iguales se termina (HAY SOLUCIÓN), de ser menor no hay solución.El resultado final es: 28(4), 41(3), 91(1), 109(2) y 37(4); (Método 2). Figura 6. Tenemos 5 segmentos fonéticos (NS = 5), en secuencia dinámica definidos. Para Modelo 1 falla, hay tres segmentos. Concluyendo fonéticamente tenemos NS = 5 que coincide con la transcripción fonética esperada (NF = NS = 5): 'sil', 'm', 'a', 's', 'sil', ver Figura 7. M2 da el resultado esperado. Ultima gráfica 'Etiquetas'. Se observa que las fronteras de los segmentos fonéticos se encuentran en la zona media de la coarticulación con una distancia media de 10 mseg. El algoritmo del Modelo2 se presenta en la Figura 8.

NOTA 2. Para algunas palabras hay resultado de etiquetado CORRECTO, NS = NF, tanto por el Modelo 1 como por el Modelo 2. La solución final se toma bien eligiendo que modelo da mejor resultado: M1 o M2 o AMBOS, promediando la

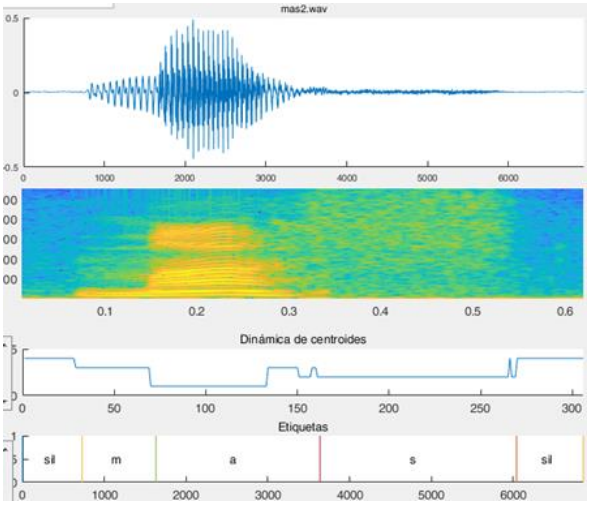

Fig. 7. Etiquetado automático de segmentos fonéticos, palabra 'mas'

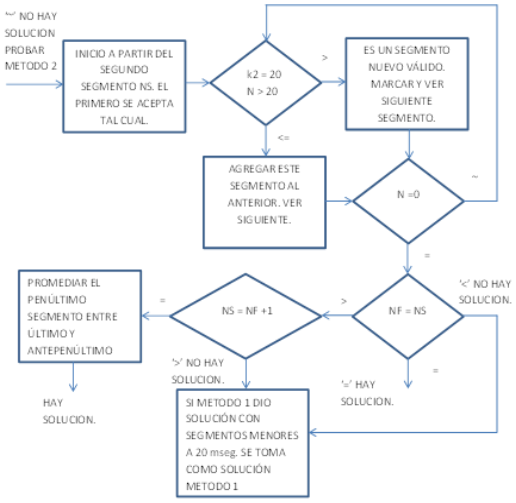

Fig. 8. Modelo 2. Diagrama del algoritmo

posición de las fronteras encontradas por ambos modelos.

El usuario escoge que resultado es el mejor. El usuario puede cambiar el umbral de k2 de 20 por defecto a 10, 15 o 25 y repetir el proceso de etiquetado automático o aplicar etiquetado manual. Por defecto K2 = 20.

Al obtenerse un resultado de etiquetado automático CORRECTO, donde NS = NF, se realiza la asignación de la transcripción fonética dada a los segmentos fonéticos NS encontrados, lo cual se presenta en las figuras 4, Modelo 1 y en figura 7, Modelo2; en la ventana inferior 'Etiquetas'.

2.1.6. Tratamiento para solucionar el etiquetado automático en palabras que contienen fonemas explosivos: b, d, g, k, p, t

Los fonemas explosivos van precedidos de un segmento de silencio sil, los cuales no son escritos en la transcripción fonética de las palabras que los contiene. Así a manera de ejemplo, las palabras: aba se representa por sil /a/ /b/ /a/ sil (NF=5) y papa por sil /p/ /a/ /p/ /a/ sil (NF=6), etc. Esto implica que el número de segmentos fonéticos NS finales correctos va a ser mayor a la cantidad de fonemas escritos en la transcripción fonética NF, al menos uno más por cada fonema explosivo que contenga la palabra.

Los resultados esperados en NS son: sil /a/ sil /b/ /a/ sil (NS=6) y sil /p/ /a/ sil /p/ /a/ sil (NS=7). Obsérvese que si el fonema explosivo está al inicio de la palabra, ya existe un 'sil' antes del mismo. Si hay varios fonemas explosivos al interior de la palabra, se incrementa el número NS a obtener en 1 por cada fonema explosivo.

Durante el proceso de selección y búsqueda de NS por los modelos M1 y M2, es muy probable que los segmentos fonéticos correspondientes a los fonemas explosivos sean considerados como ruido dada su corta duración y son absorbidos por el segmento de silencio que los precede.

Esto provoca que la solución NS = NF, encontrada como solución correcta para los modelos M1 y M2, entrega como segmentos fonéticos la unión de los fonemas explosivos con el silencio que les precede: 'sil + fonema explosivo'.

Una vez ejecutada la búsqueda del conjunto NS de la palabra a etiquetar por los modelos M1 y M2, se determina a partir de la transcripción fonética si hay fonemas explosivos en la palabra: cuantos y la posición de cada uno de ellos; para la ´b' de aba (sil /a/ /b/ /a/ sil) hay un fonema explosivo 'b' y ocupa la posición 3; para las 'p' de papa (sil /p/ /a/ /p/ /a/ sil), hay dos fonemas explosivos, uno al

inicio y otro dentro de la palabra; ocupan la posición: 2 y 4.

La metodología a aplicar en este caso tiene dos posibilidades: sin fonema explosivo al inicio o con fonema explosivo al inicio.

2.1.6.1. Sin fonema explosivo al inicio, fonemas explosivos al interior de la palabra

Luego de ejecutados los modelos M1 y M2, del resultado que aporte NF = NS, solución CORRECTA, se procede a encontrar la frontera final del segmento NS que contiene la unión 'sil-explosivo', para cada posición calculada; posición 3 para aba. Tomando esta posición como referencia, se toma el segmento centroide más corto está a la izquierda o derecha de esta frontera y se infiere que ese es el segmento que le corresponde al segmento explosivo. Se reacomoda la distribución de NS incorporando el nuevo segmento explosivo y ajustando la secuencia de los segmentos NS, el segmento explosivo se incorpora después del silencio que precede a la explosiva. Ya hemos incorporado un valor a la secuencia NS.

De haber más fonemas explosivos, el procedimiento es el mismo. Como ejemplo para la palabra aba1.wav., el Modelo 1 es el que obtiene la respuesta correcta NF = NS = 5, figura 9, segunda gráfica M1 procesado. El segmento NS(3) se corresponde con la unión 'sil-b', se busca cual es la frontera final de este NS(3) y se aplica el procedimiento señalado. El resultado se muestra en la gráfica 'Etiquetas' de la figura 10. El archivo de transcripción fonética correspondiente es aba1.lab.

2.1.6.2. Con fonema explosivo al inicio

Un fonema explosivo al inicio de palabra, durante la búsqueda de los segmentos NS, éste es absorbido por el silencio inicial de la transcripción fonética por considerarlo un ruido, lo cual implica que el número de segmentos NS final encontrados se vea afectado, debido a la unión 'sil-explosivo' inicial.

NF considera el 'sil' inicial y el fonema explosivo escrito. Para eliminar esta incongruencia, se cambia la transcripción fonética escrita por: ej., posa1 queda como /p/, /o/, /s/, /a/, sil (NF=5); sin

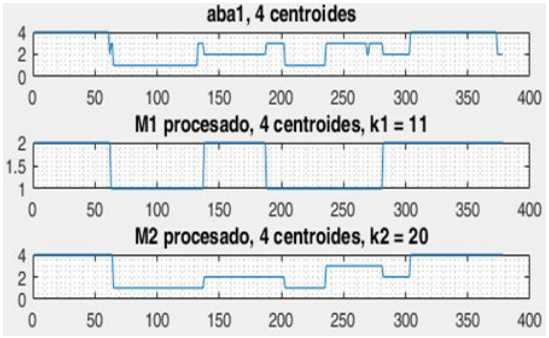

Fig. 9. Dinámica de centroides y agrupación Modelo 1 para aba1.wav

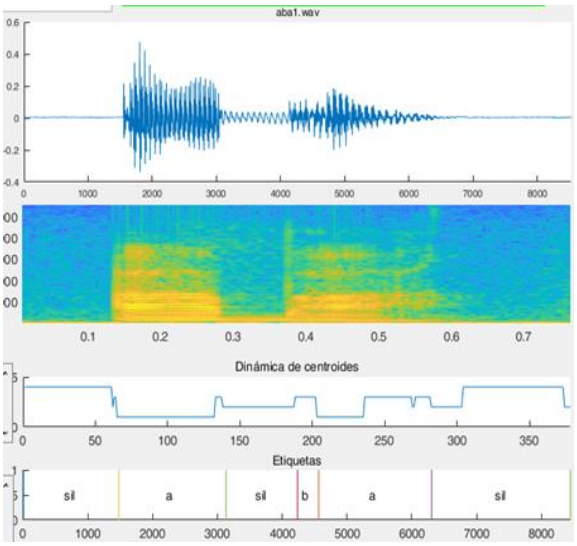

Fig. 10. Etiquetado automático de segmentos fonéticos, palabra 'aba1'

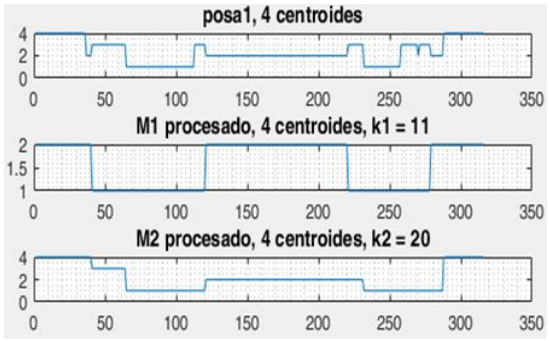

Fig. 11. Dinámica de centroides y agrupación Modelo 1 para posa1.wav

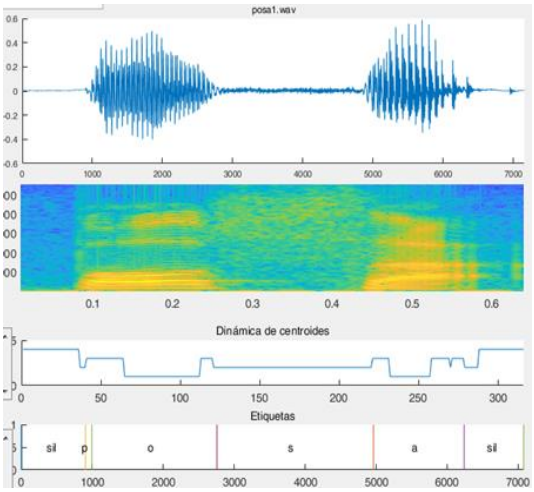

Fig. 12. Etiquetado automático de segmentos fonéticos, palabra 'posa1'

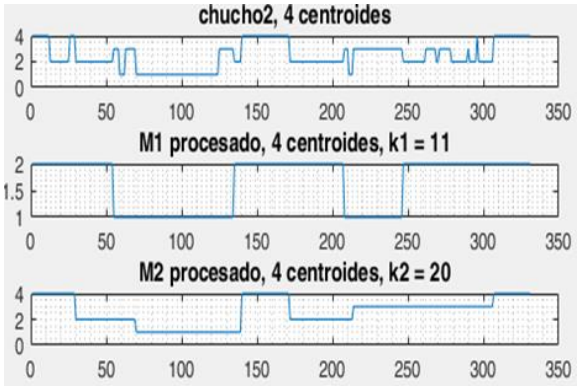

Fig. 13. Dinámica de centroides y agrupación Modelo 2 para chucho2.wav

el silencio inicial. De esta forma la cantidad de segmentos fonéticos NS a encontrar será uno menos (NS=5). Vemos que el resto queda igual. Ahora la cantidad de fonemas explosivos y su posición queda: un fonema explosivo en posición 1. Se ejecutan los modelos M1 y M2 y se procede como en el punto 2.1.5.1, para el caso NF = NS.

Al encontrar la frontera final del fonema explosivo /p/, en 1, ver figura 11 y 12, se reacomodan las fronteras de igual forma, tomado en cuenta el centroide corto más cercano a la frontera final en proceso. Se pone como restricción el máximo tamaño del segmento fonético explosivo al inicio es N=20.

NOTA 3. Para los casos con resultados NF = NS en ambos modelos: M1 y M2, se toma la decisión de búsqueda y ubicación de los segmentos fonéticos explosivos con los resultados de uno de los dos modelos, se prefiere el resultado del M2, ya que la promediación de fronteras entrega resultados incorrectos. Esto se debe a que el algoritmo trabaja sobre la selección de cambios de valores en la dinámica de los centroides iniciales. Un ejemplo de ello es en la palabra posa1.wav. Primera frontera después del 'sil' inicial.

2.1.7. El tratamiento para solucionar el etiquetado automático en palabras que contienen el fonema 'ch' (H)

El fonema 'ch' va precedido de un segmento de silencio 'sil'. Este fonema NO ES DE CORTA DURACIÓN a diferencia con los fonemas explosivos. En la transcripción fonética de una palabra que contiene este fonema, NO SE ESCRIBE EL FONEMA 'sil'. Para efectos de este trabajo al fonema 'ch' se le asignó la representación /H/. Así en las palabras: *chao* y *chucho*, la transcripción fonética se corresponde con: sil /H/ /a/ /o/ sil y sil /H/ /u/ /H/ /o/ sil, respectivamente. Observándose que al igual que para los fonemas explosivos, hay segmentos de silencio al inicio antes del fonema 'ch' y al interior de la palabra antes del mismo fonema. Metodológicamente lo que se hace es, teniendo en cuenta la transcripción fonética de la palabra, se añade a la misma el fonema 'sil' antes del fonema /H/, siempre y cuando el fonema /H/ no esté al inicio de la palabra.

Para los ejemplos mostrados, estos quedan: *chao* como sil /H/ /a/ /o/ sil y *chucho* como sil /H/ /u/ sil /H/ /o/ sil. El resto del procesamiento queda idéntico al procesamiento de palabras que no contienen fonemas explosivos. En la figura 13 y 14 vemos el resultado del etiquetado automático de la palabra chucho2. El resultado es con el Modelo 2.

3. Experimentos y resultados

Se realizaron pruebas de etiquetado automático con palabras que contienen hasta cinco fonemas diferentes y hasta un máximo de 8 fonemas a etiquetar, repitiendo fonemas al formar las palabras más largas. Se trabajó con un corpus comprendido por 105 palabras, que se describen a continuación: aba, ada, ado, agua, aja, aji, ajo, ala, alba, ale, ali, alu, alla, alli, ama, amo, ana, ano, api, aro, asa, ata, ato, ava, ave, avu, bata, boba, bota, boya, casa, casi, chao, cheo, chia, chio, choza, chucho, cola, cuna, dedo, del, dos, efe, ego, ele, eme, enano, fio, hacha, hago, hallo, hiena, higo, huele, hugo, humo, ira, ivan, ivo, lago, leo, lio, mama, mas, masa, masivo, mio, musa, nano, nene, nos, nueve, oro, osa, oso, papa, pasa, posa, sas, sazona, seis, ser, soda, sol, solo, sonso, sopa, sosa, sun, susana, suyo, tapa, tito, topo, tupa, ufo, uno, upa, usa, uva, van, viene, voy, voz, wan. No obstante se evitó usar palabras con más de cinco fonemas diferentes y en especial que estén coarticulados con fonemas que contienen descriptores característicos similares. Se usaron palabras que contienen fonemas repetidos.

Las palabras que están conformadas por fonemas ‘no sordos’ (explosivos de corta duración), así como las que contienen fonemas explosivos y el fonema ‘ch’; fueron etiquetados correctamente. Esto da un acierto de etiquetado correcto para 26 fonemas de 28 propuestos en la tabla2. Esto significa un ahorro de tiempo considerable en la tarea de etiquetado fonético.

Como característica principal del etiquetado automático, se señala que los dos modelos diseñados poseen propiedades que se complementan para dar los resultados esperados del etiquetado automático. El Modelo 1 es bueno para encontrar resultados correctos cuando los segmentos fonéticos concatenados no son semejantes en sus características: vocal-silbantes; además en presencia de ruido fonético. El Modelo 2 es bueno para encontrar resultados correctos cuando los segmentos fonéticos concatenados son parecidos en sus características: vocal-vocal, vocal-semivocales, vocal-consonantes sonoras. Ambos modelos resuelven las fronteras del etiquetado automático con una precisión de unos 10 mseg.

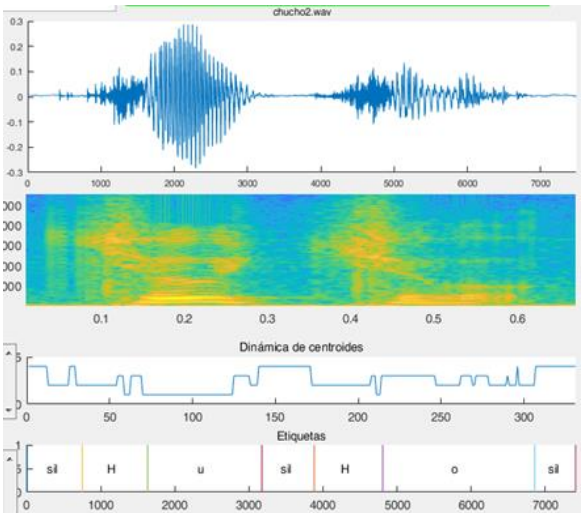

Fig. 14. Etiquetado automático de la palabra chucho2.wav

Tabla 2. Resultados obtenidos

Etiquetado/porcentaje de reconocimiento	**%**
Manual	98.96
Automático	97.9

Los fonemas explosivos: /b/, /d/, /g/, /k/, /p/ y /t/ tienden a identificarse como ruido fonético debido a su corta duración y estar precedidos de un silencio, lo que da lugar a que el fonema explosivo se una al silencio; por lo que se diseñó un procedimiento de detección automática de fronteras que involucra la dinámica de cambio de los vectores símbolos y la información de la transcripción fonética de la palabra que se procesa y que contiene los fonemas explosivos. Para el fonema ‘ch’ se añade a la transcripción fonética el silencio ‘sil’ que le antecede en su pronunciación. Para los fonemas vibrantes: /r(/ y /r/, hay dificultad en su etiquetado, no siempre se obtienen buenos resultados, son fonemas donde sus características varían fuertemente durante su producción. Se debe de realizar el etiquetado de los mismos usando la opción Manual.

La aplicación desarrollada, al realizar el etiquetado automático, da la posibilidad de re-

etiquetar manualmente las fronteras fonéticas resultantes del etiquetado automático, a criterio del usuario, posibilitando corregir por el mismo alguna frontera que a su perecer no está bien ubicada. Si el sistema no encuentra resultado plausible, NS = NF, lo informa como 'No Result' (NO CORRECTO). Esto ocurre frecuentemente con palabras que contienen fonemas explosivos o vibrantes.

El resultado NO CORRECTO, del etiquetado automático tiene que ser solucionado con la opción de Etiquetado Manual. Con el resultado del etiquetado automático del corpus de palabras utilizado, se creó el corpus de los modelos de cada uno de los fonemas, HMM (*Hidded Markov Model*) de fonemas y se crearon, por concatenación fonética, los modelos de las palabras [13].

3.1. Etiquetado promedio automático

Se añadió la opción de realizar el etiquetado de un conjunto de palabras de forma completamente automática. Esta opción realiza para cada palabra a etiquetar, las 12 pruebas de las combinaciones de los parámetros k1 y k2. Para todas las respuestas que ofrecen solución de etiquetado encontrada, se realiza el cálculo del promedio de las fronteras obtenidas y se da como resultado el etiquetado final de la palabra procesada.

4. Validación de la efectividad de reconocimiento de palabras usando los resultados del etiquetado automático vs el etiquetado manual

Se utilizó el software desarrollado por la universidad de Cambridge HTK Versión 3.4 [14-15], específicamente diseñado para actividades de reconocimiento de voz, para procesar los resultados del etiquetado del corpus de voz diseñado para la prueba de etiquetado automático vs etiquetado manual.

Este corpus de voz en idioma español consta de 106 palabras diferentes y contienen 23 fonemas. Se omite el fonema /r/, por su característica ruidosa de larga duración. Se trabajó con un usuario. Los resultados obtenidos se presentan en la tabla 2.

4.1 Metodología de creación del corpus de palabras para obtener el etiquetado automático de fonemas

El corpus de palabras que se utiliza para aplicar esta técnica de etiquetado automático usando la dinámica de cambio de las características descripticas de los segmentos fonéticos, requiere se diseñe el mismo considerando que:

- Usar palabras de corta duración. Máximo cinco fonemas diferentes.
- Las palabras deben estar precedidas y terminar con segmentos de silencio.
- Evitar en lo posible coarticulación de fonemas con descriptores fonéticos similares.
- El etiquetado de fonemas vibrantes, debe de ser realizado con la opción Manual.

5. Conclusiones

El objetivo principal de esta investigación es encontrar las fronteras fonéticas mediante la dinámica de cambio de las características de la señal en el tiempo, que como se postula, es el proceso que ocurre a nivel de la percepción del contenido fonético de las palabras en nuestro sistema nervioso central y que es independiente del idioma.

El resultado de la determinación de las fronteras fonéticas para el etiquetado automático, usando la técnica de análisis de la dinámica de cambio de las características de los fonemas es posible y representa una nueva alternativa para obtener las fronteras de los segmentos fonéticos concatenados para formar palabras. Los dos modelos desarrollados se complementan en el trabajo para determinar correctamente las fronteras fonéticas.

La opción de etiquetado manual está prevista, para aquellos casos en que la pronunciación de la palabra y el resultado de la determinación de los segmentos fonéticos encontrados para ella, no coincidan con la transcripción fonética asignada a la misma.

El porciento de reconocimiento y etiquetado correcto obtenido con esta aplicación es del 97.9% inferior en un 1.06%, con respecto al porcentaje de

reconocimiento obtenido sobre el mismo corpus de palabras, pero haciendo uso de un etiquetado manual superior al 90%. Sin embargo, el tiempo utilizado en el etiquetado del corpus de voz de forma automática, es significativamente menor que el estimado de hacerse manualmente, además de eliminar la subjetividad personal en el trabajo de etiquetado.

Agradecimientos

Se agradece al Instituto Politécnico Nacional y en especial a la Secretaría de Investigación y Posgrado, por su financiamiento para el desarrollo del presente trabajo, bajo el auspicio del proyecto SIP 20181550 y SIP 20195296.

Referencias

1. **UPV (2010).** *Desarrollo de un sistema de reconocimiento automático del habla*. Universidad Politécnica de Valencia, Escuela Superior de Ingeniería Informática. pp. 8–9.
2. **UPC (1993).** *Técnicas de procesado y representación de la señal de voz para el reconocimiento del habla en ambientes ruidos*. Universidad politécnica de Cataluña, Departamento. de teoría de la señal y comunicaciones, Capítulo 5, Sección 5.5.
3. **Galka, J. & Ziolko, M. (2008).** Wavelets in Speech Segmentation. *Electrotechnical Conference, MELECON´08, The 14th IEEE Mediterranean,* pp. No. 5-7, pp. 876–879.
4. **Ziolko, B., Manandhar, S., & Wilson, R.C. (2006).** Phoneme Segmentation of Speech. Pattern Recognition. *Conference on ICPR´06, 18th International,* Vol. 40, No. 307, pp. 282–285. DOI: 10.1109/ICPR.2006.931.
5. **Hosom, J.P. (2000).** *Automatic Time Alignment of Phonemes using Acoustic-Phonetic Information.* Institute of Science and Technology.
6. **Bansal, P., Pradhanet, A., & Arora, M. (2014).** Speech Synthesis – Automatic Segmentation. *International Journal of Computer Applications,* Vol. 98, No. 4, pp. 29–31. DOI:10.5120/17172-7253.
7. **Toledano, D.T., Gómez, L.A.H., & Grande, L.V. (2003).** Automatic Phonetic Segmentation. *Speech and Audio Processing, IEEE Transactions on,* Vol. 11, No.6, pp. 617–625. DOI: 10.1109/TSA.2003. 813579 (2003).
8. **Davis, S. & Mermelstein, P. (1980).** Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 28, No. 4, pp. 357–366, DOI: 10.1109/TASSP.1980.1163420.
9. **Hernández-Mena, C.D. & Herrera-Camacho, A. (2017).** CIEMPIESS: *A New Open-Sourced Mexican Spanish Radio Corpus*. Departamento de Procesamiento Digital de Señales. Universidad Nacional Autónoma de México (UNAM).
10. **LNCS (2015).** Automatic Phoneme Border Detection improves Speech Recognition. Springer. MICAI 2015. LNAI 9413 Part I, pp. 127–138. DOI: 10.1007/978-3-319-27060-9.
11. **Quilis, A. (1999).** *Tratado de Fonología y Fonética Españolas*. Gredos.
12. **Linde, Y., Buzo, A., & Gray, R. (1980).** An Algorithm for Vector Quantizer Design. *IEEE Transactions on Communications*, Vol. 28, No. 1, pp. 84–95. DOI: 10.1109/TCOM.1980.1094577.
13. **Rabiner, L. & Biing-Hwang, J. (1993).** *Fundamentals of Speech Recognition.* Prentice Hall.
14. **Young, S., et al. (2006).** *The HTK Book*. Cambridge
15. **Young, S., et al. (2006).** *The HTK Toolkit*, http://htk.eng.cam.ac.uk/ Cambridge.

Article received on 29/07/2019; accepted on 06/01/2020. Corresponding author is Sergio Suárez-Guerra.