GUÍA PARA LA SEGMENTACIÓN Y TRANSCRIPCIÓN FONÉTICA PARA LAS TECNOLOGÍAS DEL HABLA

Ing. Jorge A. Gurlekian (1) Dr. Humberto Torres (2) Dr. Diego Evin (3)

(1) Ingeniero Electrónico. Director del Laboratorio de Investigaciones Sensoriales. INIGEM CONICET-UBA. Hospital de Clínicas Av. Córdoba 2351. 9 Piso Sala 2. anagraf99@hotmail.com

(2) Bioingeniero, Dr. En ingeniería. Investigador Científico del Laboratorio de Investigaciones Sensoriales. INIGEM CONICET-UBA. Hospital de Clínicas Av. Córdoba 2351. 9 Piso Sala 2. hmtorres@hotmail.com

(3) Bioingeniero, Dr. En Ciencias de la Computación. Investigador Científico del Laboratorio de Investigaciones Sensoriales. INIGEM CONICET-UBA. Hospital de Clínicas Av. Córdoba 2351. 9 Piso Sala 2. diegoevin@gmail.com

RESUMEN

En este trabajo se define una guía para la segmentación fonética y su transcripción mediante el alfabeto SAMPA (Speech Assessment Methods: Phonetic Alphabet). La transcripción fonética con SAMPA es de uso creciente en las tecnologías de habla en las tareas de preparación de datos acústicos para ser utilizados en el entrenamiento de sistemas de reconocimiento de habla/hablante y en la preparación de unidades de concatenación en sistema de conversión texto a habla. Se presentan ejemplos de cada realización fonética, los criterios de segmentación y su etiquetado.

Palabras claves*: habla, transcripción, segmentación acústica, tecnologías del habla*

A GUIDE FOR PHONETIC SEGMENTATION AND LABELING IN SPEECH TECHNOLOGIES

ABSTRACT

A normalized guide is proposed here, for both phonetic segmentation and labeling using SAMPA, (Speech Assessment Methods: Phonetic Alphabet) based on acoustic productions of each speaker. SAMPA labeling is a standardized technique for speech pre-processing both for automatic speech recognition systems and text to speech concatenation systems. Speech samples with segmentation and, labeling are shown.

Keywords*: speech, labeling, acoustic segmentation, speech technologies.*

INTRODUCCIÓN

La segmentación acústica consiste en indicar los instantes de inicio y fin de los distintos eventos acústicos, presentes en un archivo de audio. Dicha información de carácter temporal resulta de interés en las disciplinas que deben asociar segmentos acústicos de distinta longitud, con etiquetas que representan unidades lingüísticas y no lingüísticas de duración segmental y supra-segmental de una manera confiable. Entre esas disciplinas se pueden mencionar la conversión de texto a voz (CTV) [1,2], el reconocimiento automático del habla (RAH) [3], la identificación automática de hablantes (IAH) [4] y la detección automática de alteraciones vocales como el índice de precisión articulatoria (IPA) y el índice de perturbación integrado (IPI) entre otros [5].

Para las aplicaciones tecnológicas desarrolladas en esas áreas, en general suele ser indispensable la creación de corpus de datos con algún tipo de segmentación acústica, lo que hace que esta tarea sea muy requerida en la actualidad. A continuación se comentan algunos detalles específicos de las segmentaciones acústicas requeridas en las disciplinas mencionadas.

Los sistemas de CTV de alta calidad basados en concatenación de unidades presupone el etiquetado manual de las unidades de habla -típicamente difonos- que son segmentos de habla que van desde un instante de mayor regularidad o estabilidad en un fono hasta otro instante semejante en el próximo fono. Aún cuando la tarea de segmentación se inicia en forma automática, la mayor precisión se alcanza con la corrección manual por parte del experto en fonética. En estos sistemas, la alta naturalidad se obtiene además con la descripción fonética de la entonación, el ritmo y el acento.

En la conversión de sonido a texto, las técnicas de reconocimiento automático utilizan algoritmos de aprendizaje o entrenamiento que pueden requieren una segmentación inicial de unidades. En este caso la segmentación se utiliza para aprender los rasgos que distinguen a los sonidos del lenguaje, los cuales constituyen los modelos acústicos. Estos sistemas de RAH además cuentan con conocimiento lingüístico a priori, que viene dado mediante diccionarios de pronunciaciones que indican cómo se pronuncian las palabras, y modelos de lenguaje que señalan las probabilidades de observar distintas secuencias de tales palabras.

Finalmente, los sistemas híbridos de IAH para aplicaciones forenses han mejorado su desempeño al combinar la segmentación fonética y grafémica con los métodos probabilísticos que utilizan representaciones espectrales de las unidades para obtener medidas de verosimilitud en forma automática.

Si bien el etiquetado acústico con segmentación manual y realizado por expertos es el estándar de oro en aplicaciones de tecnologías de habla de alta calidad, los elevados costos y tiempos requeridos, así como la necesidad de contar con personas calificadas para llevarlo a cabo, muchas veces hacen que este procedimiento resulte muy costoso. Una solución alternativa y ampliamente difundida en la actualidad consiste en combinar la aplicación de métodos automáticos de segmentación y su corrección por fonoaudiólogos con entrenamiento musical, para iniciar o verificar a posteriori la transcripción automática.

La segmentación acústica automática de calidad impone dos precondiciones. En primer lugar, contar con la transcripción léxica a nivel de frases. Esta transcripción contiene la secuencia ordenada de palabras que se pronuncian en una emisión determinada. No es necesario que éstas incluyan las anotaciones de inicio y fin de cada palabra. En segundo lugar se necesita un sistema de RAH ya sea: entrenado con material equivalente al que se necesita segmentar, adaptado con parte del nuevo material, o bien, entrenado en su totalidad con el mismo material a segmentar. Este sistema de RAH puede ser empleado para estimar cuáles son los momentos más probables de inicio y fin de palabras, trifonos y fonos. A este proceso se lo conoce como alineamiento forzado.

Las transcripciones léxicas de las frases se deben transcribir en forma manual o, como una alternativa, en forma automática con un sistema de RAH. En este último caso, las segmentaciones presentan un mayor nivel de error, ya que se suman las incertidumbres y errores inherentes a todo sistema de RAH. Esto implica un mayor costo en la revisión final de las segmentaciones obtenidas.

En cualquier caso, la segmentación automática tiene limitaciones intrínsecas dado el modo de funcionamiento del sistema de RAH. Los instantes de segmentación están condicionados por las ventanas de observación del reconocedor. Para un reconocedor fonético estándar por ejemplo se suele obtener una salida de reconocimiento a una tasa fija de 10 ms, discretizando los instantes de detección de transiciones entre segmentos. En casos donde es vital contar con un entorno controlado de los instantes de inicio y fin de cada unidad, como es el caso ya mencionado de CTV por métodos concatenativos, estas limitaciones generan la necesidad de una revisión y ajuste fino de las segmentaciones automáticas.

Los sistemas de etiquetado automático requieren de reglas de transcripción grafémico- fonética que se aplican de acuerdo al contexto en que aparece cada fonema. La tarea del experto consiste en corregir las marcas temporales de segmentación y decidir si la variante fonética que propone la regla de la convención es la que se observa. Por lo tanto estas reglas son el marco de referencia inicial para el etiquetador automático. La tarea del etiquetador manual es priorizar lo percibido en la emisión acústica. Las modificaciones a las reglas se reflejan luego mediante una estadística que permite adaptar las reglas para ese locutor. Con respecto a la marcación temporal existe un número de convenciones que serán presentadas para cada grupo de sonidos. Los casos que presentan ambigüedad deben consensuarse por los expertos que determinan la marca final.

El presente trabajo surge como resultado del entrenamiento realizado a etiquetadores durante varios proyectos de desarrollo de corpus de habla, y compendia los apuntes y recomendaciones

ofrecidas a profesionales de fonoaudiología y locución para fijar el consenso en la tarea de segmentación y corrección.

La transcripción fonética

Recientemente se ha definido el alfabeto fonético SAMPA (Speech Assesment Methods: Phonetic Alphabet) [6] para la Argentina [7] fundamentado en las realizaciones fonéticas del español hablado en esta vasta región dialectal. Este alfabeto se ha definido para varias lenguas y es de aplicación en las tecnologías de habla.

Las ventajas principales de este alfabeto son: 1. Permitir el diseño de córporas con cobertura fonética completa, 2. Permitir el uso de un conjunto de caracteres estándar que se mapea con todos los símbolos del alfabeto fonético internacional (international phonetic alphabet: IPA). El mapeo asigna las diferentes variantes fonéticas regionales a un mismo símbolo SAMPA. De este modo los córpora grabados mediante textos SAMPA pueden garantizar la riqueza fonética de cada región.

En el trabajo mencionado [7] se han presentado los 22 fonemas del español de Argentina y Latinoamérica (17 consonantes y 5 vocales). Los dos fonemas que no se utilizan en Latinoamérica y nos diferencian del español peninsular son /z/ y /v/. Se han indicado además los ocho alófonos más frecuentes utilizados en la Argentina: [h, B, D, G, N, C, w, j) identificándose un total de 30 fonos. Ver síntesis en Tabla I y denominaciones en Tabla II.

Denominación de los sonidos de habla

Las realizaciones acústico-fonéticas utilizan términos extraídos de los distintos campos de la fonética Tabla I.

Representación acústica de los sonidos de habla

La manifestación acústica es la transducción eléctrica de la variación de presión en función del tiempo –forma de onda- y la del espectrograma de banda ancha (alta resolución temporal). Las ondas de vocales y otros sonidos con vibración glótica se presentan como ciclos de presión cuasi-periódicos en el tiempo que se distinguen claramente de los sonidos ruidosos y de los sonidos que contienen silencio más ruido. En la Figura 1 se distinguen claramente los sonidos ruidosos, periódicos y oclusivos. Sin embargo cuando se emiten dos vocales seguidas o en combinación con consonantes con vibración cuasi-periódica, la diferenciación se logra con el espectrograma. La coarticulación se manifiesta mediante las transiciones de formantes. Para ver esquemas de las transiciones VCV del español ver [4].

Tabla 1. Relaciones fonémicas con la fonética SAMPA y sus grafemas para el español de la Argentina.

Vocales

/a, o, e/ se transcriben a SAMPA con sus símbolos fonéticos [a, o, e]

/u, i/ se transcriben como [u, i,]. Si forman diptongo dentro o fuera de la palabra se transcriben como [w] y [j]. Grafemas a, e, o, u, i, y.

Consonantes

Ruidos

/s/ se transcribe como [s]. Cuando precede a una consonante puede transcribirse como [h]. Si el fonema es omitido se indica como [0]. Grafemas: s/c/z.

/x/ se transcribe como [x] antes de [a, o, u, w] y como [C] antes de [e, i/j]. Grafemas j, g.

/f / se transcribe como [f]. Grafema: f

/ʒ, dʒ, ʃ / se transcriben como [Z]. Grafemas: ll, y

Silencio + Ruido

/p,t,k/ se transcriben como [p, t, k]. Grafemas p, t, k/c/q.

/t ʃ, / se transcribe como [H]. Grafema: ch.

Nasales

/n/ se transcribe como [n]. Si precede a un sonido velar o palatal, se transcribe como [N].

o labial o bilabial, puede transcribirse como [m]. Grafema n

/m, ɲ/ se transcriben como [m, J]. Grafemas m, ñ.

Interrumpidos

/ɾ, r/ se transcriben como [r, R]. Grafemas: r y r/rr.

Continuos

/b, d, g/ se transcriben como [B,D,G] salvo que sigan a una nasal o sean inicio de frase después de pausa, en cuyo caso se transcriben [b, d, g]. Grafemas b/v, d, g.

/l/ se transcribe igual.

Tabla 2. Denominaciones derivadas de la fonética articulatoria, acústica y perceptual de los sonidos de habla. 1La clasificación perceptual es una adaptación de Guirao [8].

SAMPA	Articulatoria	Acústica	Perceptual[1]
p, t ,k (3)	Oclusivos sordos	Silencio + ruido corto, medio y largo	Pulsos fuertes no continuos compactos agudos
b, d, g (3)	Oclusivos sonoros	Un primer formante de baja frecuencia	Pulsos suaves no continuos difusos graves
B, D, G (3)	Aproximantes (no fricativos) [4]	Formantes transicionales de baja amplitud	(no evaluados) Transicionales periódicos
s, f, x h, C (5)	Fricativos sordos	Bandas de ruido (h,f) en todo el espectro	Ruidos en bandas Continuos
H (1)	Africada sorda	Silencio + ruido	Pulso de Ruido Fuerte
r (1)	Liquida vibrante simple	Una interrupción entre dos vocales	Periódico+ruido Cortos Suave No continuo
R (1)	Liquida vibrante múltiple	Múltiples interrupciones	Periódico+ruido largos continuos
l (1)	Liquida Lateral Pre-palatal lateral	Formantes de amplitud media	Periódico Continuos de baja amplitud
Z (1)	Fricativa sonora Palatal y pre-palatal Africada sonora	Periodicidad + banda de ruido de alta frecuencia.	Periódico+ruido Largo Agudo
i, e, a, o, u (5)	Vocales	Periódicos de amplitud alta con formantes de alta int.	Periódicos Continuos de alta amplitud
w, j (2)	Semivocales	Formantes transicionales de amplitud media	(no evaluados) Idem vocales
m, n, N, J (4)	Nasales	Formantes fijos de baja frecuencia y amplitud	Periódicos Continuos de baja amplitud

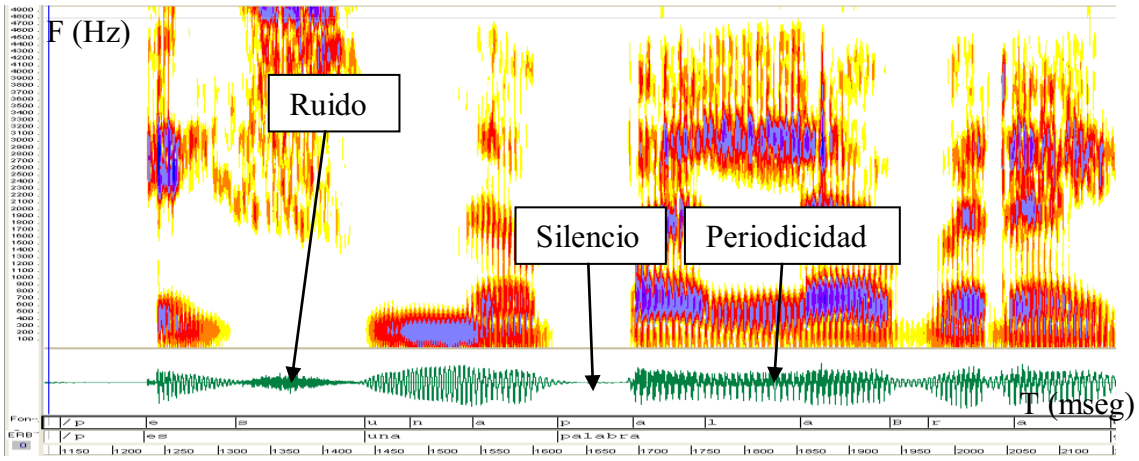


Figura 1. En la forma de onda se distinguen los sonidos periódicos [e,u,a,l] de los ruidos [s] y silencios de un sonido pulsante [p]. En el espectrograma se distinguen la [u] y la [n] de una [a]. También se distingue la [l] de las dos [a] adyacentes.

Sonidos Vocálicos

Los fonemas vocálicos /i,e,a,o,u/ se corresponden con los fonos [i,e,a,o,u] Figura 2 y los alófonos [w, j] Figura 3. A medida que el habla es más descuidada o tiene mayor velocidad los valores prototípicos de los formantes de las vocales aisladas se modifican al punto de desaparecer quedando como indicación de la vocal una representación acústica aproximada o gruesa que permite su identificación.

Como rasgos de rescate vocálico se observa para la vocal /a/ un 1er formante más alto que el resto de los formantes de todas las vocales presentes, con un 2do formante lo más próximo a este 1er formante. Para la vocal /i/ los rasgos distintivos gruesos son un 1er formante más bajo que el resto de los formantes y un 2do formante lo mas separado posible de este 1ro. La vocal /u/ tendrá el 1er formante más bajo (como la /i/) y un 2do formante lo más cercano posible. Con este nivel de descripción podrán diferenciarse las secuencias continuas de vocales. Las vocales /e/ y /o/ se diferenciaran de la /i/ y la /u/ al considerar su configuración de formantes como más cercano al patrón de una /a/.

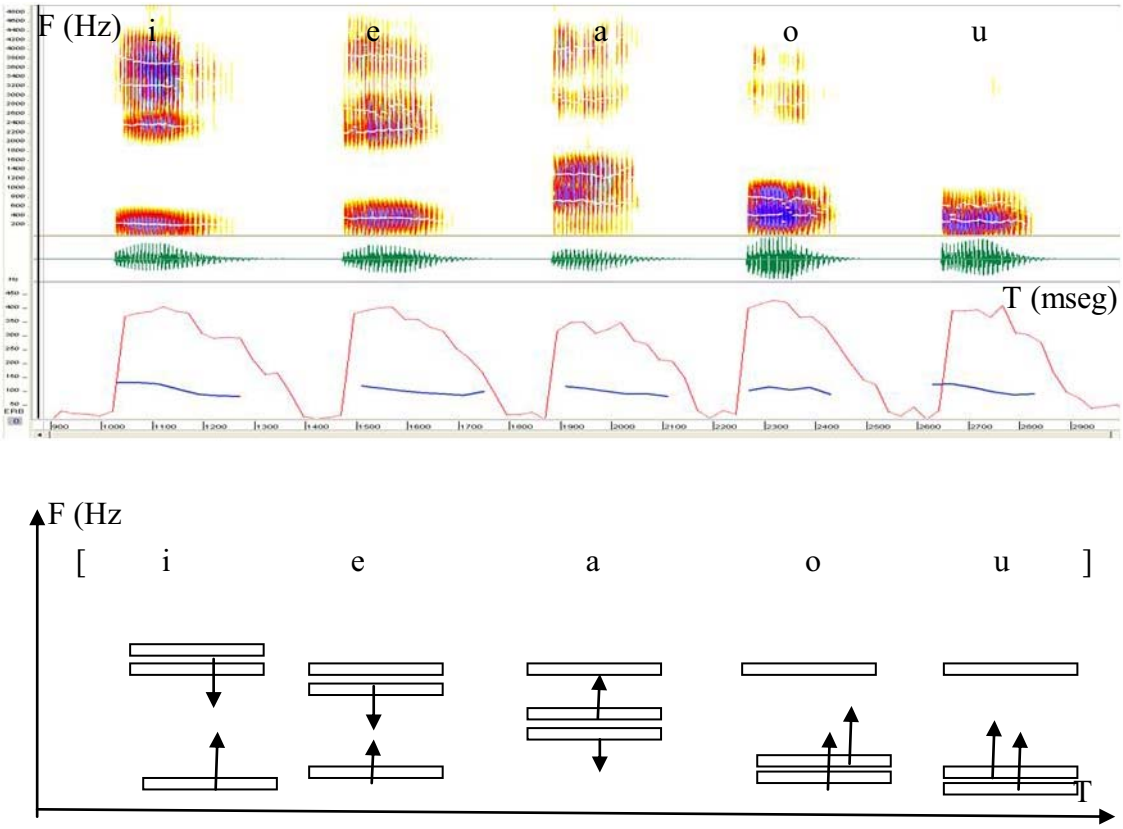


Figura 2. Representación de las vocales aisladas. Espectrograma, forma de onda y contornos de Energía Total y Frecuencia Fundamental (arriba) emitidas por un locutor masculino y de un esquema idealizado de las mismas vocales. (abajo).

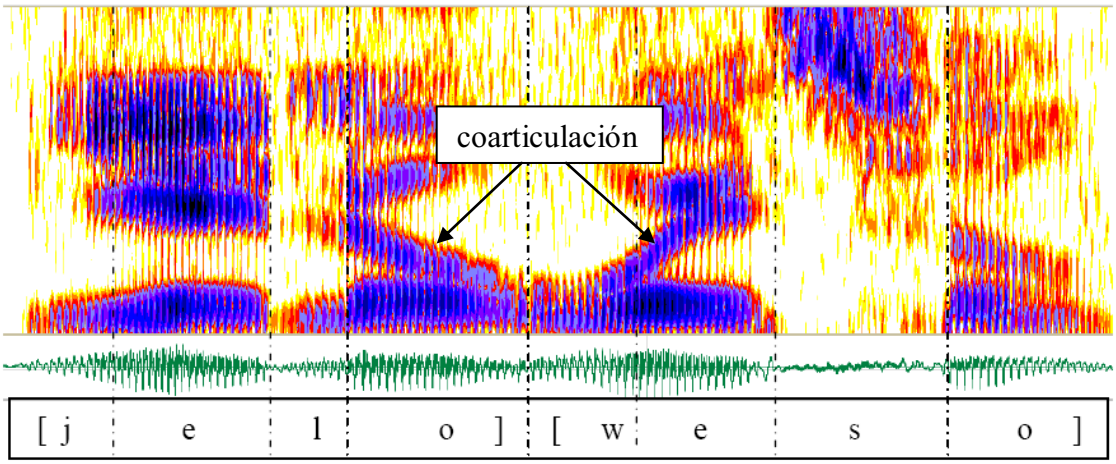


Figura 3. Representación espectrográfica de [j,w] en hielo, hueso. Obsérvese la menor duración y atenuación de estos sonidos.

Segmentación

En la segmentación en secuencias VV, el criterio es dividir la transición vocálica en dos partes iguales y asignar cada parte a un fono. Ver figura 4. En la segmentación de secuencias CV o VC el criterio es asignar toda la transición vocálica a la vocal. Si la vibración glótica de baja amplitud de la vocal continúa breve ó totalmente en el segmento que le corresponde al silencio o en el segmento correspondiente de ruido por falta de oclusión completa de la vocal previa, se asignará igualmente ese segmento a la consonante.

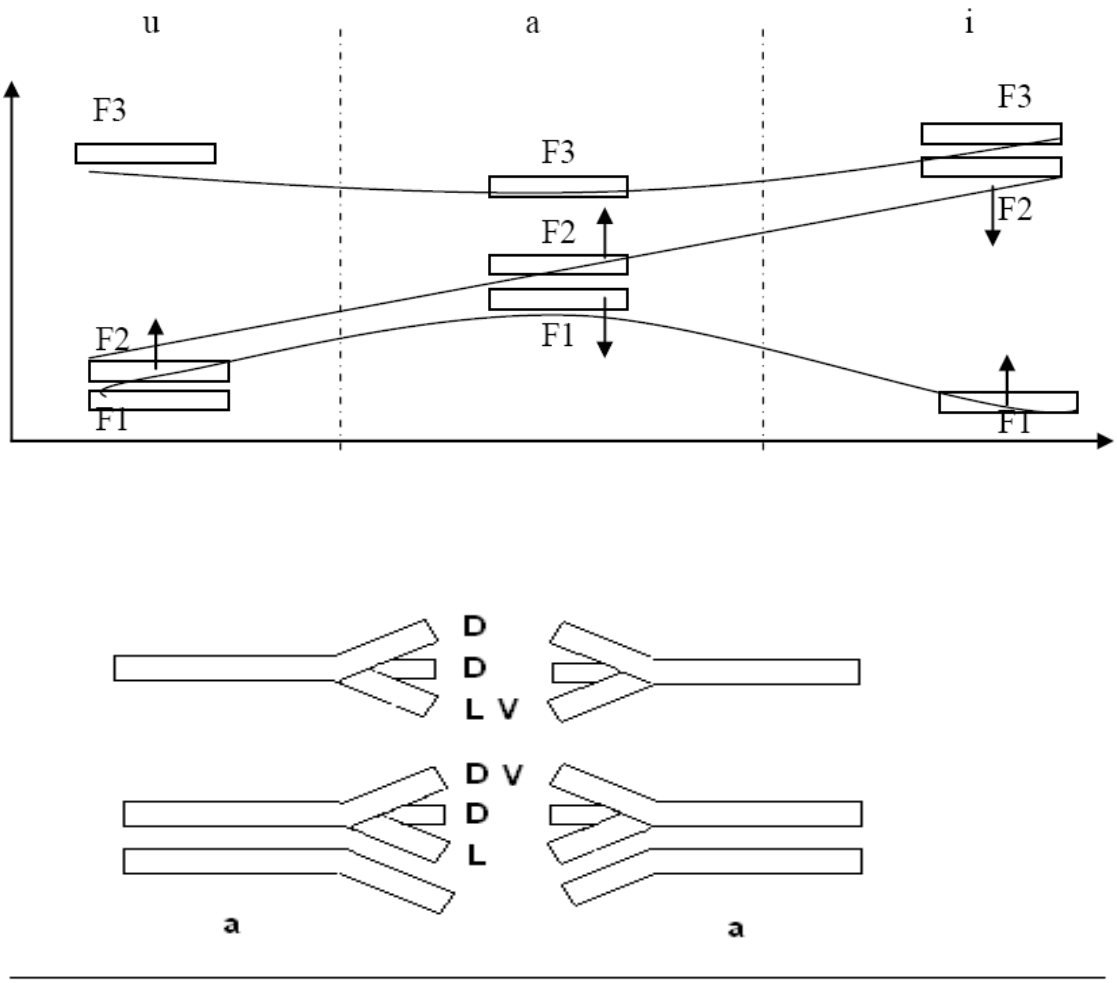


Figura 4. Esquema de coarticulaciones. Arriba: Vocálicas, se indica la trayectoria de los tres primeros formantes. Los formantes 1 y 2 de la vocal [u] tienden a elevarse, en la [a] a separarse y en la [i] a acercarse. Las transiciones se dividen por la mitad y se asignan a cada vocal. Abajo: Consonánticas, se indican las trayectorias para tres puntos de articulación amplios, L: labial, D: dental y V: velar, en contexto con la vocal /a/.

Sonidos [p, t, k, H]

Los fonemas /p,t,k,t ʃ/ se corresponden con los sonidos con silencio y ruido tales como los fonos oclusivos sordos [p,t,k] y la africada [H]. Estos presentan un segmento de silencio típicamente de 30 a 60 mseg, resultado de la oclusión al paso del aire para finalizar con un ruido de explosión de 1-5 mseg para [p] de 5 a 15 mseg para [t], de 15 a 30 mseg para [k] y de 30 a 60 mseg para [H]. La terminación del ruido de explosión es el instante que debe ser marcada como límite final de estas consonantes. Ver Figura 5. Como se mencionó en el inicio el segmento de silencio puede presentar ciclos amortiguados en amplitud del sonido sonoro anterior que son visibles por no existir una oclusión completa. Cuando estos sonidos están al inicio de una oración o de una frase melódica delimitada por silencio debe marcarse el inicio dejando aproximadamente 60 mseg de silencio previo al ruido de explosión. El sonido [H] se considera como la unión de dos sonidos: la [t] y [s].

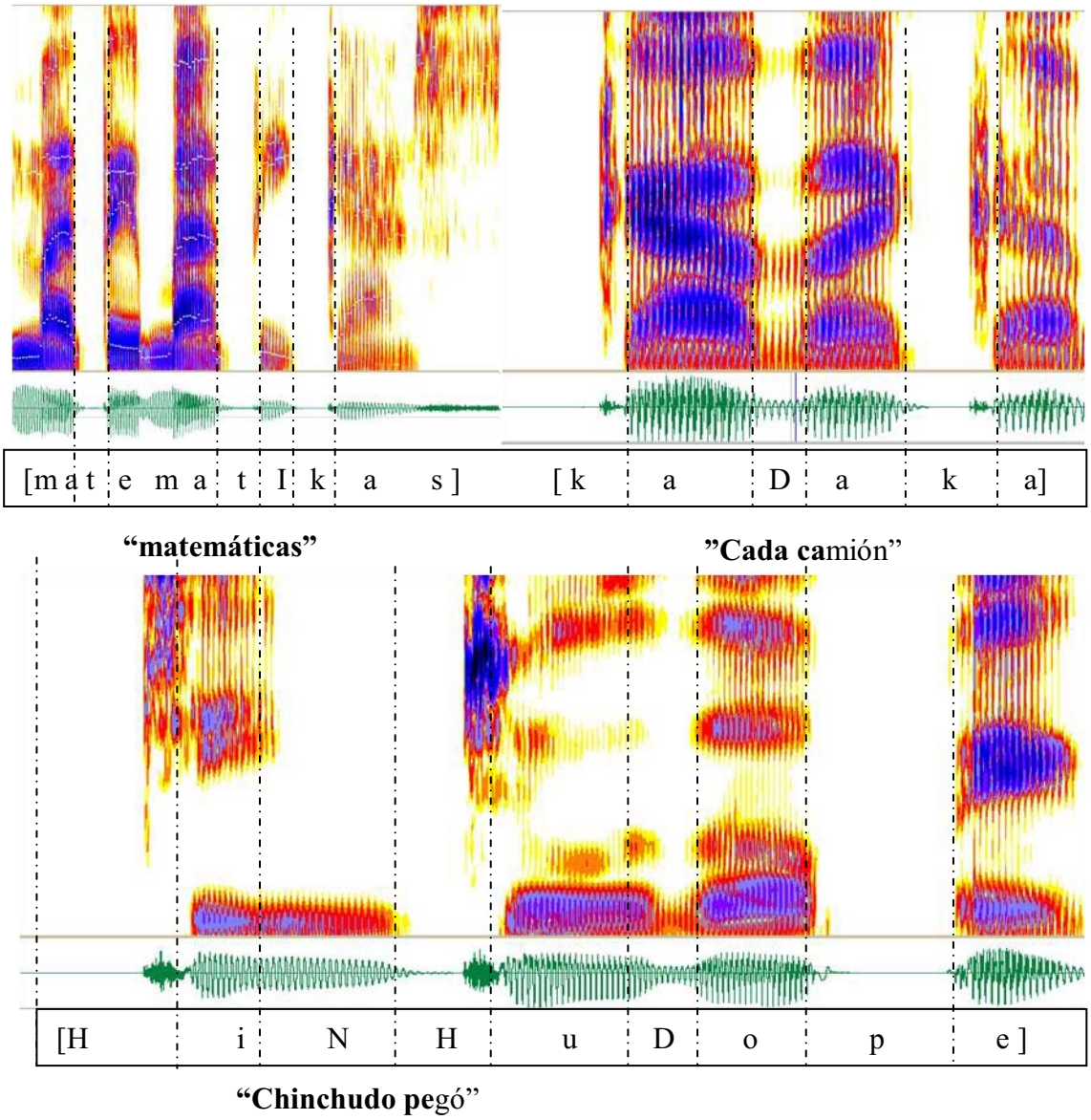


Figura 5. Representación espectrográfica de [p], [t], [k] y [H]

Sonidos [s,f,x,C,Z]

Son sonidos compuestos solo por ruido. Se distinguen claramente de los sonidos periódicos y de los pulsos. Ver Figura 6. El sonido [s] tiene una duración típica de 100 a 200 mseg con unas concentración de alta energía en la banda de 4000 a 5000 Hz.

Cuando se produce en secuencias pre-consonánticas se produce la variante aspirada [h] [9], donde la duración disminuye notablemente a 30-60 mseg y su energía se distribuye en todo el rango de frecuencias. El sonido [f] es de una duración similar a la [s] pero de baja intensidad a lo largo de todo el espectro de frecuencias. El sonido [x] y su alófono [C] presentan una concentración de alta energía en la banda de frecuencias media, entre los 2000 y 3500Hz.

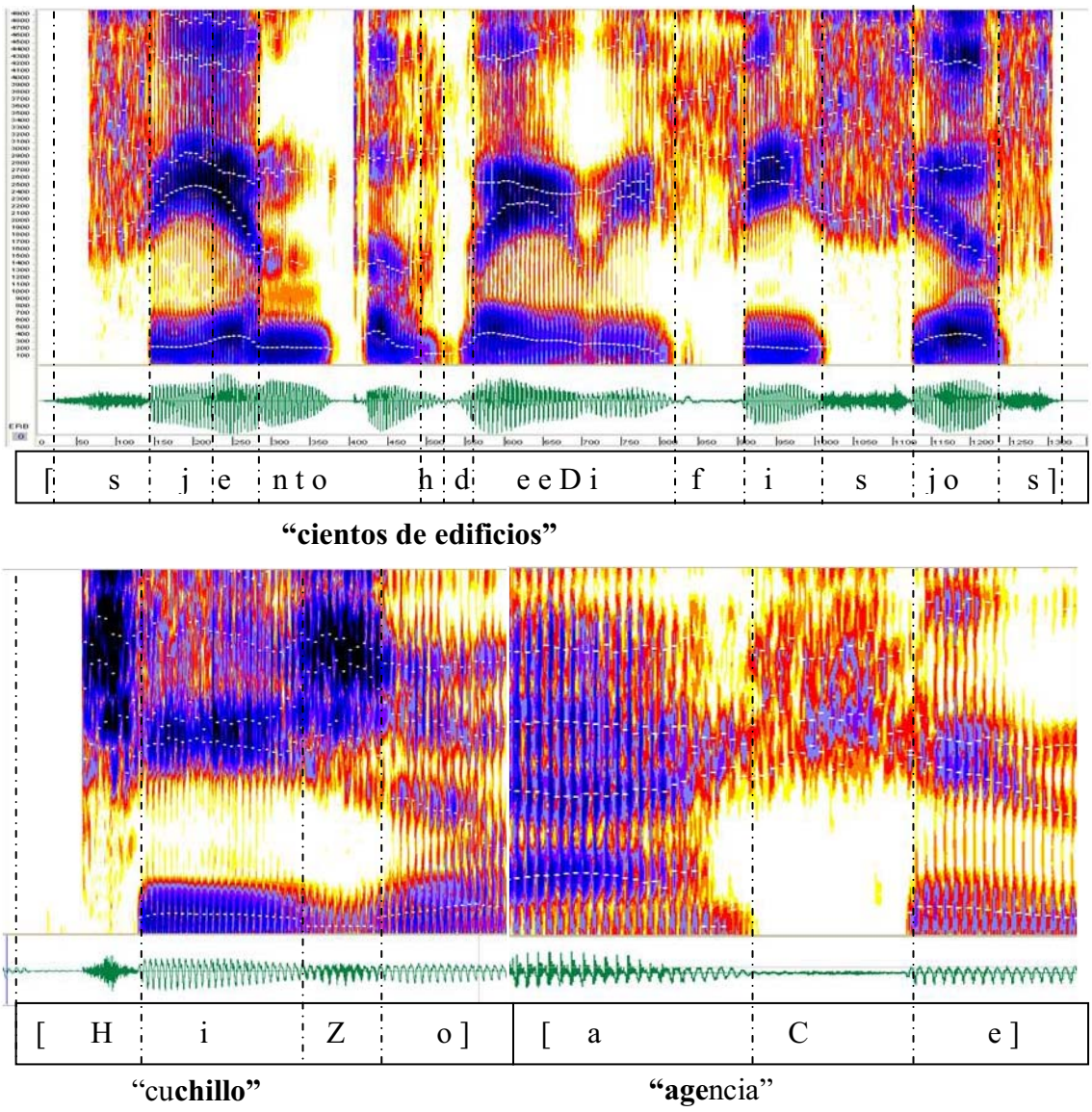
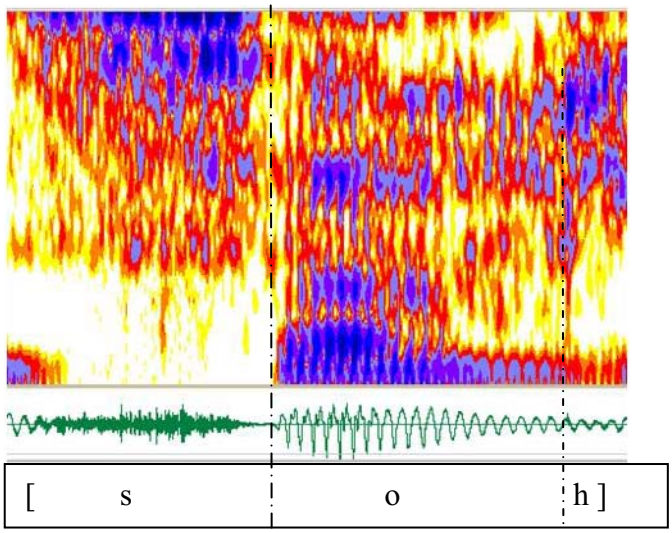


Figura 6. Representación espectrográfica de [s], [f], [x], [C], [Z]

Alófonos [h] y [C]

El sonido [h] presenta un espectro distribuido en frecuencias y puede aparecer superpuesto a una amortiguación periódica, en cambio [s] posee una concentración en 4500 Hz con mayor duración y sin vibración periódica. Ver Figuras 7 y 8.



"sos"

Figura 7. Representación espectrográfica de [s] y[h].

El fonema /x/ presenta una banda de ruido cuya posición es muy dependiente de la vocal. Esta dependencia vocálica divide las realizaciones en [C y x]. Ver Figura 8.

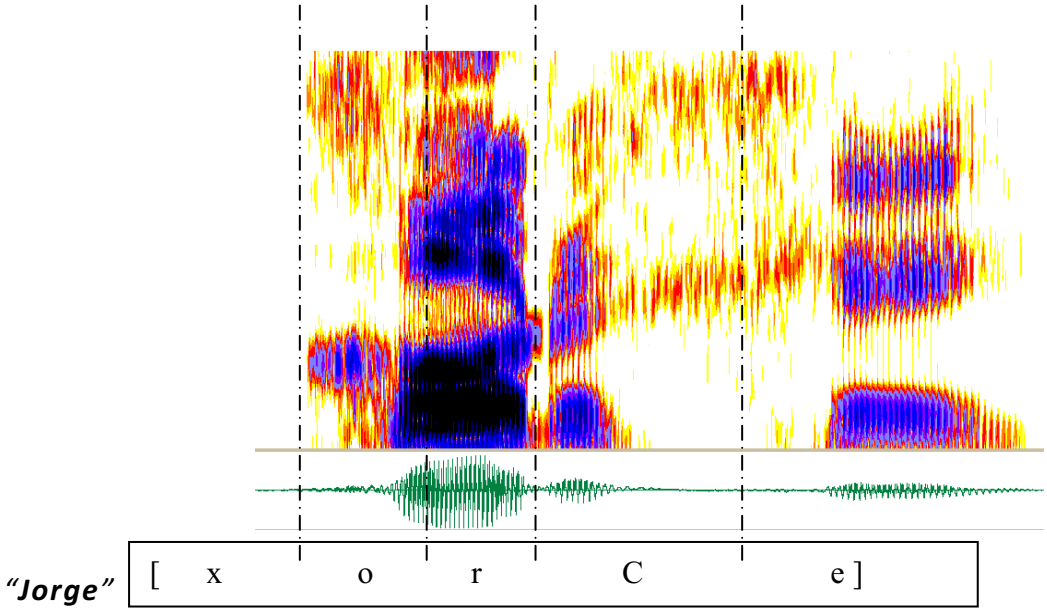


Figura 8. Representación espectrográfica de [x] y [C].

Sonido [r]

El sonido [r] puede presentarse como una breve interrupción de 10 a 20 mseg entre dos sonidos vocálicos. Ver Figura 9. En secuencias de grupos consonánticos CC donde una de las C es [r] se visualiza una vocal epentética que se asignará a la consonante [r]. Ver Figuras 10 y 11. Esta vocal no audible aparece entre la consonante [r] y la consonante con la que forma la secuencia. Ejemplo: [tres] → [teres], [arte] →[arate]. En posición final de [r] seguida de silencio también aparecerá una vocal epentética. Ejemplo: [ver]→[vere], [mar]→[mara].

Opcionalmente otras consonantes pueden presentar una vocal epentética como finalización o inicio. En estos casos también se asignarán a la consonante que produce la epéntesis. La vocal epentética posee una duración menor a la de las vocales con una estructura de formantes y energía equivalentes a la vocal más cercana. De acuerdo a estos contextos hay tres realizaciones de [r]: 1. sin vocal epentética, 2. con vocal epentética a la derecha de la interrupción y 3. con vocal epentética a la izquierda de la interrupción.

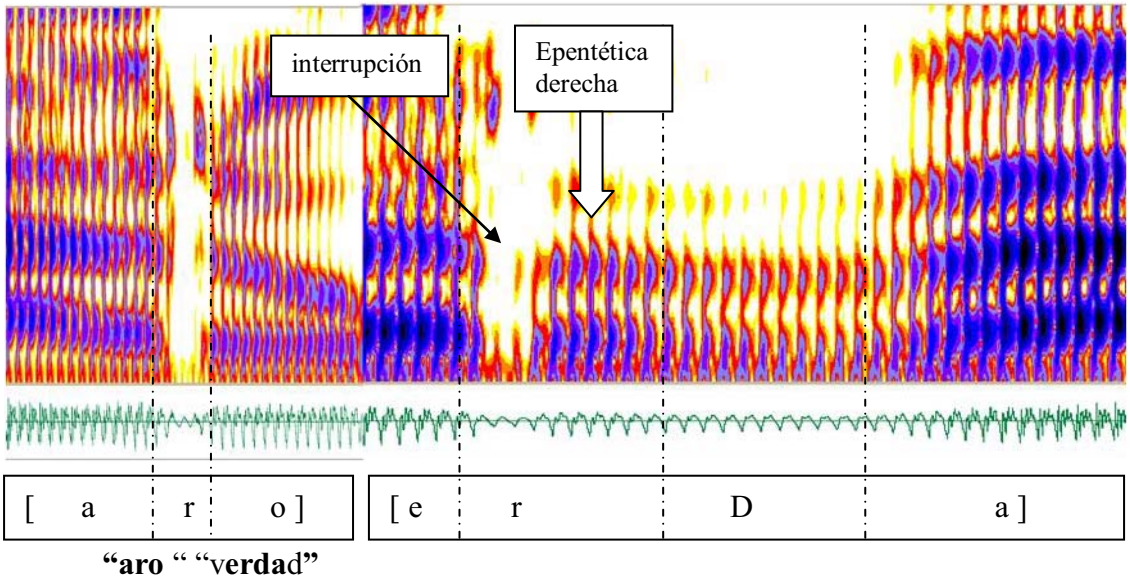


Figura 9. Realización del sonido /r/. En posición intervocálica y en grupo r+consonante y consonante +r.

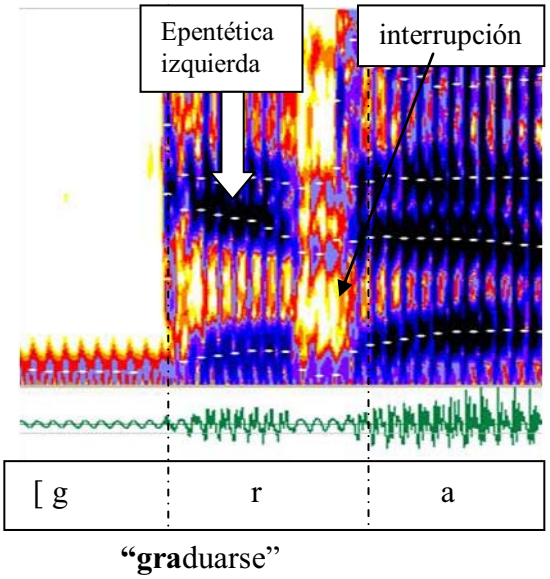


Figura 10. Realización del sonido /r/. En grupos consonante + r .

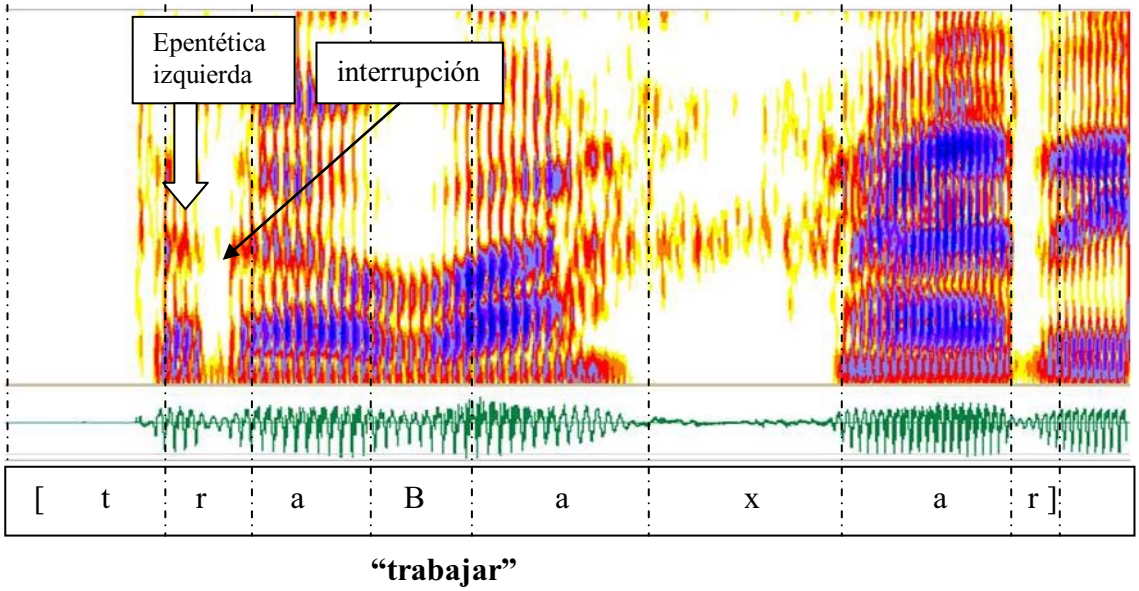


Figura 11. Realización del sonido /r/. En grupos consonante + r y en posición final.

Sonido [R]

Este sonido presenta más de una interrupción rápida de 10 a 20 mseg en la secuencia de formantes vocálicos. Ver Figura 12. Posee distintas variantes: ruidosa, aproximante y combinadas según la región dialectal. [10]

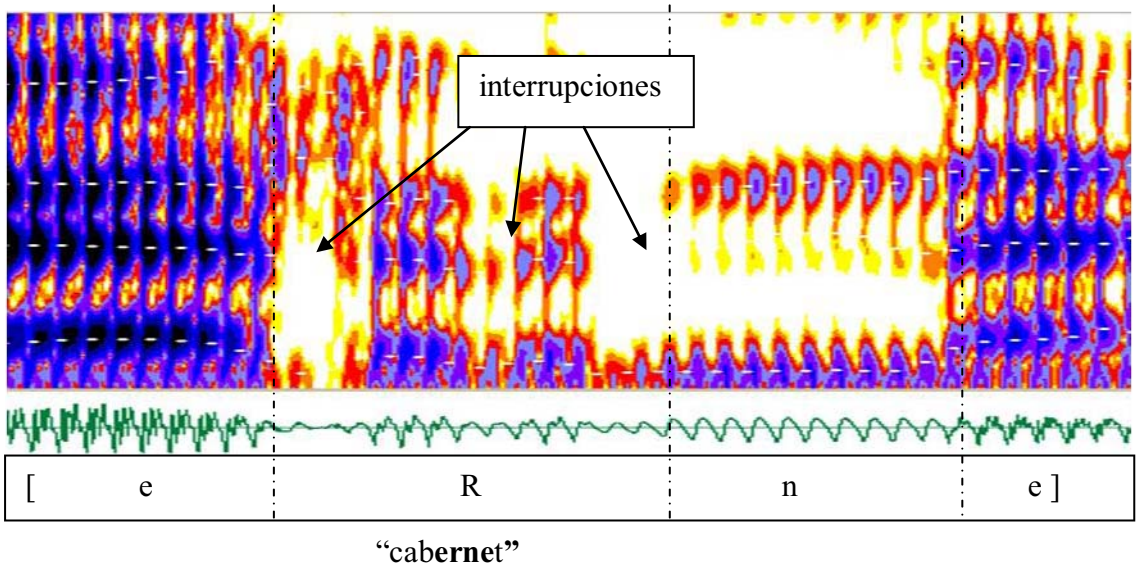


Figura 12. Representación espectrográfica de [R].

Sonido [l]

Es un sonido con una estructura de formantes bien definida similar a una vocal /e/. Ver Figura 13.

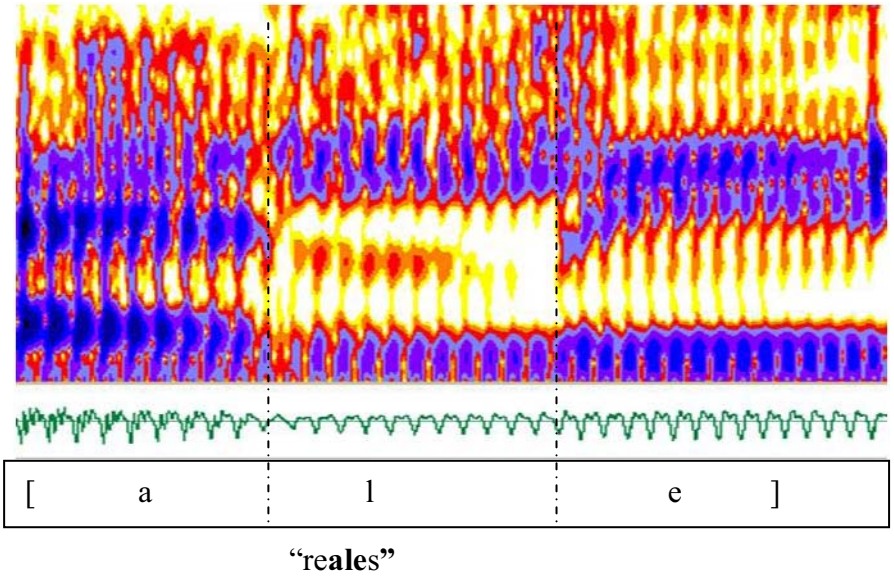


Figura 13. Representación espectrográfica de [l].

Sonidos [b,d,g] y [B, D,G]

Los fonemas /b,d,g/ se realizan como los fonos [b,d,g] y sus alófonos [B,D,G]. Son sonidos con vibración glótica y una oclusión incompleta por lo que presentan ciclos de muy baja energía. En [b, d, g] el espectro se limita a una resonancia de muy baja frecuencia y un ruido de explosión breve muy débil. En [B,D,G] se visualiza energía en los formantes superiores -como mínimo en el 2do

formante- sin ruido [11]. Las transiciones a la vocal son lentas 30-40 mseg. La segmentación debe asignar las transiciones de los formantes de mayor amplitud a la vocal y las transiciones atenuadas la consonante. Ver Figuras 14-17.

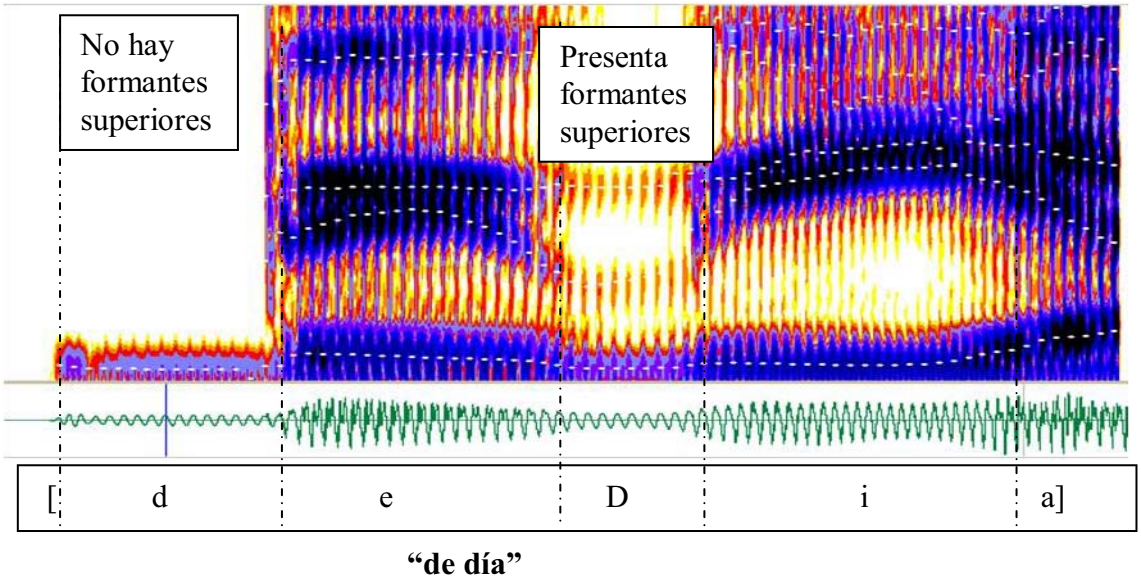


Figura 14. Representación espectrográfica de [d] y [D]

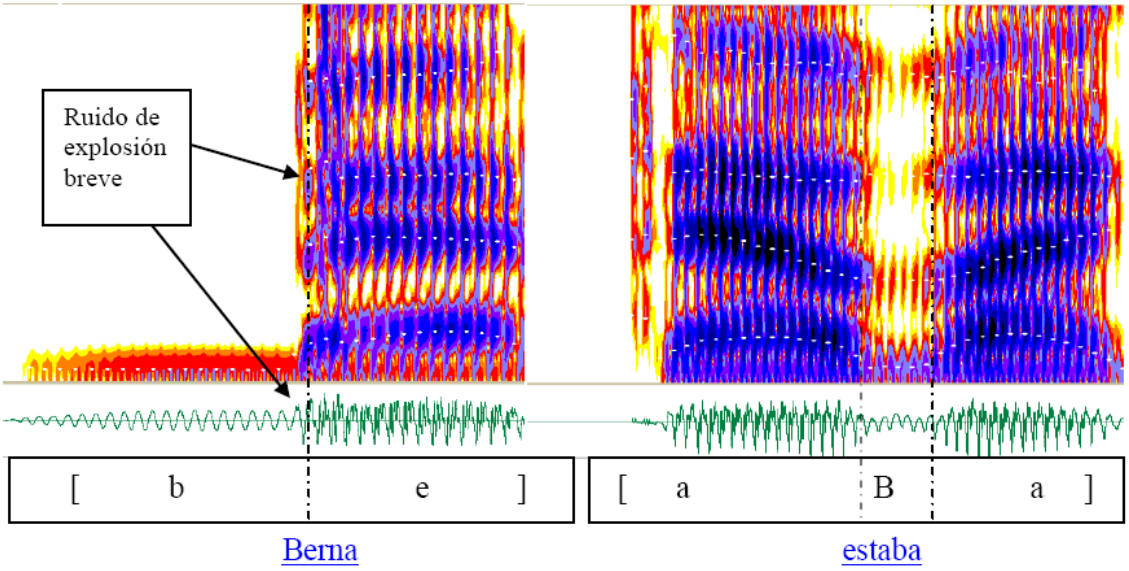


Figura 15. Representación espectrográfica de [b] y [B]. "**Be**rna", "est**aba**".

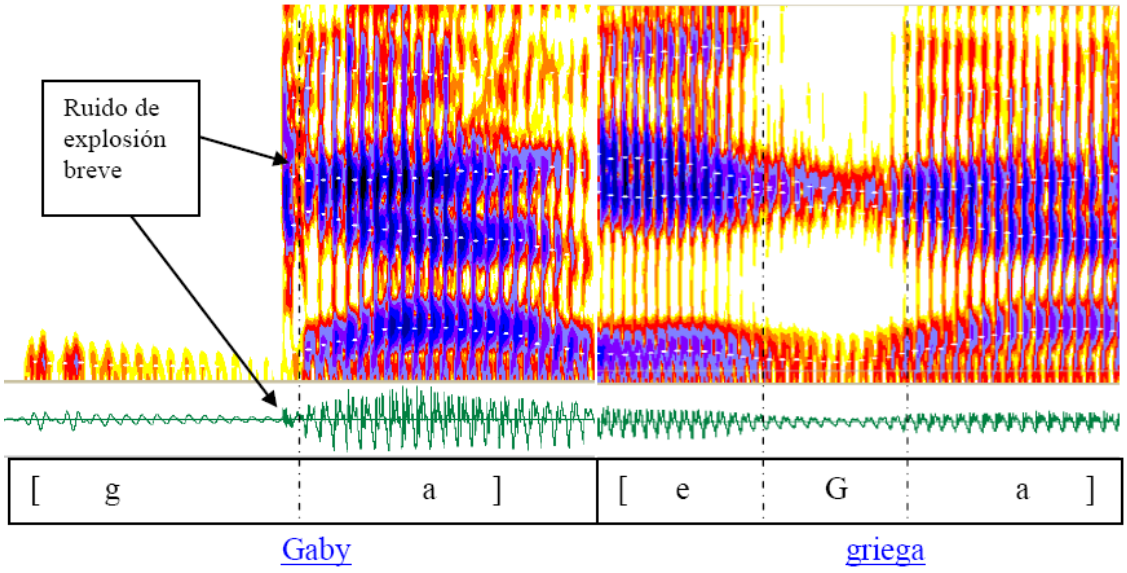


Figura 16. Representación espectrográfica de [g] y[G]. "**Ga**by, "gri**ega**".

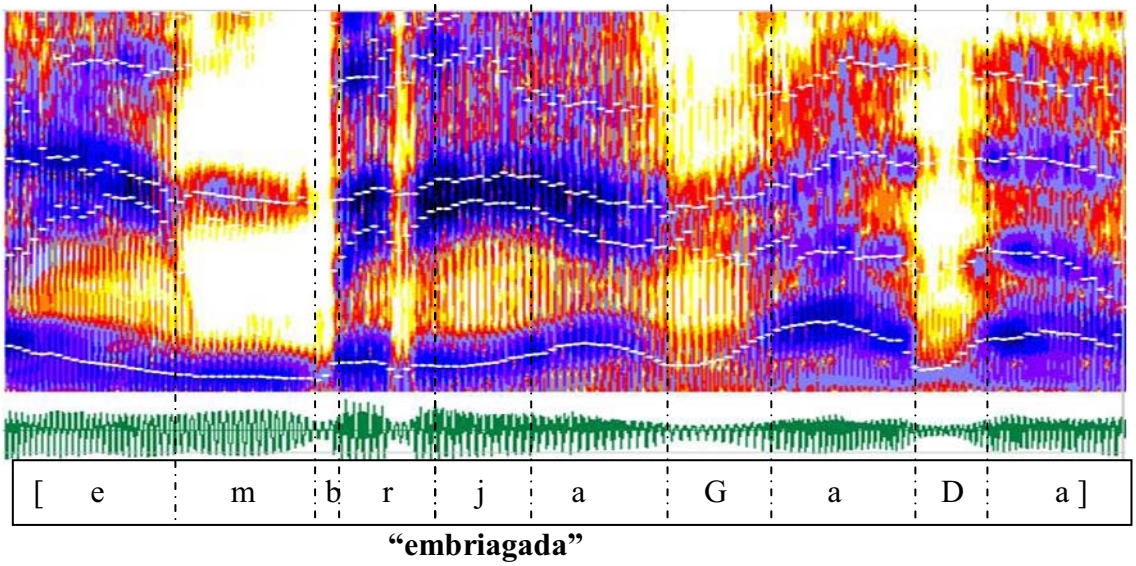


Figura 17. Representación espectrográfica de [b], [G] y [D].

Sonidos [Z] y [L]

[Z] es un sonido con ruido de alta intensidad en la banda de 3000-4000 Hz puede aparecer superpuesto a ciclos de vibración glótica cuasi-periódica. El segmento con ruido se asigna completamente a la consonante. Ver Figura 18. [L] es un sonido con vibración glótica y formantes sin ruido. Las transiciones de menor amplitud corresponden a la consonante.

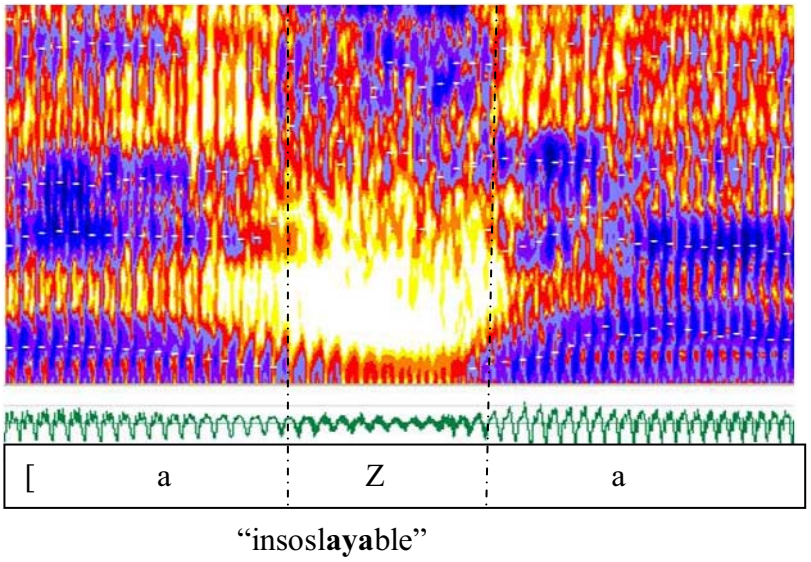


Figura 18. Representación espectrográfica de [Z].

Sonidos [m, n, N, J]

Se caracterizan por presentar dos formantes nasales de menor energía que las vocales entre 250 y 550 Hz para FN1 850 y 1250 para FN2. Los formantes nasales poseen una rápida transición con los sonidos orales adyacentes. Este cambio abrupto y la atenuación o ausencia de continuidad de los formantes orales son un buen indicador del inicio y fin del sonido nasal. El sonido n presenta el alófono [N] que aparece antes de una consonante velar [g, x, C, k, h]. Este sonido se transforma en [m] antes de [p,b,f]. Ver Figuras 19, 20 y 21.

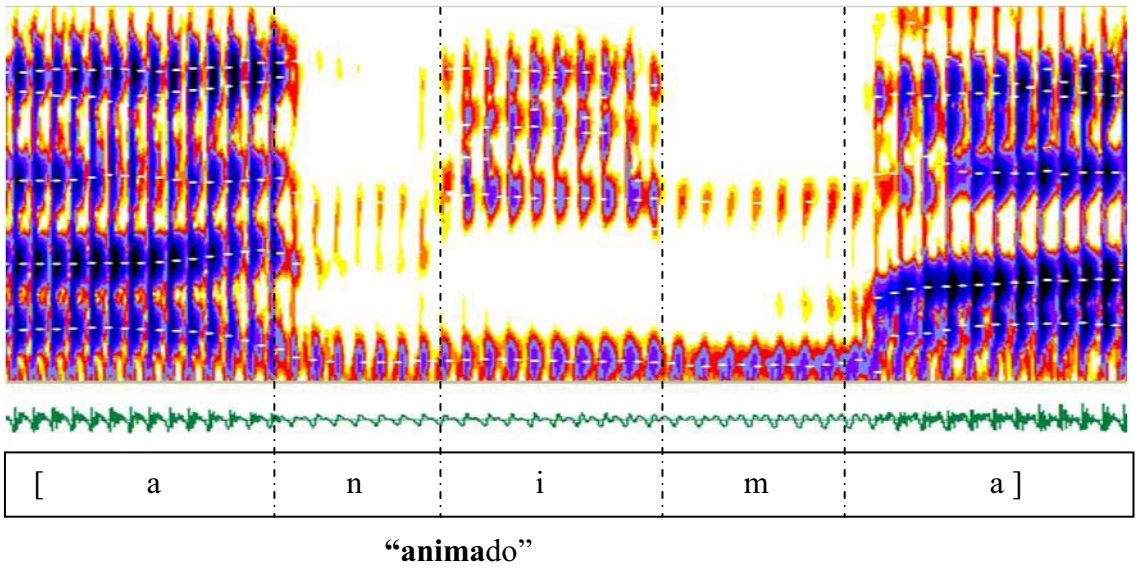


Figura 19. Representación espectrográfica de [n] y [m]. Se observa en el segundo formante la coarticulación dental versus la labial.

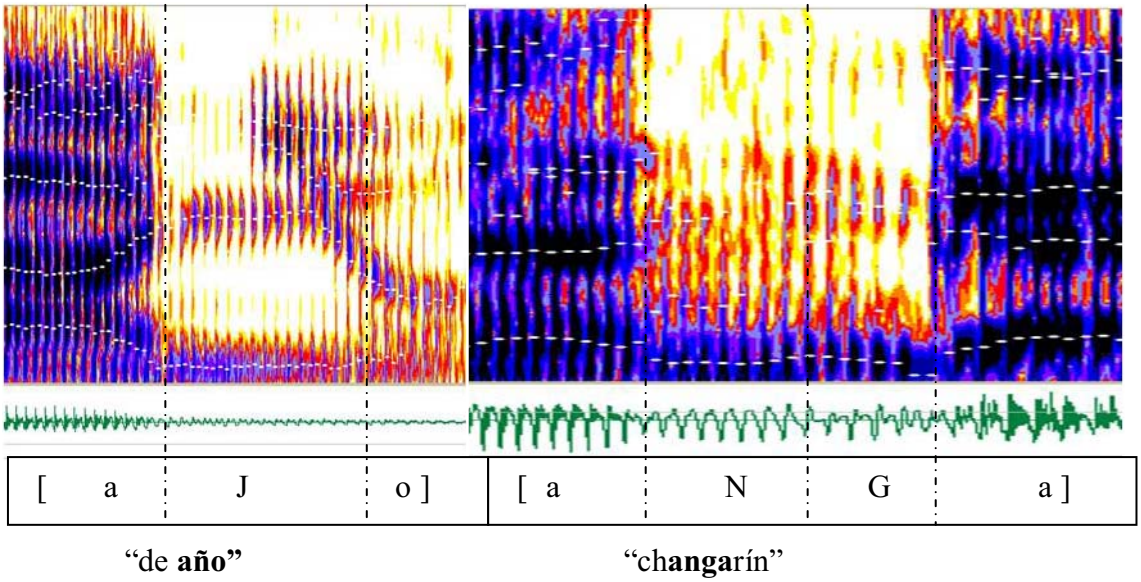


Figura 20. Representación espectrográfica de [J] y [N]. Obsérvese la realización de G después de nasal que contradice la regla de una producción oclusiva.

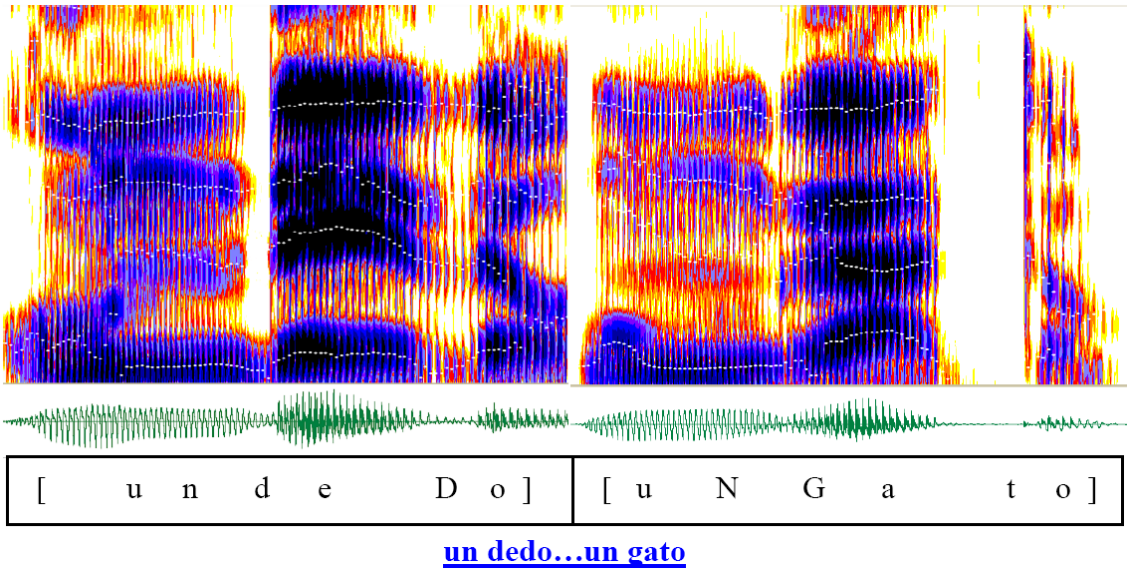


Figura 21. Representación espectrográfica de [n],[d], [D], [N] y [G]. Obsérvese la realización de G después de nasal que contradice la regla.

Segmentación Prosódica

La segmentación prosódica manual o su corrección se requieren para el diseño de los modelos de entonación y ritmo. Se indican con las marcas de pausa de frases y grado de separación entre palabras, los acentos de frases y tonos de juntura y acentos tonales. [12]

Pausas

Las frases se delimitan por cambios de tono abruptos o silencios, en ambos casos la sensación producida es la de una pausa. La notación es con un grado numérico de 3 ó 4. Dentro de las frases se indica el grado de pausa entre palabras con los números 0, 1, 2. Valores orientativos en función de la consonante C y vocal V al final de una palabra e inicio de la siguiente palabra son los siguientes: grado 0 para V1V1 y C1C1, grado 1 para CV, V1V2 y VC y grado 3 para C1C2. Los números 1 y 2 representan distintas vocales y consonantes.

Acentos de Frase

Se observan al final de la frase entonativa en el contorno de F0. Se indican como H- (alto), M- (medio) y L- (bajo).

Acentos tonales

Se perciben en la sílaba alrededor de cada acento léxico de la palabra de contenido. Si la palabra recibe foco estrecho, el máximo de F0 estará en correspondencia temporal con el segmento acústico de la vocal de la silaba acentuada lexicalmente. El acento de frase no cambiará esta asociación. Se indica con un acento mono-tonal H* o L*, bitonales L*+H, H*+L, L+H*, H+L* y con tritonales [13].

Si la palabra recibe foco amplio, el máximo de F0 se observa en las cercanías de la sílaba con acento léxico o puede no existir. La posición de la palabra, el tipo de palabra, el acento de frase y las variaciones estilísticas del hablante influyen en esta localización. Por ejemplo: si el acento de frase es alto H-, el acento tonal tenderá a desplazarse hacia la derecha, acercándose al acento de frase en las palabras finales con acento nuclear como L*+ H en las palabras esdrújulas y graves, ó fundiéndose con él como L+H* en las palabras finales agudas. Este desplazamiento también se observa con claridad en palabras no finales con acentos pre-nucleares.

Si el acento de frase es bajo L-, los acentos tonales pueden desplazarse hacia la izquierda, alejándose de acento L- o más aun desapareciendo en la palabra final con acento nuclear. El alejamiento es mayor en palabras finales con acento lexical agudo H+L* y menor o sin corrimiento en los otros tipos de acentos.

Tonos de Juntura

Se observan en el extremo del contorno de F0 en la última frase de la oración. Pueden ser H%, L% ó M%. Hay notaciones en la literatura donde el tono de Juntura puede contener al acento de frase en una única secuencia. Se indica como H-H% ó HH%.

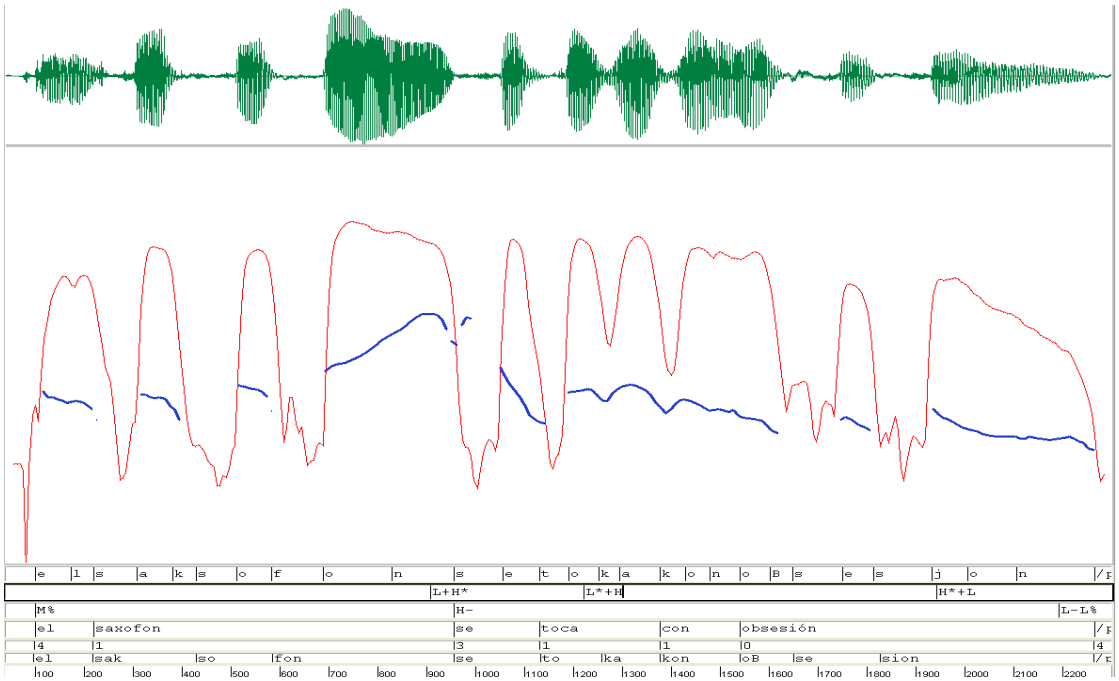


Figura 22. Oración Declarativa. “El saxofón se toca con obsesión”.

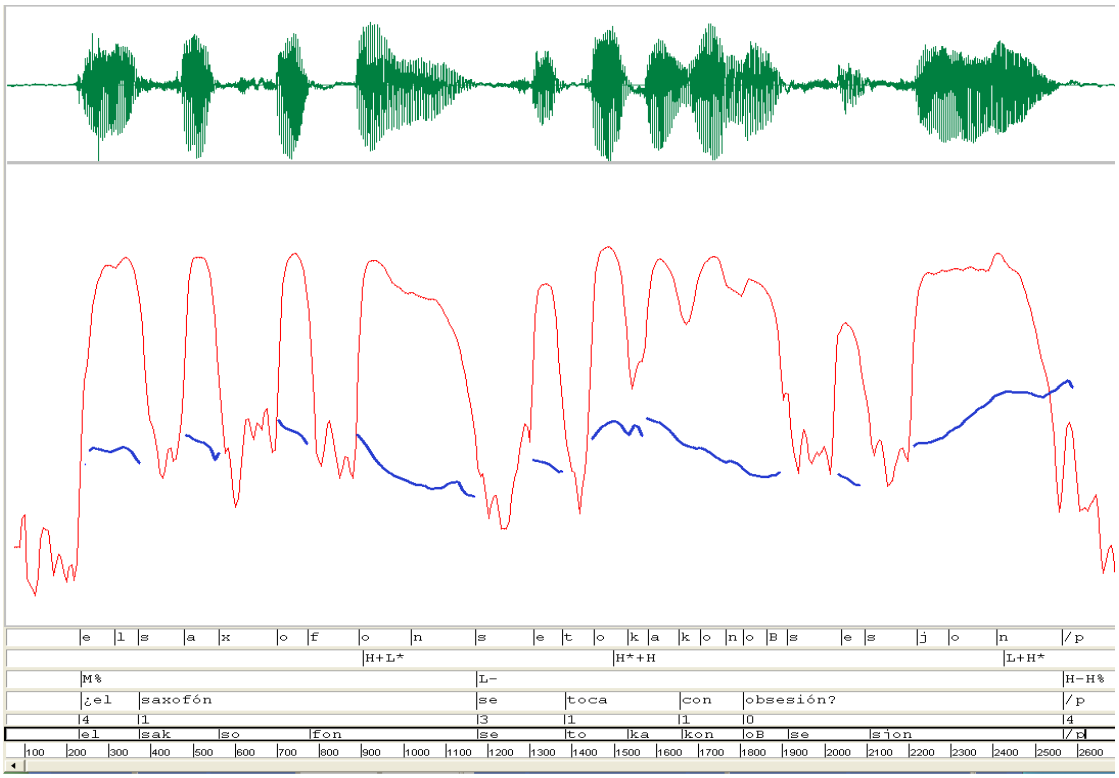


Figura 23. Oración Interrogativa. “¿El saxofón se toca con obsesión?”.

CONCLUSIONES

Se han propuesto los criterios para realizar la segmentación acústico-fonética y el etiquetado fonético al nivel segmental y suprasegmental requeridos para caracterizar las bases de datos con aplicación a las tecnologías de habla. Los casos que contradicen las reglas deben indicarse para estimar estadísticamente sus frecuencias de aparición y asociarlos eventualmente como características de un locutor particular. La forma de onda permite delimitar los segmentos acústicos con mayor precisión. En segundo término se empleará el espectrograma y la evaluación perceptual de sílabas o palabras.

AGRADECIMIENTOS

A la Fga. Lic. M.E. Vaccari por la revisión del manuscrito. A la Agencia de promoción Científica y Técnica (ANPCyT) por el apoyo financiero a través del proyecto PAE PID 094. Al Fontar (Fondo Tecnológico Argentino) por el proyecto FONSOFT a la Empresa Global News Group.

BIBLIOGRAFÍA

- *Torres, HM, Gurlekian, JA and Cossio M, C. 2012. Aromo: Argentine Spanish TTS System". In: Proc. VII Jornadas en Tecnología del Habla and Iberian SLTECH Workshop, IberSpeech 2012, 416-421. Madrid.*

- *Gurlekian J. A., Rodríguez, H., Colantoni, L. and Torres, H. "Development of a Prosodic Database for an Argentine Spanish Text to Speech System", Proceedings of the IRCS Workshop on linguistic databases. Bird, Buneman &Liberman (eds). Philadelphia, pp. 99-104, December 2001.*

- *Gurlekian JA, Colantoni, L. , Torres, H. , Rincón, A. , Moreno, A. and Mariño, J. "Database for an Automatic Speech Recognition System for Argentine Spanish", Proceedings of the IRCS Workshop on Linguistic Databases. Bird, Buneman & Liberman (eds). Philadelphia, pp. 92-98, December 2001.*

- *Univaso, P. Martinez Soler, M., Evin, D.A Gurlekian JA "A preliminary approach to forensic speaker recognition using phonemes". IberSpeech 2012. Madrid.*

- *Sigal, L y Gurlekian, JA. 2014. Aplicación de los índices de perturbación integrado y de precisión articulatoria en pacientes con disfonía espasmódica. Disponible on-line en: http://revistalogopedia.uclm.es. Revista de Investigación en Logopedia 2:132-150.*

- *Wells, J.C., 1997. 'SAMPA computer readable phonetic alphabet'. Handbook of Standards and Resources for Spoken Language Systems. Berlin and New York: Mouton de Gruyter.*

- *Gurlekian J. A., Colantoni, L. y Torres, H., 2001 "El alfabeto fonético SAMPA y el diseño de córpora fonéticamente balanceados". Fonoaudiológica, Ed. Asalfa, 47, 3, 58-69.*

- *Guirao, M., 1980. Categorías psicoacústicas de señales de habla, Acústica 1 (2):20-49.*

- *Borzone de Manrique, A.M., 1980. La /s/ pre- consonántica en el español de Buenos Aires. Fonoaudiológica, Ed. Asalfa, 26, 1, 245-250.*

- *Martínez Soler M., Univaso P., Gurlekian J. A., "A Dissimilatory Process on the Phoneme /r/ in the Province of Corrientes, Argentina", en preparación.*

- *Santagada, M.A. and Gurlekian J.A., 1993. "Spanish Voiced Stops in VCV contexts: Are they fricative variants or aproximants?", Revue de Phonetique Appliqué, Nº 90-91-92, pp. 90-100.*

- *Beckman y Gayle A. E., 1997. "The guidelines for ToBI labeling", The Ohio State University Research Foundation.*

- *Gurlekian J. A., Torres, H. y Colantoni, L., 2004. "Evaluación de las descripciones analítica y perceptual de la entonación de una base de datos de oraciones declarativas de foco amplio para el español hablado en Buenos Aires. Estudios de Fonética Experimental, Nro: XIII, pp. 275-302.*

ANEXO

Símbolos utilizados para la marcación de eventos acústicos para el entrenamiento de sistemas de reconocimiento de habla.

Hablante/Locutor: [sn] con n=1,2….etc. Simultáneos [s1s2]

Ruidos: [r]

Publicidad (separadores): [p]

Música o cortina: [m]

Jingle: [pm]

Silencios (baches): [sil]

Palabras no inteligibles: [**]

Palabras Cortadas: palabr-

Mal pronunciadas: ***mal-pronunciada bien-pronunciada***

Siglas deletreadas: +Y +P +F

Palabras extranjeras o no existentes en el diccionario de la RAE: @iu @tub

Falso arranque: (cua-) (cua-) cuando

Ruidos del hablante identificados: [riw] "risa", [jaw] "jadeo", [tow] "tos", [chw] "chasquido"

Ruidos del hablante no identificados: [spk]

Ruidos cortos no humanos: [int]

Pausas llenas (muletillas): [eeh],[esteh],[estemh],[mmh]

Alargamiento de la vocal final de palabra: [i], [e], [a], [o], [u]

Tanto los ruidos identificables del locutor como las pausas llenas no descriptas pueden agregarse respetando el último carácter, respectivamente w y h.