UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO
FACULTAD DE FILOSOFÍA Y LETRAS
LICENCIATURA EN LENGUA Y LITERATURAS HISPÁNICAS

MANUAL DE ETIQUETADO FONÉTICO E IMÁGENES ACÚSTICAS DE LOS ALÓFONOS DEL ESPAÑOL DE LA CIUDAD DE MÉXICO, PARA SU USO EN LAS TECNOLOGÍAS DEL HABLA

TESIS QUE, PARA OBTENER EL TÍTULO DE
LICENCIADA EN LENGUA Y LITERATURAS HISPÁNICAS,
PRESENTA

MONTSERRAT ALEJANDRA CHAVARRÍA AMEZCUA

Asesores: Mtro. Javier Octavio Cuétara Priede
Dr. Luis Alberto Pineda Cortés

Ciudad de México, 2010

"que toda la vida es sueño, y los sueños, sueños son"

Calderón de la Barca

Pienso que en ocasiones ésta es la parte más difícil. Decir gracias no es sencillo, pero lo digo con todo el corazón. GRACIAS:

Dios por permitir que llegara este momento, por haberme puesto en este tiempo, en este lugar, con todas las a las que tanto amo…

Papá y mamá: por confiar en mí, por enseñarme y darme la libertad, por soportar cada una de mis rebeldías, por enseñarme a crecer, a ser fuerte y a ser autosuficiente, por comprender mis locuras y mis sueños, por preocuparse por mí, por cuidarme y apapacharme cuando lo necesito, y, sobre todo, por el AMOR y el apoyo incondicional que me han dado.

Hermanas Lola, Angélica y Giovana: por crecer conmigo, por ser mis compañeras de juegos, por todos los más hermosos y mejores momentos, vivencias y recuerdos de mi vida; porque sin ustedes, en definitiva, la vida no sería igual. Las amo tanto como a nuestros papás, que es más que amar a mi propia vida, que a mí misma.

Ma. Hor: por quererme tanto y nunca olvidarse de mí.

Amigas Fer, Yu, Lu, Lau, Mon, Dino y Aure: por todos buenos momentos de la universidad que compartimos, por todos los años que hemos estado y estaremos juntas, por ser las mejores amigas, por ser las mejores confidentes, por todo el apoyo y la comprensión, por todo. Las amo nenas.

Javier: Por creer y confiar en mí, por aceptar ser mi asesor en este trabajo, por hacerme crecer y madurar, por todo el apoyo, por ser el mejor maestro, por enseñarme tantas cosas, por ser un ejemplo a seguir, por todos los cocos, los jalones de orejas y los manazos tan oportunos; por todas la risas, por preocuparte por mí, por abrirme las puertas de tu casa, por toda tu amistad.

Doctor Luis: por confiar en mí, por su amistad, por tenderme la mano, por darme todo lo necesario en mi estancia en el departamento, por permitirme ser parte de su proyecto, por no dejarme sola y preocuparse por mí.

Dra. Margarita, Dra. Lilian, Dra. Axel: por ser parte de mi formación académica, por ser excelentes profesoras y mostrarme nuevos mundos con sus conocimientos. Por leer este trabajo que se nutrió de sus observaciones.

Compañeros y amigos del DDC-IIMAS: por aguantarme todos los días. Siempre he dicho que somos una familia a la que le doy los buenos días, el buen provecho y las buenas noches. Por las porras y todos los hermosos momentos que hemos pasado juntos.

A los que ya no están y aún viven en mi corazón: Abuelo Ángel y padrino Vic, los extraño y los recuerdo con mucho amor.

Gary: por todas las fiestas y carreras que haces al verme llegar.

A todos aquellos que no he nombrado, pero que son sumamente importantes en mi vida.

Agradezco inmensamente a la Universidad Nacional Autónoma de México que me ha dado la oportunidad de una excelente educación y que me ha permitido ser una universitaria orgullosamente puma.

Sin dejar de ser yo: "voy a llorar".

Agradezco al Proyecto CONACYT "GOLEM-II: Un asistente conversacional situado con lenguaje hablado y visión computacional", con número de referencia 81965.

Y al proyecto PAPPIT-UNAM "Desarrollo de módulo de procesamiento semántico estocástico para el robot Golem con un corpus con etiquetación mínima", con número de referencia IN-115710.

Ambos coordinados por el Doctor Luis A. Pineda Cortés.

Índice

Índice de figuras

1. Introducción

La lingüística computacional se dedica a la construcción de herramientas computacionales para el análisis y estudio del habla. En las últimas décadas, esta disciplina ha tenido un gran auge en la ciencia y en la tecnología debido a los grandes avances de sus estudios. Una de las ramas de la lingüística computacional son las tecnologías del habla; éstas se enfocan, entre otras cosas, a la construcción de sistemas de síntesis y reconocimiento del habla: los primeros están orientados a la recreación del habla y los segundos al reconocimiento de la misma. El objetivo de estos sistemas es lograr la interacción humano-máquina por medio del habla.

En la actualidad se ha podido lograr la comunicación entre humanos y máquinas, aunque de forma muy limitada debido a diferentes factores, entre ellos, los lingüísticos como la polisemia, la conversación diversa, un léxico limitado, la falta de un sistema de reconocimiento más eficaz, entre otros. Para solucionar esta limitación, las tecnologías del habla están desarrollando sistemas de síntesis capaces de recrear un habla cada vez más parecida a la humana y reconocedores que la capten con más facilidad, especialmente el habla espontánea. En esta tarea las tecnologías del habla trabajan conjuntamente con la fonética, pues para que los sistemas de síntesis y reconocimiento sean más eficaces requieren de una descripción detallada de cada sonido que emitimos los humanos al hablar (Llisterri 1991).

La construcción de sistemas de síntesis y reconocimiento de habla conlleva un largo proceso. En primer lugar, es necesario compilar un corpus oral bastante amplio, el cual se analiza acústicamente con herramientas computacionales que permiten ver la señal sonora, tales como los oscilogramas y los espectrogramas. Posteriormente, se pasa a la etapa de transcripción del corpus. Esta etapa representa un trabajo largo y, sobre todo, laborioso, ya que cada oración o palabra del corpus se segmenta en alófonos y éstos, a su vez, se transcriben en símbolos que pertenecen a un alfabeto fonético; dichos símbolos deben

describir las características acústicas de los sonidos, de manera que al terminar de segmentar y transcribir cada uno de los alófonos en las etiquetas, la oración queda transcrita fonéticamente. A este proceso, en la jerga de la lingüística computacional, se le conoce como etiquetado fonético computacional.

La transcripción del corpus se hace manualmente con ayuda de herramientas computacionales que incluyen espectrogramas y oscilogramas. Éstas, entre otras funciones, permiten reproducir las grabaciones obtenidas de los corpus, segmentar la imagen espectrográfica de cada alófono y hacer su transcripción fonética, lo que facilita en gran medida el trabajo del transcriptor.

A simple vista, el proceso de transcripción puede parecer un trabajo sencillo; sin embargo no es así, pues la transcripción de un corpus representa una gran dificultad debido a varios aspectos. En primera instancia, un corpus está constituido por varias horas de grabación, por lo que se requiere de mucho tiempo para transcribirlo y, aunque existen herramientas computacionales que de alguna manera facilitan este trabajo, se necesita un grupo numeroso de personas destinadas a esta labor. En muchas ocasiones, las personas que van a transcribir el corpus suelen pertenecer a distintas formaciones académicas como lingüistas, ingenieros, informáticos o matemáticos; por esta razón muchos son inexpertos y es necesario darles una capacitación previa sobre el etiquetado fonético computacional; esto hace que el tiempo de etiquetado del corpus se retrase, se haga complicado y tedioso. A lo anterior se suma que, cuando un corpus se transcribe por varias personas, parte de ellas principiantes, provoca que haya inconsistencias y falta de uniformidad en el etiquetado, como la mala delimitación en las fronteras de los sonidos, la confusión de alófonos y etiquetas mal colocadas, por mencionar algunos casos.

A partir de las dificultades que se presentan durante el proceso de transcripción, se buscó una solución que contribuyera para que el proceso sea más consistente y estable. Por lo anterior, este trabajo de tesis hace una revisión sobre la fonética y las tecnologías del habla, y su mayor contribución es proponer un *Manual* d*e etiquetado fonético e imágenes acústicas de los alófonos del español de la ciudad de México, para su uso en las*

tecnologías del habla. El *Manual* consiste en la presentación de la imagen acústica de los alófonos del español de la ciudad de México, pues se espera que al reconocerlos visualmente en un espectrograma sea posible segmentarlos de manera más precisa.

Esta tesis se enmarca dentro del *Proyecto DIME* (*Diálogos Inteligentes Multimodales en Español*). Este proyecto se desarrolla en el Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas (IIMAS) de la UNAM. El proyecto tiene como objetivo lograr la comunicación entre humanos y máquinas, mediante el lenguaje natural; para ello, una de sus tareas ha sido la construcción de sistemas de reconocimiento del habla en español de México (Pineda 2008). Para llevar a cabo esta tarea, el proyecto se dio a la labor de diseñar, recopilar y transcribir fonéticamente el *Corpus DIMEx100* con el fin de crear los modelos acústicos y los diccionarios de pronunciación para la programación de los reconocedores de habla; por esos motivos fue grabado cuidadosamente, dentro de una cabina de audio para evitar el ruido y con un habla cuidada (Pineda *et al.* 2004, 2010). Por lo anterior, este corpus fue elegido para obtener las imágenes ejemplos del *Manual* presentado en esta tesis.

La elaboración de esta tesis responde a varios intereses. En primer lugar, busca contribuir al desarrollo de las tecnologías del habla en México que, al ser una disciplina relativamente joven en nuestro país, carece de herramientas que auxilien, guíen y enseñen a reconocer los sonidos del lenguaje en un espectrograma, a diferencia de otros países que cuentan con diversos manuales y libros que muestran un análisis sobre la onda sonora con esos fines. En segundo lugar, se ha hecho para mejorar y facilitar el proceso de transcripción de los corpus del *Proyecto DIME*; por ello, se limita a describir los sonidos del español de la ciudad de México y se apega a las normas y estándares de transcripción del *Corpus DIMEx100*. En tercer, lugar se espera que sea una base y motivación para futuros trabajos que aborden los distintos dialectos del español de México.

El trabajo aquí presentado se divide en dos partes. La primera parte se divide en cuatro capítulos, incluida esta Introducción. El segundo capítulo comienza con una breve introducción a la lingüística computacional. Continúa con una descripción de las tecnologías del habla y sus aplicaciones: el funcionamiento y la arquitectura de los sistemas

de síntesis y reconocimiento del habla, y finaliza resaltando la importancia de la colaboración de la fonética en la construcción de dichos sistemas, así como las herramientas de las que se ha valido a lo largo del tiempo para los estudios del habla, tales como instrumentos eléctricos y alfabetos.

El tercer capítulo trata acerca de la lingüística computacional en México, específicamente del *Proyecto DIME*. Se presentan los proyectos que se han desarrollado dentro de las tecnologías del habla, en especial del *Corpus DIMEx100*, pues de éste se extrajeron las imágenes ejemplos para el *Manual*. Continúa con una descripción general sobre la recopilación de los corpus orales, del proceso de transcripción y de las herramientas que se usan para llevar a cabo esta tarea. Cabe mencionar que la descripción de estos procesos se apega a la recolección y etiquetado de los corpus del *Proyecto DIME*, y precisamente en estos procesos se pretende aplicar el trabajo aquí presentado, con el propósito de mejorar la transcripción; sin embargo, cualquier otro proyecto puede retomarlos, ya que han mostrado ser consistentes para la elaboración de sistemas de reconocimiento del habla. Posteriormente, se presenta el alfabeto fonético computacional *Mexbet* (Cuétara 2004) y sus diferentes niveles de transcripción, el cual se hizo para el español de la ciudad de México, con el objetivo de usarse en proyectos de tecnologías del habla en nuestro país. Este alfabeto es parte importante de esta tesis, pues el *Proyecto DIME* lo ha utilizado para la transcripción de sus corpus; por lo tanto, las representaciones simbólicas de los alófonos del *Manual* están transcritas con *Mexbet*. Para finalizar, se presenta el uso del programa computacional *Speech View* y sus diferentes herramientas para transcribir.

En el cuarto capítulo se presenta el *Manual de etiquetado fonético e imágenes acústicas de los alófonos del español de la ciudad de México, para su uso en las tecnologías del habla.* Para la construcción de este *Manual*, primero se hizo una revisión en el *Corpus DIMEx100* del comportamiento de la imagen espectrográfica de cada uno de los alófonos del español de la ciudad de México que propone *Mexbet* en su inventario. Esta revisión se hizo con el propósito de hacer un análisis con los datos que arrojaron las distintas imágenes, y con base en él seleccionar la imagen ejemplar para cada alófono. Posteriormente, se recurrió a la teoría ya existente sobre la fonética acústica, la fonética del español de México y las

tecnologías del habla, con el fin de complementar el análisis obtenido a partir de la observación de las imágenes espectrográficas de los alófonos del español de la ciudad de México, y de unificar criterios sobre su comportamiento tanto en el espectrograma como en el oscilograma.

Una vez obtenidos todos los materiales para la elaboración del *Manual*, se organizó su distribución. Para fines más didácticos, el *Manual* está organizado por grupos fonemáticos, es decir, cada capítulo está compuesto por los distintos alófonos de un fonema. Los capítulos comienzan proporcionando datos teóricos sobre cada alófono. Posteriormente se divide en distintos apartados, según el número de alófonos; cada apartado contiene las reglas de realización de los alófonos, la descripción e imagen para su reconocimiento en el espectrograma y, finalmente, se instruye cómo segmentar la imagen espectrográfica.

El *Manual* presentado en esta tesis pretende brindar diferentes aplicaciones prácticas. Las principales son las siguientes:

1) Tener una herramienta que enseñe cómo segmentar y transcribir los corpus orales y, de esta manera, sea posible subsanar las dificultades que se presentan durante el proceso de transcripción
2) Facilitar, agilizar y uniformar el proceso de transcripción de los corpus del *Proyecto DIME* y de otros proyectos interesados en las tecnologías del habla en México.
3) Ser una guía para las personas inexpertas en el etiquetado fonético computacional. Así como, un medio de aprendizaje para los alumnos de la asignatura de fonética.

Por lo anterior, el *Manual* fungirá como una herramienta de ayuda en el etiquetado fonético computacional, pues cuando alguna persona se enfrente al proceso de etiquetado o el etiquetador –experto o inexperto– tenga alguna duda, podrá consultar el *Manual* y tener las herramientas necesarias para segmentar y transcribir correctamente, al mismo tiempo que obtendrá cocimientos sobre fonética acústica.

Finalmente, la segunda parte del trabajo es el *Manual de etiquetado fonético e imágenes acústicas de los alófonos del español de la ciudad de México, para su uso en las tecnologías del habla* que aparece como Apéndice. Se tomó está decisión porque el *Manual* es una unidad que puede prescindir de los capítulos anteriores; de esta manera, cuando el

lector recurra a este trabajo podrá leer únicamente el *Manual*, pues contiene la información precisa sobre la segmentación y la transcripción fonética computacionales.

2. La lingüística computacional

En este capítulo se da una introducción a la lingüística computacional. Se explicará brevemente cada una de sus ramas, las cuales son el procesamiento del lenguaje natural, el procesamiento de textos y las tecnologías del habla; así como en qué consiste cada una de ellas. Se ahondará en la rama de las tecnologías del habla; se presentarán dos de sus aplicaciones, que son los sistemas de síntesis y reconocimiento del habla; de ellos se describirá su arquitectura y funcionamiento. Posteriormente, se resalta la colaboración de la fonética en las diferentes tareas de la construcción de los sistemas de síntesis y reconocimiento, tales como el diseño, la recolección y análisis de los corpus orales.

En este mismo capítulo se describe brevemente la historia de la fonética instrumental, y se hace un recorrido por los instrumentos de los que se ha valido a lo largo del tiempo para el estudio del habla. Finalmente, se mencionan los alfabetos fonéticos que son utilizados actualmente para estudios del español, ya que son otro tipo de herramientas para los estudios fonéticos. Del mismo modo, se mencionan algunos alfabetos fonéticos computacionales que se han elaborado a partir de los otros, y que son utilizados por proyectos de las tecnologías del habla.

La lingüística computacional es un área científica en la que trabajan conjuntamente dos disciplinas: la lingüística y la computación. Esta área se ha interesado en desarrollar programas para el modelado y estudio de la lengua; por esta razón, se dedica a la construcción de sistemas computacionales lingüísticos, los cuales tiene como objetivo el procesamiento de la estructura lingüística humana.

Esta ciencia comenzó a desarrollarse en la década de los cincuentas, tras la necesidad de contar con máquinas que tradujeran automáticamente textos de diferentes idiomas. Más adelante, se comenzaron a construir programas de traducción y se observó que la lingüística era sumamente importante para su elaboración, pues primero era necesario conocer y comprender los diferentes niveles de la lengua de cada idioma (su sintaxis, semántica,

pragmática, etc.), para así poder hacer traducciones fieles. A más de cinco décadas de aquellos comienzos, hoy se puede contar con diversas herramientas lingüísticas computacionales, como los traductores automáticos para diversos idiomas, los correctores de estilo, los sistemas de reconocimiento de habla para las computadoras y los teléfonos celulares, etc. (Llisterri 2003a).

Actualmente, la lingüística computacional se divide en tres ramas de estudio: el procesamiento de lenguaje natural, el procesamiento de textos y las tecnologías del habla (Moreno 1998); en cada una de ellas, participan de distinta manera las diversas ramas de la lingüística, así mismo se benefician de las herramientas creadas por la computación. Por un lado, el procesamiento del lenguaje natural tiene por objetivo hacer conversiones automáticas de distintas representaciones lingüísticas. Algunas de sus aplicaciones son los sistemas de traducción automática, de recuperación y extracción de información; en otras palabras, sistemas que resumen el contenido de un texto y que extraen información para crearlo.

Por otro lado, el procesamiento de texto tiene como objetivo la creación de herramientas computacionales que ayuden a facilitar el manejo del contenido de un texto. Las aplicaciones de esta rama, entre otras, son los sistemas de corrección de escritura como los correctores ortográficos, sintácticos y de estilo; los sistemas de análisis textual que sirven para contabilizar datos como la frecuencia de aparición de palabras o construcciones sintácticas; los sistemas para análisis de corpus como los transcriptores lingüísticos, que transcriben para cada palabra de un texto su categoría gramatical, y los analizadores sintácticos. Estas herramientas ayudan a los lingüistas a realizar tareas complejas como analizar, contabilizar y tener un control sobre grandes bases de datos (Moreno 1998).

Finalmente, las tecnologías del habla tienen por objetivo obtener procedimientos con los que se pueda lograr la comunicación entre personas y máquinas por medio del habla (Llisterri 1991). Para ello, esta rama se dedica a la construcción de sistemas que generan y reconocen el habla, llamados de síntesis y reconocimiento. A continuación se profundiza en las tecnologías del habla, ya que es el área de estudio donde se inserta este trabajo.

2.1. Las tecnologías del habla

Las tecnologías del habla tienen por objetivo "alcanzar la interacción con los sistemas informáticos mediante la voz, con el fin de que las personas podamos hacer uso eficiente de éstos, eliminando las restricciones que, en muchas ocasiones, nos imponen las herramientas actuales" (Llisterri 2003a:249). Para lograr su objetivo, las tecnologías del habla se dedican a desarrollar los sistemas de síntesis y reconocimiento del habla.

Desde hace ya muchos años, el hombre ha trabajado en la creación de sistemas que puedan hablar. Llisterri menciona que en los setentas las tecnologías del habla habían tenido grandes avances en sus estudios. Por una parte, se pudo tener control de la producción del habla por medio de una computadora; por la otra, también se crearon técnicas digitales para manipular las señales acústicas del habla (Llisterri 2003a). Conforme pasó el tiempo, se fueron perfeccionando ambos sistemas de síntesis y reconocimiento de manera que, en la década de los ochentas, los sistemas de reconocimiento pudieron captar el habla fluida, es decir, reconocieron un enunciado sin la necesidad de hacer pausas entre cada palabra. Para la década de los noventas, las tecnologías del habla ya tenían un gran desarrollo en sus sistemas; los sintetizadores se habían comercializado y el vocabulario de los reconocedores se había ampliado.

En la actualidad, las tecnologías del habla tiene sistemas capaces de hablar de una manera más fluida y de reconocer el habla espontánea con más eficacia, sin importar el número de palabras que se pronuncien durante la comunicación. En varios países hay compañías que se dedican al desarrollo de estos sistemas. En Europa hay compañías como *Loquendo*[1] y *Acapela*[2] que, entre otros, han construido sistemas de síntesis capaces de emitir cualquier texto con un habla casi natural. En América existe *Bell Labs*[3] que también se dedica al desarrollo de sistemas de síntesis para uso de las tecnologías. En los sistemas de reconocimiento destaca el sistema *Sphinx* desarrollado por la *Universidad de Carngie Mellon*. Este sistema ha sido utilizado para crear algunos reconocedores del *Proyecto*

[1] http://www.loquendo.com
[2] http://www.acapela-group.com
[3] http://www.bell-labs.com

DIME. Otras compañías que desarrollan sistemas de reconocimiento son *IBM* y *Macintosh* para los sistemas de dictado en las computadoras.

A pesar de las ventajas que tienen actualmente los sistemas de reconocimiento y síntesis, existen algunas limitaciones en su funcionamiento. Una de ellas es que el sistema únicamente reconoce y reproduce palabras u oraciones que tiene en su dominio,[4] motivo por el cual la comunicación está orientada a un solo tema. Otra dificultad es "la falta de naturalidad de los sistemas que permiten que un sistema informático tenga una salida vocal, en las dificultades para reconocer el habla de cualquier persona que se exprese de forma corriente y hablando de cualquier tema en un entorno real" (Llisterri 2003a:253). Aún así, dicha tecnología se ha puesto en uso para productos comerciales como los programas de dictado para la computadora, los de pronunciación para la enseñanza de lenguas extrajeras, la marcación de voz en la telefonía y la implementación de servicio vía telefónica de algunas empresas.

En los siguientes dos apartados se describe la arquitectura de los sistemas de síntesis y reconocimiento, y los procesos que se llevan a cabo para su funcionamiento. Esto, con el fin de observar de qué manera colabora la lingüística –específicamente la fonética– en la construcción de dichos sistemas.

2.1.1. La síntesis del habla

La síntesis del habla se encarga de recrear artificialmente el habla humana; para ello, elabora sistemas computacionales llamados sintetizadores. El funcionamiento de los sintetizadores se basa en la conversión de un texto escrito en habla, por medio de "diversos módulos con información fonética y lingüística que realizan una serie de transformaciones a la cadena inicial de caracteres ortográficos hasta convertirla en una onda sonora y constituye tanto una aplicación informativa útil para acceder mediante el habla a información almacenada en forma escrita como una herramienta de investigación en

[4] Por *dominio* me refiero a las palabras que le han sido programadas a los sistemas y que, por lo tanto, conoce y puede reproducirlas o, bien, reconocerlas.

fonética que permite validar las hipótesis sobre producción y percepción del habla realizadas desde diversos marcos teóricos" (Llisterri *et al.* 1999:459).

A partir de la búsqueda por obtener inteligibilidad y naturalidad en los sintetizadores, se han creado varias formas para generar el habla artificial; entre ellas, las dos principales son la síntesis por formantes y la síntesis por concatenación. La síntesis por formantes toma parámetros del habla natural –como la frecuencia fundamental– y a partir de ello se diseña reglas para crear una onda sonora artificial. La síntesis por concatenación consiste en el encadenamiento de segmentos del habla previamente grabados. Existen diferentes tipos de unidades para concatenar, entre los que destacan la síntesis por concatenación de unidades y la síntesis por concatenación de difonos. La primera consiste en la unión de unidades lingüísticas como alófonos, sílabas, frases, etc., obtenidas a partir de un corpus oral. La segunda se caracteriza por hacer el encadenamiento de unidades que abarcan de la mitad de un primer sonido hasta la mitad del que le precede. A este tipo de unidades se les conoce como difonos, los cuales también son obtenidos a partir de corpus orales. El propósito de este tipo de síntesis es conservar la naturalidad en la transición entre dos sonidos (Llisterri 2003a).

El procedimiento que sigue un sintetizador es largo, ya que el texto escrito pasa por diferentes módulos hasta generar la señala sonora para su emisión. En la Figura 1 se muestra el esquema de la arquitectura de un sintetizador.

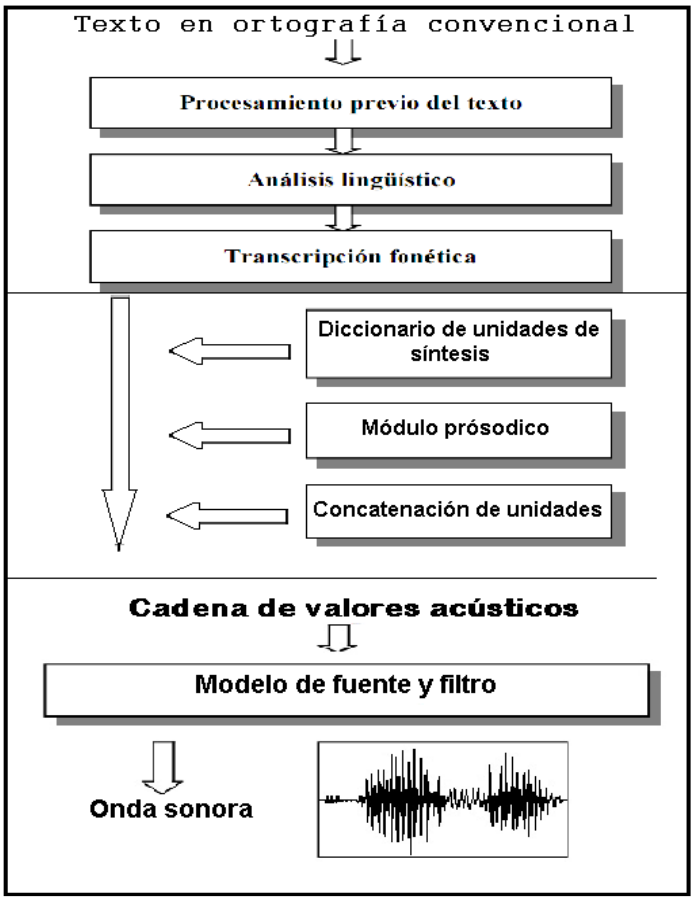

Figura 1. Arquitectura de un sintetizador de habla (Llisterri 2003a:14).

En este esquema se puede observar que el proceso de conversión de texto a habla se divide en tres fases y, a su vez, cada una de ellas se divide en distintos módulos. En la primera, de arriba abajo, se hace un procesamiento previo del texto y se cambia a una ortografía convencional. Posteriormente, el texto preprocesado se transcribe fonéticamente; en esta fase es importante que el texto que la máquina va a enunciar esté correctamente escrito, es decir, sin palabras que pueden causar ambigüedades como las siglas, los números, las abreviaturas, etc. Por ejemplo, si en una oración se encuentran escritas las siglas *PRD* es necesario escribirlas en el texto a procesar, tal como se pronuncian *perredé*. Esto, con el propósito de evitar problemas durante la conversión del texto a su representación fonética.

En la segunda fase, aparece el diccionario de unidades, el cual tiene almacenadas palabras con su correcta pronunciación. Cada palabra del texto escrito pasa por el diccionario de unidades, el cual determina su pronunciación fonética y remplaza el texto ortográfico por la

transcripción fonética de la palabra. Una vez obtenida la pronunciación de cada una, el texto pasa a un módulo prosódico, donde es dividido en distintas unidades: oraciones, frases, proposiciones, etc. Posteriormente, estas unidades pasan al módulo de concatenación. Finalmente, en la tercera fase toma el texto transcrito y concatenado para transformarlo en señal sonora y emitir el habla sintética.

2.1.2. El reconocimiento del habla

El reconocimiento del habla, al igual que la síntesis, es un área interesada en la comunicación entre humanos y máquinas, motivo por el cual se dedica a la construcción de sistemas reconocimiento del habla. De cierta manera, el funcionamiento de estos sistemas son inversos a los de síntesis, pues su tarea es reconocer la señal acústica del habla humana y transformarla en texto, o bien, a su representación simbólica lingüística. De esta manera, las máquinas pueden conocer las necesidades o peticiones del usuario y, del mismo modo, entablar un diálogo con él.

En sus inicios, los sistemas de reconocimiento estaban bastante limitados por diversos factores; por ejemplo, sólo reconocían palabras aisladas y emitidas por un sólo interlocutor. En la actualidad, la interacción humano-máquina ha tenido grandes avances, ya que los sistemas de reconocimiento son capaces de captar el habla de distintos usuarios y pueden reconocer un gran número de oraciones. En la Figura 2 se muestra un esquema de la arquitectura de un sistema de reconocimiento de habla.

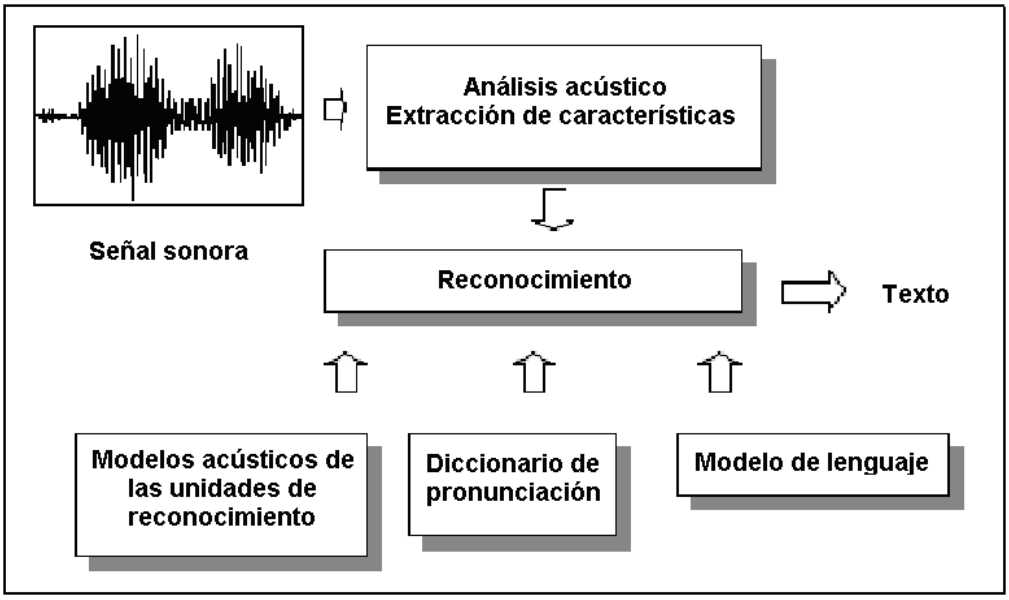

Figura 2. Arquitectura de un reconocedor de habla (Llisterri 2003b:15).

Como se puede observar, el reconocimiento se basa en 6 módulos. El primero es el de la señal acústica, ubicado en la esquina superior izquierda del esquema. En este módulo el sistema capta la señal sonora, la reconoce como habla y la envía al módulo de análisis acústico (ubicado a la derecha del anterior), donde selecciona las características acústicas que posee la señal como los formantes, la imagen espectral, el tono y la entonación. Posteriormente, esas características son enviadas al descodificador de mensajes o módulo de reconocimiento (que está en el centro del esquema), el cual desentraña el mensaje junto con los módulos de los modelos de lenguaje, los modelos acústicos y el diccionario de pronunciación (ubicados en la parte inferior del esquema).

Cada uno de los módulos anteriores desempeña un trabajo diferente. En primer lugar, los modelos acústicos contienen diversas palabras transcritas fonéticamente, con el propósito de hacer una comparación entre los rasgos acústicos de las palabras almacenadas y los de la señal sonara emitida por el hablante. En segundo lugar, el módulo del diccionario de pronunciación contiene las palabras que fueron utilizadas para crear los modelos acústicos, con su pronunciación correcta y las posibles pronunciaciones por parte del usuario, con el fin de indicar la secuencia exacta de los sonidos que debe interpretar el reconocedor. Finalmente, en el módulo de modelos del lenguaje están las palabras que pueden preceder a las palabras utilizadas en el diccionario de pronunciación (Tapias 2002).

El reconocimiento del habla se lleva a cabo cuando el sistema capta una señal sonora y ésta se alinea con su representación fonética en una "secuencia de símbolos fonéticos formada a partir de la concatenación de una o más pronunciaciones previamente almacenadas en el diccionario de pronunciación. Cuando se logra realizar esta alineación, se resuelven de manera simultánea el problema de segmentar la señal en una secuencia de palabras y el reconocimiento de la voz propiamente" (Pineda 2008:20).

Pineda (2008:21) menciona que para que los recursos destinados a la construcción de reconocedores sean de utilidad "deben recopilarse de manera empírica y analizarse a partir de una base fonética sólida para el lenguaje. Dado el estado de la tecnología, para crear reconocedores de calidad se requiere, además de conocer las herramientas y algoritmos de computo mencionados, contar con un recuso lingüístico con una buena base empírica".

2.2. La fonética en las tecnologías del habla

En las tecnologías del habla, la fonética se encarga de describir los sonidos, de asignarles una representación simbólica y de establecer "un conjunto de reglas que determinan las características acústicas de cada unidad y la forma de concatenarlas" (Llisterri 2003a:264). Su participación en las tecnologías del habla es de suma importancia, pues para que se logren mayores avances tecnológicos en la construcción de sistemas de síntesis y reconocimiento, se requiere tener un nivel superior de conocimientos fonéticos, con los cuales se puede seguir desarrollando y perfeccionando dichos sistemas (Aguilar *et al.* 1994).

La fonética participa en varias fases de la elaboración de los sistemas de reconocimiento. En primer lugar, es necesario recopilar corpus orales (de los cuales se hablará más detalladamente en §3.2.), para ello se necesita de la colaboración de la dialectología, la cual determinará los diversos dialectos de una lengua con el fin de obtener registros de cada una de ellas para compilarlos dentro del corpus; de esta manera, el reconocedor poseerá amplias muestras y no tendrá dificultades para reconocer distintos acentos del habla (Llisterri

2003b). Si por el contrario, el reconocedor es para una comunidad en particular, basta con recopilar muestras del dialecto para el cual va a ser destinado.

Una vez recopilado el corpus, se transcribe fonética y ortográficamente y se alinea temporalmente con su imagen acústica. En esta fase, el fonetista participa en la segmentación la señal sonora y en la transcripción de cada uno de los sonidos con base en un alfabeto fonético computacional. Cabe mencionar que muchas veces no se cuenta con gente especializada para llevar a cabo esta labor, y es allí donde este trabajo de tesis tendrá uso, pues contiene la información necesaria para aprender a segmentar la señal sonora y hacer la transcripción de los alófonos.

Cuando el corpus ya ha sido transcrito, se comienzan a crear los distintos módulos del reconocedor. La fonética participa directamente en dos de ellos: en los modelos acústicos y en el diccionario de pronunciación. En el primero, se extraen del corpus los distintos rasgos acústicos que se obtuvieron y se representan con símbolos que forman parte de un alfabeto fonético computacional. En el segundo, se almacenan las palabras del corpus transcritas fonéticamente con su pronunciación correcta y las distintas pronunciaciones que el hablante puede realizar.

En la construcción de sistemas de síntesis, la fonética también tiene amplias labores como:

1) La implementación de reglas de transcripción fonética automática que establecen las correspondencias entre grafías y alófonos, la silabicación y la acentuación, complementadas por diccionarios de pronunciación para el tratamiento de excepciones.
2) La elaboración de distintos modelos para los sintetizadores, por ejemplo:
    - Modelos de duración segmental, que consideren los diversos factores que influyen en la duración y basados en datos procedentes de corpus representativos.
    - Modelos de intensidad segmental que, igualmente, consideren los factores que inciden en la intensidad y procedan de corpus representativos.
    - Modelos de asignación de pausas, que contemplen tanto las marcadas mediante signos de puntuación como las no marcadas.
    - Modelos de entonación que permiten generar una curva melódica natural, teniendo en cuenta factores fonéticos, sintácticos, semánticos y pragmáticos (Llisterri 2007:25).

Como podemos observar, la fonética tiene una participación fundamental en las aplicaciones de las tecnologías del habla, ya que brinda diversas herramientas para el tratamiento del habla, tales como alfabetos fonéticos o datos acústicos para su registro, sin las cuales no seria posible la construcción de los sistemas de síntesis y reconocimiento (Llisterri 1991). Una de las ramas de la fonética que participa de manera relevante en las tecnologías del habla es la fonética instrumental, la cual basa sus investigaciones en los resultados obtenidos mediante instrumentos que registran el habla y permiten hacer un análisis acústico de ella.

2.2.1. La fonética instrumental

En los últimos años, la fonética ha recurrido a varios recursos para el estudio de la lengua. De la misma manera, otras disciplinas han necesitado de ella para llevar a cabo su trabajo, como es caso de la computación y de la informática, entre otras.

La fonética tiene varias ramas de estudio, una de ellas es la fonética instrumental, llamada así porque basa sus estudios en distintos instrumentos para la medición y el análisis del registro acústico del habla (Solé 1985). Esta rama de la fonética ha crecido junto con la tecnología, pues ésta le ha brindado instrumentos que le han permitido estudiar de manera más exacta los fenómenos de la lengua, favoreciendo la creación de nuevas hipótesis y teorías (Llisterri 1991). Otras disciplinas también se han beneficiado de este avance, como las tecnologías del habla y sus aplicaciones.

La fonética instrumental no es necesariamente una disciplina joven y ha pasado por un largo proceso para obtener los instrumentos de los que goza hoy en día. A continuación se hace una breve exposición de algunos de los instrumentos que se han desarrollado a lo largo de la historia.

2.2.1.1. Antecedentes

Los primeros instrumentos que se utilizaron para el estudio de la fonética fueron poco sofisticados, pero marcaron un gran comienzo para las tecnologías. Uno de los primeros instrumentos fue el palatógrafo. Este instrumento constaba de un paladar artificial que, cuando se utilizaba, se cubría con una sustancia especial. Posteriormente, el hablante metía el paladar dentro de su boca y cuando pronunciaba cualquier sonido, diferentes partes del paladar se quedaban sin la sustancia, delimitando la zona en la que la lengua tenía contacto con el paladar. Este instrumento únicamente brindaba datos sobre la articulación de la cavidad oral, sin determinar la forma de la boca durante la pronunciación. En la década de los ochentas, aun se seguía utilizando el método palatográfico con el mismo fin, pero se dejó de usar el paladar artificial y la sustancia se colocaba en el paladar del hablante

Posteriormente, nació el método plastográfico, que consistía –al igual que en el palatógrafo− en introducir un paladar artificial en la boca. Este paladar estaba formado por hilos de estaño, los cuales se doblaban según la forma de la lengua al emitir los sonidos. Este método aportaba registros sobre la articulación de la cavidad oral, brindaba imágenes de la articulación de la cavidad bucal y daba información más completa que la obtenida con el palatógrafo (Gil 1999).

Otros instrumentos para el estudio del habla fueron aquellos que buscaban la manera de medir la onda sonora, como el quimógrafo, inventado en 1847 por Karl Ludwing. Este era un aparato cilíndrico que medía el movimiento muscular, al mismo tiempo que lo grababa con un estilete apoyado en papel ahumado. Actualmente, el quimógrafo usa mecanismos eléctricos que miden el flujo del aire y los movimientos de los órganos, aunque, en gran parte, ha sido desplazado por el polígrafo. Tiempo después se crearon los resonadores de Helmboltz, que se inventaron por el científico de quien llevan su nombre. El objetivo de éstos era recrear los sonidos de las vocales por medio de resonadores naturales (Gil 1999). Los resonadores eran unos objetos esféricos con dos orificios en los polos, uno de ellos de gran tamaño, y el otro pequeño, parecido a una boquilla de embudo, adaptado para introducirlo en el oído (Helmholtz 1877/1954). El uso de estos aparatos consistía en emitir

un sonido por el orificio superior; mientras que por el otro se percibía su resonancia. En la Figura 3 se puede observa la imagen de un resonador.

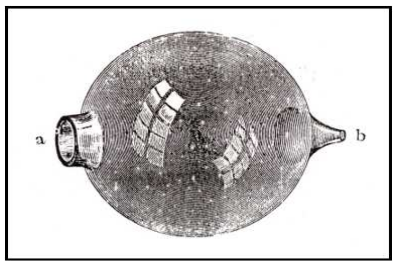

Figura 3. Imagen de un resonador (Helmholtz 1877/1954:43).

Entre los aparatos actuales y sofisticados se encuentran el oscilógrafo y el espectrógrafo. El oscilógrafo es un instrumento electrónico que capta las ondas sonoras y las reproduce gráficamente en una pantalla; en ella las ondas aparecen como una línea luminosa que se desplazan vertical y horizontalmente en el tiempo. De esta menara, se obtiene una descripción del período y la amplitud de la onda sonora (Gil 1999).

El espectrógrafo fue desarrollado en "los laboratorios americanos de Bell Telephone Co. durante el decenio 1930-1940, el principal objetivo de sus diseñadores fue encontrar un método para volver visible el habla de forma que los sordos pudieran leerla y vieran, así, aumentadas sus posibilidades de comunicación" (Gil 1999:53). Con este instrumento se puede analizar la onda compleja, ya que la descompone en sus distintos armónicos; también, proporciona "información sobre la intensidad glotal de los sonidos y sobre la amplitud de cada uno de sus componentes" (Gil 1999:57). El espectrógrafo consta de tres módulos: el primero es un sistema de grabación, el segundo un filtro de frecuencias y el tercero una aguja que dibuja la onda sonora sobre papel.

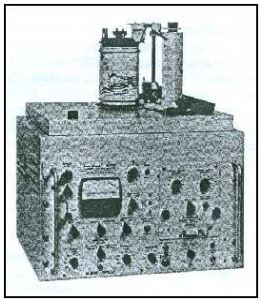

Figura 4. Espectrógrafo (Gil 1990:53).

Para utilizar este aparato es necesario contar con una grabación, la cual puede hacerse desde el micrófono del espectrógrafo, o bien, tener hecha una de manera previa. Primero, esa grabación pasa por el módulo de filtrado, el cual selecciona algunas bandas de frecuencia, dependiendo de los márgenes del filtro, para establecer la frecuencia fundamental y convertirla en corriente eléctrica. Posteriormente, la corriente eléctrica es transmitida a la aguja y ésta, a su vez, transcribe en el papel la frecuencia fundamental y los armónicos de las ondas complejas.

Actualmente, tanto el oscilograma como el espectrograma han dejado de ser aparatos y se han consolidado como programas computacionales, con los que se puede trabajar de manera simultánea y son de fácil acceso. Además, cuentan con una banda de tiempo, que "permite presentar la trayectoria de los formantes y por tanto obtener información sobre las transiciones de un sonido a otro" (Llisterri *et al*. 1999:453). Cualquier persona puede instalar en su computadora un programa de análisis de habla que cuente con oscilogramas y espectrogramas. Uno de ellos es el *Speech View*, creado por el *Center for Spoken Lenguaje Undestanding* (CSLU) del Oregon Graduate Institute (OGI).[5] En el Capítulo 3 se hablará acerca del uso de esta herramienta, debido a que es la que se utilizó para las imágenes espectrográficas que ilustran el *Manual*.

2.2.1.2. Alfabetos fonéticos

Los alfabetos fonéticos son otro tipo herramienta que sirven para el estudio de la lengua. Cada lengua posee un gran número de sonidos, los cuales pueden estar o están representados por símbolos que forman parte de alfabetos fonéticos. A lo largo del tiempo ha habido distintas propuestas de alfabetos fonéticos para el español; sin embargo, en la actualidad dos son los más conocidos en el mundo hispánico: el alfabeto de la *Revista de Filología Española* (RFE) y el *International Phonetic Alphabet* (IPA). Cabe mencionar que el primer alfabeto RFE es un alfabeto únicamente para el español y sus variantes, mientras que el IPA documenta un gran número de sonidos de distintas lenguas del mundo.

[5] http://cslu.cse.ogi.edu/

La *Revista de Filología Española* se fundó en 1914. Una de sus tareas fue elaborar un alfabeto fonético que describiera los sonidos del español y de algunas de sus variantes; la elaboración de este alfabeto era necesaria para que se utilizara en estudios lingüísticos llevados a cabo por la misma revista. En 1915 *La Revista* publicó el alfabeto que, hasta la fecha, es conocido bajo el nombre del alfabeto de la *Revista de Filología Española*, o bien, *alfabeto hispanista*. Este alfabeto ha sido lo suficientemente consistente para la transcripción de los sonidos del español, motivo por el cual ha sido utilizado para varios estudios; sin embargo, en los últimos años, el uso del alfabeto IPA ha incrementado de forma importante en los estudios lingüísticos del español (Mota *et al.* 1995).

La Asociación Fonética Internacional (AFI) se fundó en 1986. Uno de sus objetivos fue crear un alfabeto fonético para las diversas lenguas del mundo, el cual fue llamado *International Phonetic Alphabet* (IPA). Con este alfabeto se buscaba la unificación y la consistencia en la representación de los sonidos para las distintas lenguas del mundo; para ello la Asociación tomó las representaciones del alfabeto románico. En 1999 la AFI publicó un libro llamado *Handbook of the International Phonetic Association. A guide to the use of the International Phonetic Alphabet*; en éste se justifica haber tomado el alfabeto romano para las representaciones de su alfabeto: "The IPA is based on the Roman alphabet, which has the advantage of being widely familiar, but also includes letters and additional symbols from variety sources. These additions are necessary because de variety of the sound in languages is much grater than the number of letters in the Roman alphabet" (IPA 1993:3). En este libro, también da una serie de usos para el alfabeto aplicables a las tecnologías emergentes: "The IPA can be used for many purposes. For instances, it can be used as a way to show pronunciation in a dictionary, to record language in linguistic fieldwork, to form the basis of a writing system for a language, or annotate acoustic and other displays in the analysis of speech. For all these tasks it is necessary to have a generally agreed set of symbols for designating sounds unambiguously, and the IPA aims to fulfil this role" (IPA 1993:3). Hasta la fecha, la Asociación Fonética Internacional se ha preocupado por los avances de su alfabeto y lo ha llevado al margen de las tecnologías, de manera que ahora, se puede trabajar con él en las computadoras, pues ha hecho un sistema de fuentes

fonéticas, las cuales pueden ser descargadas gratuitamente de la red electrónica mundial. Estas fuentes también cuentan con códigos numéricos decimales y hexadecimales.

En las investigaciones de la lingüística computacional se utiliza otro tipo alfabetos llamados alfabetos fonéticos computacionales. Existen varios alfabetos como el *Speech Assessment Methods Phonetic Alphabet* (SAMPA), el cual fue creado por el proyecto ESPRIT para las lenguas europeas. Posteriormente se creó el alfabeto *X-SAMPA*, que fue una ampliación de *SAMPA* para los elementos prosódicos del habla.

Posteriormente, se creó el alfabeto *Worldbet* por James Hieronymus (1994), para su uso en los proyectos del habla de los *Laboratorios Bell*. La creación de *Worldbet* fue motivada a partir de que los alfabetos como *SAMPA* y *X-SAMPA* estaban hechos para lenguas europeas, y no abarcaban representaciones para los sonidos de otras. Con *Worldbet*, Hieronymus buscaba representar todas las lenguas del mundo; para ello tomó símbolos del código ASCII, de manera que las representaciones fonéticas fueran parecidas a las del IPA. También, incluyó otros símbolos que resultaban consistentes para la transcripción de los sonidos de las distintas lenguas; con el propósito de que su alfabeto fuera lo suficientemente robusto para poder hacer transcripciones en varias lenguas (Hieronymus 1994).

Otro alfabeto fonético computacional es *OGIbet* (Oregon Graduate Institute Alfabet), creado por el Oregon Graduate Institute, con el propósito de transcribir fonéticamente los corpus orales, para estudios lingüísticos del inglés desarrollados por el *Center for Spoken Lenguaje Undestanding* (CSLU) (Lander 1997).

En México se creó un alfabeto fonético computacional llamado *Mexbet* (Cuétara 2004). Este alfabeto documenta los alófonos del español de la ciudad de México y se creó para utilizarse en proyectos de tecnologías del habla en nuestro país. En el Capítulo 3 se hablará más detalladamente sobre el alfabeto *Mexbet*.

En este capítulo se ha hecho una revisión de las diferentes ramas de la lingüística computacional. De ellas, se profundizó en las tecnologías del habla, mencionando dos de sus aplicaciones que son los sistemas de síntesis y reconocimiento del habla. Esto, con el propósito de destacar los trabajos en los que colabora la fonética para la construcción de dichos sistemas. Posteriormente, se hizo mención a la fonética instrumental, abordando los diversos instrumentos que ha utilizado para los estudios del habla. Finalmente, se describieron los principales alfabetos fonéticos para el español y, a partir de ellos, se mencionaron algunos alfabetos fonéticos computacionales. En el siguiente capítulo se dará un panorama al *Proyecto DIME*, ya que es en el que se inserta este trabajo de tesis. Después, se hablará sobre los procesos de recolección y transcripción de corpus en el marco del *Proyecto*, con el objetivo de exponer las labores fonéticas dentro del mismo.

3. Las tecnologías del habla en México: El *Proyecto DIME*

En este capítulo se presenta el *Proyecto DIME*, se habla sobre los recursos lingüísticos y computacionales que ha desarrollado para la construcción de sistemas de reconocimiento. De estos recursos, se describe con más detenimiento el *Corpus DIMEx100*, por ser éste el corpus con el que se ilustra el *Manual* que se propone en esta tesis. Posteriormente, se describen los procesos de segmentación y transcripción fonética computacionales, con el propósito de precisar las tareas en las que tendrá uso el *Manual*. Para finalizar, se presentan las herramientas que ha utilizado el *Proyecto DIME* para la transcripción de sus corpus orales, tales como el alfabeto fonético *Mexbet* y el programa *Speech View*.

Dentro de los proyectos de las tecnologías del habla en México, se resalta la participación del *Proyecto de Diálogos Inteligentes Multimodales en Español* (DIME). Este proyecto comenzó a desarrollarse en el Departamento de Ciencias de la Computación del Instituto de Matemáticas Aplicadas y en Sistemas (IIMAS) de la Universidad Nacional Autónoma de México, en 1998, bajo la dirección del Dr. Luis A. Pineda Cortés, quien coordina un equipo de investigadores, técnicos y estudiantes (Pineda 2008). El objetivo principal de *DIME* fue desarrollar sistemas computacionales conversacionales para el español de México; es decir, desarrollar sistemas tecnológicos del habla para lograr la interacción entre humanos y máquinas. Para ello, se fijaron tres tareas iniciales: "The first consist in the compilation, transcription and tagging of a multimodal *corpus* in a design domain; the second is the development of a Spanish speech recognition system specialized in the language employed in the application domain; the third is a definition of a Spanish grammar and parser for Spanish comprehensive enough to interpret the lexicon and grammar observed in the DIME corpus" (Pineda *et al*. 2002:166); que posteriormente se extendieron para completar todos los módulos necesarios en los sistemas de diálogo multimodales.

Como se puede notar, desde sus inicios el *Proyecto DIME* ha llevado a cabo diferentes investigaciones enfocadas al desarrollo de los sistemas conversacionales. Estas

investigaciones han sido multidisciplinarias, ya que en ellas han intervenido diferentes campos de estudio. Por ejemplo, dentro de la lingüística resaltan diversas investigaciones fonéticas, fonológicas, gramaticales, semánticas y pragmáticas; en el área de la computación destaca la construcción de sistemas de reconocimiento del habla. En algunas de las investigaciones, el *Proyecto* ha trabajado en conjunto con otros grupos nacionales, como el Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE); e internacionales, como el *Institute for Human and Machine Cognition* de la Universidad de Rochester en el Estado de Nueva York, en los Estados Unidos, y el Departamento de Filología Española, de la Universidad Autónoma de Barcelona, en España. La colaboración con estos y otros grupos ha servido para que las investigaciones del *Proyecto* se enriquezcan con los avances de otros desarrollos científicos similares.

Actualmente, el *Proyecto DIME* cuenta con diversos recursos lingüísticos y computacionales; por ejemplo, corpus orales y diversos sistemas de reconocimiento. Estos recursos, a parte de servir como base para diversas investigaciones, han sido implementados en el *Proyecto Golem*; cuyo objetivo ha sido el desarrollo de un robot capaz de conversar en el español hablado en México, entre otras aplicaciones.

Los corpus orales que ha compilado el *Proyecto* son: El *Corpus DIME* y el *Corpus DIMEx100*. Éstos han servido para hacer investigaciones lingüísticas y, de manera específica, para la elaboración de los diccionarios de pronunciación y los modelos acústicos de los sistemas de reconocimiento.

El primer corpus que se recolectó fue el *Corpus DIME*, el cual se hizo con el objetivo de obtener diálogos en español de México para el análisis de los actos del habla y para elaborar los modelos acústicos de los sistemas de reconocimiento (Villaseñor *et al.* 2001). Este primer corpus se recopiló con base en el protocolo Mago de OZ, el cual propone un escenario en el que se tiene a una persona (el mago) que interpreta el papel de un sistema y a diferentes usuarios, quienes, uno a uno, interactúan oralmente con el sistema para resolver determinada tarea con ayuda del mago. La persona que juega el papel del sistema es previamente instruida para actuar como un sistema computacional, no como humano; es

decir, debe limitar su expresividad y puede ayudar al usuario sólo cuando esté ignorando o transgrediendo las reglas, entre otras características.

De esta manera, el usuario piensa que realmente está interactuando con un sistema inteligente (Villaseñor *et al.* 2001). Este protocolo permite observar las características del habla espontánea, como el léxico, la sintaxis, etc. De igual forma, permite estudiar la interacción del humano con la máquina.

La tarea que se determinó para la recolección del *Corpus DIME* fue diseñar una cocina; esta tarea se eligió por ser un tema sencillo y común para todos lo usuarios. Además, esta acción generaría varias referencias espaciales y direccionales que posteriormente resultarían útiles para analizar y diseñar un sistema de diálogo hombre-máquina. Para el diseño de la cocina, el sistema mostraba en la pantalla un menú con varios elementos para la decoración y dos planos de la cocina: uno de dos dimensiones y otro de tres; éstos servían para que el usuario viera la disposición del lugar y las medidas de los muebles. Durante la grabación, el usuario interactuaba oralmente con el sistema para elegir, acomodar y mover los muebles a un determinado lugar del plano.

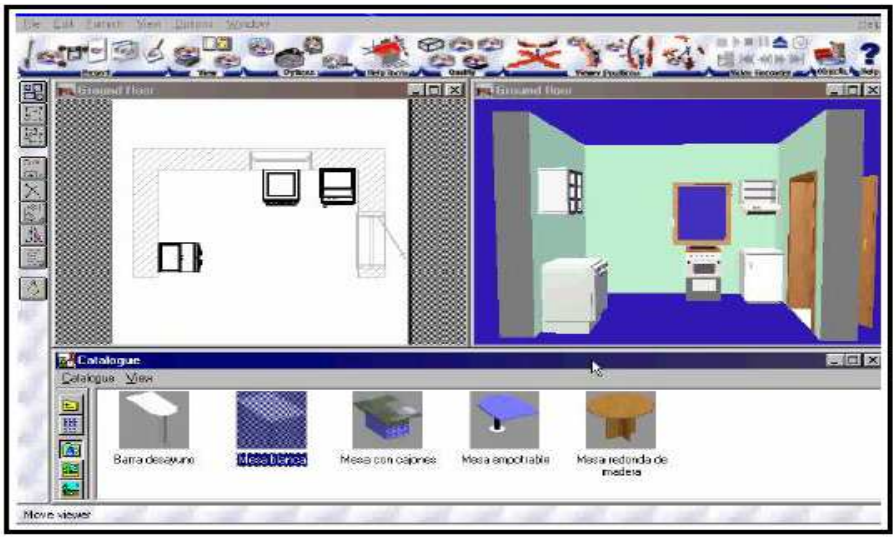

Figura 5. Imagen de la interfaz del programa de diseño de cocinas (Pineda 2008).

El *Corpus DIME* quedó constituido por 31 diálogos y un total de 7:10 horas de grabación. Los diálogos fueron grabados por 16 personas, con una edad promedio de 30 años (Pineda

et al. 2002). La zona geográfica fue tomada al azar, con la finalidad de obtener muestras generales de la lengua. Este corpus ha servido para modelar los actos de habla durante el discurso, los cuales han sido utilizados para el módulo de interpretación del sistema de reconocimiento del habla y para tener una guía del léxico y la gramática española (Villaseñor *et al.* 2001).

3.1. El *Corpus DIMEx100*

Como se mencionó en el apartado anterior, una de las tareas del *Proyecto DIME* fue la construcción de los modelos acústicos y los diccionarios de pronunciación para la programación de sistemas de reconocimiento, con ese fin se recopiló el *Corpus DIMEx100* (Pineda *et al.* 2004, 2010). Como el corpus estaba destinado para esos propósitos, se elaboró conforme a características específicas. Por una parte, el registro del habla no debía de ser espontáneo, sino por el contrario con una pronunciación cuidada y controlada; es decir, el hablante debía pronunciar las palabras claramente, para obtener muestras optimas de cada sonido para los modelos acústicos. Por otra parte, la grabación no debía presentar ruidos externos a la voz para obtener buenas muestras de los alófonos; por ese motivo, fue grabado dentro de una cabina de audio.[6]

Para la recopilación del *Corpus DIMEx100*, en primer lugar, se hizo una recolección de oraciones de la Web, se tomó este medio porque se le consideró un recurso lingüístico amplio, consistente y equilibrado. El resultado de esa recopilación fue el *Corpus230*, el cual quedó conformado por 344,619 oraciones (Pineda *et al.* 2004, 2010). A partir de ese primer corpus, se extrajeron aquellas oraciones que estaban conformadas por un rango de 5 a 15 palabras, quedando con un total de 15000 oraciones, las cuales fueron ordenadas de menor a mayor de acuerdo a su valor de perplejidad. La perplejidad es "a commonly used measure of the goodness of a language model that could be intuitively thought of representing the

[6]"The corpus was recorded in a sound study at CCADET, UNAM, with a Single Diaphragm Studio Condenser Microphone Behringe B-1, a Sound Blaster Audigy Platimun ex (24 bit/96khz/100db SNR) using the Wave Labe program; the sampling format is mono at 16 bits, and the sampling rate is 44.1 khz" (Pineda *et. al* 2004:976).

average number of word choices at every predictive step; the lower the number, the better" (Pineda *et al.* 2010:349). Así, el grupo de menor perplejidad estaba compuesto por las palabras a las que podían precederle un menor número de palabras distintas dentro de una frase, y que contenía un alto grado de información; y el de mayor perplejidad estaba formado por aquellas palabras que podían sucederles un mayor número de palabras. Se tomaron las oraciones que poseían el valor más bajo de perplejidad, quedando con un total de 7000; de esas oraciones, se eliminaron aquellas que contenían palabras en otro idioma, abreviaturas o siglas inusuales, y se conservaron algunas con siglas usuales para facilitar el proceso de lectura (Pineda *et al.* 2004, 2010). Finalmente, se constituyó un corpus de 5010 oraciones, las cuales estaban destinadas para ser leídas y conformar el *Corpus DIMEx100* (Pineda *et al.* 2004, 2010). Estas frases se seleccionaron porque mostraban equilibrio entre la frecuencia de aparición y el número de muestras por cada unidad fonética en relación con el inventario alofónico del alfabeto *Mexbet* (Pineda *et al.* 2004, 2010).

Para la recopilación del corpus oral se utilizaron 100 hablantes, los cuales leyeron 60 oraciones cada uno: 50 distintas y 10 iguales para todos. Estas 10 oraciones iguales fueron grabadas por todos los hablantes con el fin de obtener diferentes muestras de las realizaciones de los alófonos, para su uso en estudios orientados a la identificación y caracterización de hablantes (Pineda *et al.* 2004, 2010). Así, el corpus quedó constituido por 6000 archivos de audio, de los cuales 5000 son diferentes y 1000 idénticos. Las variantes sociolingüísticas que se tomaron en cuenta fueron la edad: con un rango de 16 a 36 años; el nivel de educación: de secundaria en adelante; y el lugar de nacimiento: ciudad de México. La edad promedio de los hablantes fue de 23.82 años; el 87% eran de nivel licenciatura y los demás graduados. El 82% de los hablantes nacieron en la ciudad de México, y 18% eran de otros lugares, pero residían en la ciudad. Cabe mencionar que una parte de los hablantes que constituyeron el corpus fueron investigadores, profesores, estudiantes y trabajadores de la UNAM (Pineda *et al.* 2004, 2010).

Dentro de la variante "género", el corpus quedó balanceado: el 49% de los hablantes fueron del sexo masculino y el 51% femenino. En cuanto a la variable lingüística, se controló que aparecieran todas las unidades fonéticas del español de la ciudad de México y que su

ocurrencia dentro del corpus ascendiera a un número significativo y equilibrado, para que pudiera funcionar como un recurso eficiente para la construcción de sistemas de reconocimiento de habla (Pineda *et al.* 2004, 2010).

Una vez recopilado el corpus, se transcribió ortográfica y fonéticamente; para esta tarea se utilizó el alfabeto fonético computacional *Mexbet* (Cuétara 2004), el cual describe los sonidos del español hablado en la ciudad de México. El corpus se transcribió en cuatro niveles diferentes: T54, T44, T22 y Tp (Pineda *et al.* 2004, 2010). En los primeros tres niveles, la *T* hace referencia a transcripción, y el número (54, 44 y 22) a la cantidad de alófonos con los que cuenta cada nivel. El nivel Tp hace referencia a la transcripción ortográfica por palabra. Los niveles de transcripción T54 y T44 corresponden a una transcripción fonética, mientras que el nivel T22 corresponde a una transcripción fonológica (Pineda *et al.* 2010). En §3.4. se hablará más a fondo sobre el alfabeto *Mexbet* y los niveles de transcripción.

La herramienta que se usó para la segmentación y transcripción del corpus fue el *Speech View* del *CSLU TOOLKIT*, elaborado por el *Center for Spoken Language Understanding* (CSLU) del Oregon Graduate Institute (OGI) (Pineda *et al.* 2010). Esta herramienta computacional permite el registro, estudio y análisis del habla; para ello, cuenta con varias herramientas como oscilogramas y espectrogramas que ofrecen la imagen de la señal acústica, también cuenta con otras ventanas para la transcripción en etiquetas y su alineación con la señal acústica y su imagen.

Para comprobar que el *Corpus DIMEx100* es un recurso lingüístico completo y fiable, tanto para su uso en investigaciones lingüísticas como para la elaboración de los sistemas de reconocimiento, se hicieron varios experimentos que consistieron en la construcción y evaluación de los modelos acústicos de cada nivel de transcripción (Pineda *et al.* 2010). Los datos que se utilizaron para dichos modelos fueron las 5000 oraciones del *Corpus DIMEx100*; no se usaron las 10 oraciones idénticas, grabadas por todos los hablantes. Se usaron los mismos datos para la construcción de los modelos acústicos y los diccionarios de pronunciación, con el fin de poder comparar los resultados de las pruebas (Pineda *et al.*

2010). Los experimentos de reconocimiento y la evaluación de los mismos se llevaron a cabo con el sistema de reconocimiento Sphinx speech recognizer (Sphinx 2006).

A partir de estos experimentos, se obtuvieron diversos resultados. Entre ellos, se comprobó que la información fonética que aporta el corpus permite la construcción de los modelos acústicos y de los modelos de lenguaje en los tres niveles de transcripción. Del mismo modo, dichos niveles de transcripción tiene un porcentaje similar en el desempeño del reconocedor. La mayor contribución de estos resultados fue probar que los modelos acústicos tienen un alto porcentaje en el reconocimiento, a pesar de que los modelos de lenguaje sean básicos. También se encontró que un nivel de transcripción fonética estrecha representa más tiempo para el reconocimiento, debido a que se cuenta con amplias muestras en el modelo acústico[7] (Pineda *et al.* 2010). Con estos resultados se espera que este recurso pueda ser utilizado para estudios fonéticos y para la construcción de diversas herramientas computacionales, como son la construcción de fonetizadores con variación alofónica, aplicados a la elaboración automática de diccionarios fonéticos y a los sistemas de transcripción automática (Pineda *et al.* 2010).

Con los resultados de los experimentos anteriores se construyó el sistema de reconocimiento *DIMEx100*, que se implementó para un nuevo proyecto, llamado *El Proyecto Golem*. Este proyecto consistió en la programación de un robot, que es capaz de interactuar con las personas por medio del habla, la vista y el movimiento, ya que se le implementó un conjunto de sistemas computacionales: un manejador de diálogo, sistemas de reconocimiento y síntesis para la conversación; así como sistemas de visión, navegación y actuadores, para la interacción. La función del robot *Golem* es dar una guía a los visitantes del departamento, exponiendo mediante carteles las investigaciones que se desarrollan en el Departamento de Ciencias de la Computación del IIMAS de la UNAM (Pineda 2008). Para que *Golem* tuviera conocimiento, específicamente en el ámbito conversacional, "se creó un conjunto de modelos de diálogo en los que se representa tanto el esquema conversacional para llevar a cabo las visitas como el contenido conceptual al

[7] En Pineda *et al.* 2010 se reportan otros tipos de resultados y estadísticas, como el número y la frecuencia de alófonos en el corpus, la representación de alófonos en los tres niveles de transcripción, el desempeño del sistema de reconocimiento en cada nivel de transcripción, entre otros.

que es posible referirse durante la conversación" (Pineda 2008:25). De esta manera, el robot interactúa con el usuario por medio de preguntas, respuestas, peticiones y explicaciones. El robot *Golem* es una muestra de la aplicación de las tecnologías del habla en la sociedad y de cómo pueden ser implementadas en la vida cotidiana. Un sistema conversacional, que no debe ser necesariamente un robot, puede ser usado para dar guías en museos, a turistas, en un tutorial, en ventas vía telefónicas, en máquinas, en juegos, etc. (Pineda 2008).

3.2. Corpus orales

Los resultados de los estudios de muchos investigadores de la lengua están basados en la recolección de datos reales; para obtener dichos datos es necesario hacer la recopilación de un corpus. Torruella y Llisterri (1999:52) definen un corpus como "un conjunto homogéneo de muestras de la lengua de cualquier tipo (orales, escritos, literarios, coloquiales, etc.) los cuales se toman como un modelo de un estado o nivel de la lengua predeterminado".

En general, las muestras textuales u orales tomadas del habla común suelen reflejar y brindar datos sobre el fenómeno a estudiar; para que estos datos sean eficientes y muestren en mayor medida resultados fidedignos es necesario hacer grandes recolecciones de muestras, lo que resulta un trabajo duro y de difícil control para contabilizarlas y tener manejo sobre ellas manualmente. Sin embargo, en la actualidad la computación ha brindado valiosas herramientas que de alguna manera facilitan la recopilación y la organización de datos, tales como los programas de almacenamiento para archivos de texto o de audio. Este tipo de herramientas ha ayudado a que el investigador que se enfrenta a documentos con grandes cantidades de datos pueda tener un mejor control de ellos (Torruella y Llisterri 1999).

Los corpus pueden tener diferentes finalidades de estudio dentro de las diversas ramas de la lingüística como la sintaxis, la semántica, la fonología, la fonética, etc. Así mismo, puede tener distintas utilidades, pero su función principal "es establecer la relación entre la teoría

y los datos; el corpus tiene que mostrar a pequeña escala cómo funciona una lengua natural; pero para ello es necesario que esté diseñado correctamente sobre unas bases estadísticas apropiadas que aseguren que el resultado sea efectivamente un modelo de la realidad" (Torruella y Llisterri 1999:45-46). Un corpus suele estar diseñado para un interés específico, aunque puede tener diversos usos. Los corpus para los estudios fonéticos deben ser forzosamente orales, para que reflejen los fenómenos que ocurren durante el habla; por lo tanto, lo óptimo es que sean conversaciones de habla espontánea; sin embargo, también puede ser corpus constituidos por oraciones leídas.

Algunas de las investigaciones de las tecnologías del habla también requieren de compilar corpus orales, pues estos son herramientas de estudio imprescindibles. Para los sistemas de síntesis del habla (conversión texto a habla), los corpus sirven para obtener grandes cantidades de unidades fonéticas, con las cuales se hará la conversión de la representación simbólica a la onda sonora (Rafel y Soler 2003). Para los sistemas de reconocimiento del habla, los corpus sirven para extraer las unidades fonéticas que se utilizarán para los modelos acústicos, pues a partir de la onda sonora se genera una representación (Rafel y Soler 2003); por ello, generalmente, "se basan en palabras aisladas o en frases fonéticamente equilibradas" (Llisterri y Poch 1994:§2.3). También "son esenciales para el entrenamiento y la validación de los sistemas de reconocimiento y de diálogo en entornos de comunicación persona máquina, cuyas aplicaciones se extienden desde la oferta de servicios telefónicos automatizados hasta las ayudas para personas con discapacidades" (Torruella y Llisterri 1999:49); por lo tanto, los corpus que requieren las tecnologías del habla deben ser grabados en formato electrónico para su utilización informática.

El diseño de los corpus debe estar confeccionado conforme a la tarea para la que será destinado, como es el caso del *Corpus DIMEx100*. Este corpus se recopiló para la creación de diccionarios de pronunciación y de modelos acústicos; por lo tanto, se diseñó conforme a parámetros técnicos, lingüísticos y sociales. Dentro de los parámetros técnicos, se aseguro que los archivos de audio no tuvieran ruidos del exterior; es decir, únicamente se debía escuchar la voz del hablante y éste, por su parte, debía leer fluidamente y sin errores las oraciones que se le habían asignado. En los parámetros lingüísticos se cuidó que el corpus

contuviera un número representativo y equilibrado de muestras de los alófonos. Finalmente, en los parámetros sociolingüísticos, se tomó en cuenta la edad, el nivel escolar y lugar de nacimiento de los hablantes.

En las últimas décadas, los corpus orales y textuales en formato electrónico han tenido gran auge en cualquier tipo de estudio, por la facilidad que ofrecen en la organización, contabilización y control de grandes bases de datos. En la fonética las grabaciones en formato pueden ser analizadas con herramientas computacionales especializadas en el registro del habla, lo que ha permitido que se obtengan resultados más exactos de los fenómenos que ocurren en el habla.

3.3. Segmentación y transcripción fonética computacionales de corpus orales

La transcripción de un corpus o de algún documento es una labor que se lleva a cabo en varias disciplinas, como en la psicología, la antropología, la sociología, entre otras (Gibbon *et al.* 1997). En la lingüística, la transcripción de un corpus es un trabajo esencial, ya que por medio de ella se obtienen datos para el estudio de la lengua. Dependiendo de sus fines, la transcripción se puede hacer en diferentes unidades lingüísticas, como sílabas, unidades sintácticas, léxicas, fonológicas, fonéticas, prosódicas, actos de habla, etc. (Rafel y Soler 2003).

En los corpus hechos para sistemas de reconocimiento también se puede optar por algún nivel de la lengua para transcribirlo, aunque esto dependerá de las necesidades en particular. Por un lado, si se quiere un análisis más general se puede optar por la transcripción fonológica y prescindir de la fonética, ya que los fonemas “se configuran como una unidad natural que dota de gran flexibilidad al sistema, y que resulta económica desde el punto de vista del número de unidades; sin embargo, es una unidad abstracta y sometida a variaciones contextuales, lo que origina problemas importantes de concatenación” (Aguilar y Machuca 1995:§2.1). Por el otro, si se quiere que el sistema de

reconocimiento tenga una buena calidad es necesario hacer un análisis fino sobre los rasgos acústicos de los alófonos. Para ello, es recomendable hacer una transcripción fonética, pues como ésta "pretende representar de la manera más exacta de lo posible lo que se ha dicho realmente" (Rafel y Soler 2003:57), posee un repertorio más amplio de unidades alofónicas para el sistema. Esta transcripción, a diferencia de la fonológica, implica un trabajo manual exhaustivo que requiere de muchas horas para llevarlo a cabo; además, se debe contar con un inventario fonético pertinente para el idioma en el cual se va a transcribir, de lo contrario, se puede correr el riesgo de confundir formas y representaciones alofónicas.

La segmentación de un corpus consiste en identificar y marcar los límites de unidades lingüísticas, físicas o auditivas. (Crystal 1980). Ésta, al igual que la transcripción, puede hacerse en diferentes unidades lingüísticas, como fonemas, sílabas, palabras, oraciones etc., dependiendo de sus propósitos. En el caso de un corpus escrito, la segmentación se puede hacer por medio de símbolos que marquen el límite de la unidad lingüística (Autesserre *et al.* 1989). En los corpus orales, la segmentación consiste en la división de un archivo de voz, delimitando temporalmente cada unidad lingüística (Gibbon *et al.* 1997).

Actualmente, la segmentación y la transcripción de corpus orales pueden hacerse por medio de programas computacionales, los cuales cuentan con varias herramientas para el análisis, la segmentación y la transcripción de la señal sonora. Este tipo de herramientas permite la visualización de la onda sonora y la colocación de etiquetas, con el propósito de que mientras se segmenta la señal se pueda ir haciendo su transcripción alofónica en las etiquetas y, a su vez, alinearlas temporalmente con su imagen acústica. A este proceso de transcripción fonética suele llamársele *etiquetado* en la jerga de la lingüística computacional; Barry y Fourcin (1992:2) dan una definición acerca de ese término "The «labelling» of a recorded utterance involves the temporal definition and naming of its parts with reference to the physical signal. These «parts» may be temporarily discrete or over-lapping, and may be defined in acoustic, physiological, phonetic or higher level linguistic terms".

Los procesos de segmentación y de transcripción son muy importantes para la construcción de los sistemas de reconocimiento, ya que por medio de ellos se obtienen distintos módulos para su funcionamiento, como los diccionarios de pronunciación y los modelos acústicos (Barry y Fourcin 1992); por lo tanto, ambos procesos deben estar hechos con mucho cuidado. Por un lado, la segmentación de la imagen acústica debe delimitarse con exactitud para evitar la contaminación entre alófonos; del mismo modo, se debe tener cuidado al alinear la etiqueta con su respectivo segmento visual y temporal (Llisterri y Poch 1994). Por otro lado, la transcripción "responde a la exigencia de materializar o fijar de alguna manera la información lingüística y comunicativa presente en una onda sonora esencialmente transitoria" (Llisterri 1997:§2); por lo que, debe describir correctamente y precisar los rasgos acústicos de la señal sonora, y evitar la confusión de representaciones alofónicas. Ambos procesos se hacen simultáneamente, aunque se comienza con la segmentación. A continuación se detalla paso a paso cómo se segmenta y se transcribe fonéticamente un corpus; cabe mencionar que el proceso de transcripción que se detalla sigue, de alguna manera, la transcripción del *Corpus DIMEx100*.

Generalmente, la primera segmentación que se hace es por oraciones, ya que cuando nos comunicamos, pronunciamos varias palabras de manera continua, hasta que necesitamos hacer una pausa; por cada pausa que hacemos vamos formando oraciones. De la misma forma, cuando se graba a un hablante, éste va formando varias oraciones mientras habla, de manera que la grabación queda compuesta por varias de ellas dentro un mismo audio. En el caso específico de la grabación del *Corpus DIMEx100*, los hablantes leyeron 60 oraciones cada uno y, entre ellas, hacían las pausas correspondientes para que cada una se guardara en un archivo independiente. De esta manera fue más fácil hacer la segmentación de cada oración en unidades lingüísticas más pequeñas.

Una vez que se cuenta con un archivo de audio por cada oración, se toma cada una para segmentarla y transcribirla por alófonos. Para comenzar esta transcripción, primero se hacen dos segmentaciones correspondientes al primer sonido de la oración. La primera se hace al inicio de la onda sonora y la segunda al final del primer sonido; de esta segmentación se forma una primera etiqueta, donde se transcribe el símbolo del alófono que se emitió. Para los siguientes alófonos, únicamente, se hace una segmentación al final de

cada uno, ya que su inicio lo marca el final del alófono que lo antecede. En esta segmentación es muy importante la ayuda del espectrograma y del oscilograma, pues éstos guiarán la delimitación, para hacerla con más exactitud y alinearla con su etiqueta. En la Figura 6 se puede observar la transcripción fonética de la palabra *ello*, en el nivel T54 del alfabeto *Mexbet*, con el programa *Speech View*.

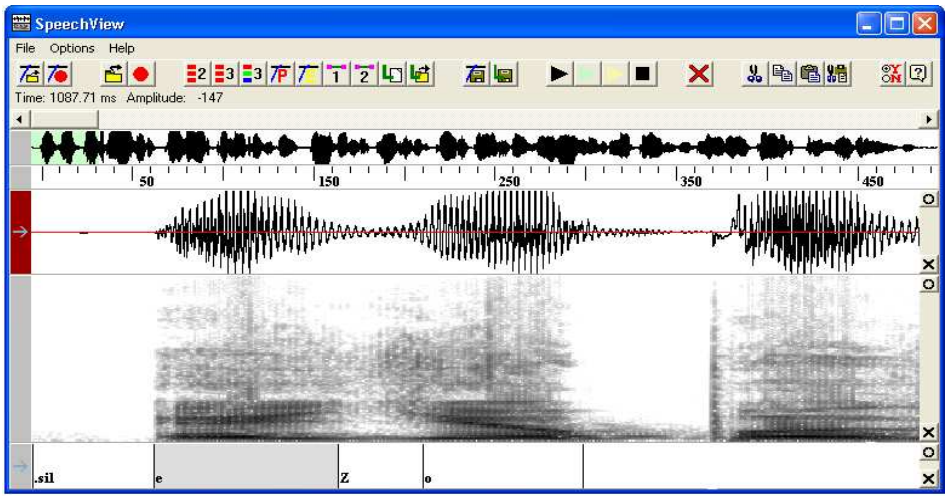

Figura 6. Segmentación y transcripción fonética, en el nivel T54.

El procedimiento anterior se hace con cada uno de los alófonos que componen la oración, de manera que, una vez transcritos todos los alófonos, queda transcrita fonéticamente. De igual forma, se hace con cada oración que forma parte del corpus para que al finalizar se tenga un corpus cuya transcripción esté sincronizada con la señal sonora, y que al ser consultado se pueda disponer de las etiquetas de cada nivel de transcripción junto con la grabación correspondiente (Llisterri 1997). Posteriormente, se hace la transcripción ortográfica de palabras. Para esta transcripción, la segmentación se hace a partir del inicio de la palabra y hasta el final de la misma; para ello, se toma como guía la transcripción alofónica, porque ya aparecen marcados los límites del primero y último alófono. Las etiquetas de los dos niveles deben de estar perfectamente alineadas entre ellas. En Figura 7 se puede observar la segmentación y transcripción de la primera palabra del enunciado: *ello*.

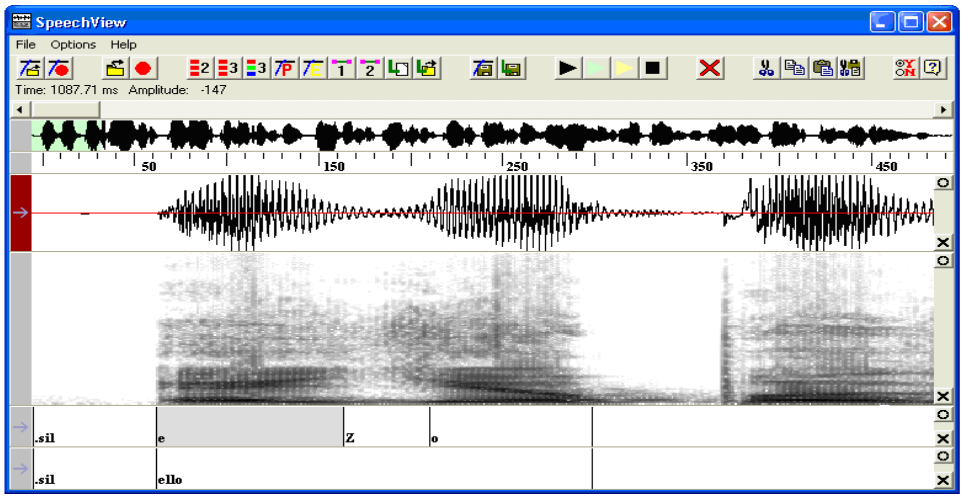

Figura 7. Segmentación y transcripción en T54 y Tp.

En este apartado se describieron los procesos de segmentación y transcripción fonética, siguiendo la transcripción del corpus *DIMEx100*; sin embargo, cabe mencionar que, dependiendo de sus fines, un corpus puede transcribirse en muchos otros niveles, como léxico, sintáctico, silábico, etc. También se señaló que en el caso de los corpus compilados para los sistemas de reconocimiento es más recomendable que se trascriban a nivel fonético, ya que éste representa un análisis acústico fino, lo cual proveerá al sistema de datos más precisos para un mejor reconocimiento. Para este tipo de transcripción se debe de contar un alfabeto fonético adecuado para el idioma en el cual se va a transcribir, de lo contrario, puede haber confusión entre las formas alofónicas. En el caso del *Corpus DIMEx100,* la transcripción se hizo en tres niveles distintos de unidades lingüísticas, basadas en el alfabeto computacional *Mexbet*, el cual se describe a continuación.

3.4. El alfabeto *Mexbet* y los niveles de transcripción (T22, T44, T54)

Como se ha mencionado en capítulos anteriores, el alfabeto *Mexbet* es un alfabeto fonético-computacional para la variante dialectal del español de la ciudad de México. En la creación de este alfabeto han contribuido varios investigadores y a lo largo del tiempo ha tenido varias versiones. Las primeras versiones se hicieron en 1999, 2000 y 2002 en el marco del *Proyecto DIME*. Posteriormente, se hizo una nueva revisión sobre el inventario alofónico y

su representación, con ello se obtuvo una versión final (Cuétara 2004). Esta última versión del alfabeto es la que se utiliza actualmente para la transcripción de los corpus *DIME* y *DIMEx100*. Hasta ahora, el alfabeto *Mexbet* ha probado ser un alfabeto lo suficientemente consistente para describir los sonidos de español de la ciudad de México.

En sus inicios, los alófonos, así como las representaciones gráficas de *Mexbet*, se elaboraron con base en los alfabetos computaciones *OGIbet* y *Worldbet*. Por tal motivo, era necesario hacer algunos ajustes para la versión mexicana. Cuétara menciona que "Mexbet requería de correcciones básicas y puntualizaciones lingüísticas" (Cuétara 2004:63). En la primera versión, el alfabeto *Mexbet* (Uraga 1999) tenía un inventario que constaba de 23 fonemas: 18 consonánticos y 5 vocálicos. Éste presentaba una primera inconsistencia, ya que incluía como fonema al alófono lateral palatal dentro de su inventario consonántico, dejando ver que, al igual que *OGIbet*, *Mexbet* tenía confusión entre los alófono de la vocal cerrada anterior /i/ y el alófono del fonema africado fricativo sonoro palatal /Z/ (Cuétara 2004).

Para la segunda versión del alfabeto *Mexbet* (Uraga y Pineda 2000), se hizo una nueva revisión en la cual se subsanó la confusión de los alófonos palatales; sin embargo, se mantuvo al alófono paravocal velar [w] dentro de las formas fonemáticas. Esta versión del alfabeto contaba con un inventario de 20 fonemas consonánticos y cinco vocálicos. En la tercera versión de *Mexbet* (Pineda y Uraga 2002) hubo un exceso en el inventario de representaciones; además, no se especificaba si las formas eran alofónicas o fonémicas. Por ejemplo, se incluyeron alófonos castellanos, como el alófono interdental alveolar /θ/; otros alófonos dentales; los alófonos aspirados de los fonemas /x/ y /s/, y la variante labiodental fricativa sonora /v/, como señala Cuétara (2004). Después de revisar exhaustivamente las versiones anteriores del alfabeto *Mexbet*, dicho trabajo propuso 4 puntos a elaborar para la nueva versión del alfabeto (Cuétara 2004:67):

1) Especificar el inventario fonológico para un alfabeto fonético computacional del español de México, para subsanar confusiones y marcar límites (como la eliminación de los fonemas hispánicos y la delimitación alofónica de las paravocales).

2) Establecer convenciones certeras y prácticas, para concretar las formas alofónicas a aquellas que tuvieran una presencia recurrente en el habla y que a la vez se pudieran modelar por reglas.
3) Eliminar formas inexistentes —como [v], por ejemplo—, e incluir alófonos con mayor pertinencia y ocurrencia en el habla de México —como la vocal /a/ velarizada y palatalizada, o la consonante /k/ palatalizada.
4) Procurar que este alfabeto fuera lo suficientemente sencillo y claro, para facilitar en la medida de lo posible el etiquetado fonético de corpus orales extensos y el análisis y procesamiento de grandes volúmenes de datos.

Una vez llevados a la práctica los puntos anteriores, la nueva versión de *Mexbet* propuso un inventario de 22 fonemas: 17 consonánticos y 5 vocálicos. Cabe mencionar que la elección de fonemas y alófonos en distribución complementaria para el alfabeto está basada en un estudio estadístico de los alófonos, hecho con base en el *Corpus DIME*. Para la elección de alófonos se tomó en cuenta su aparición dentro del corpus; Cuétara (2004:80) menciona que "Hay alófonos que por su frecuencia de aparición y la consiguiente repercusión en la lengua son candidatos óptimos para ser considerados en los alfabetos fonéticos y para los modelos acústicos y, en general, para todas las aplicaciones de las tecnologías del habla"; por lo tanto, aquellos alófonos que presentaron una alta frecuencia de aparición y podían ser moldeados por reglas, se tomaron como representativos del habla de la ciudad de México; y los alófonos que fueron poco estables o tuvieron una baja frecuencia de aparición, se prefirió mantener, únicamente, su forma prototípica; de esta manera se llegó a un inventario de 37 alófonos.

En cuanto a la representación, Cuétara (2004:68) señala que para la nueva versión de *Mexbet* buscó que los "los niveles de representación fonético y fonológico fueran claros". Para ello, mantuvo algunas representaciones que ya se habían establecido en las versiones anteriores, tal como [V] para el alófono aproximante bilabial y [N] para el nasal velar; y propuso otras nuevas representaciones, en las que procuró seguir "la convención de marcar el fonema base seguido del diacrítico que señala el proceso de asimilación fonética" (Cuétara 2004:69). Para ello, se designó un diacrítico para cada rasgo acústico, así "la persona que etiqueta puede detallar que una vocal abierta está tomando rasgos palatales o velares, y la persona que revisa el etiquetado puede darse cuenta a simple vista de los fenómenos que están sucediendo en la lengua" (Cuétara 2004:69). En la Figura 8 se pueden observar los diacríticos usados en *Mexbet*.

Fenómenos	Símbolos	Ejemplos
Dental	**_ [**	**s_ [**
Palatal	**_ j**	**k_ j**
Velar	**_2**	**a_2**
Cierres de oclusivas	**_c**	**p_c**
Acentos	**_7**	**e_7**

Figura 8. Representaciones de los diacríticos de *Mexbet*.

Como se mencionó al inicio de este apartado, la última versión de *Mexbet* se ha utilizado para la transcripción de los corpus *DIME* y *DIMEx100*. En el caso del *Corpus DIMEx100* se etiquetó en tres niveles: dos fonéticos y uno fonológico, con el objetivo de enriquecer el conocimiento sobre datos fonéticos del español de la ciudad de México (Pineda *et al.* 2010). Dichos niveles han recibido los nombres: T54, T44 y T22; en el que la *T* hace referencia a la palabra *transcripción*, y el número a los alófonos con los que cuenta.

El nivel T54 corresponde a la transcripción más estrecha. Tiene un inventario de 54 alófonos, de los cuales 37 son alófonos consonánticos, 8 son cierres de los alófonos oclusivos y africados (p_c, t_c, k_c, b_c, d_c, g_c, tS_c, dZ_c) y 9 son vocálicos tónicos y átonos (Pineda *et al.* 2010). En la Figura 9 se puede observar el inventario de alófonos consonánticos, y en la Figura 10 el de los vocálicos.

Consonantes	Labiales	Labiodentales	Dentales	Alveolares	Palatales	Velares
Oclusivas sordas	p/p_c		t/t_c		k_j /k_c	k/k_c
Oclusivas sonoras	b/b_c		d/d_c			g/g_c
Africada sorda					tS/tS_c	
Africada sonora					dZ/dZ_c	
Fricativas sordas		f	s_[	s		x
Fricativas sonoras	V		D	z	Z	G
Nasales	m		n_[	n	n~	N
Vibrantes				r (/ r		
Lateral				l		

Figura 9. Inventario de alófonos consonánticos del nivel T54 (Cuétara 2004:69)

Vocales átonas	Anteriores	Central	Posteriores
Paravocales	j		w
Cerradas	i		u
Medias	e		o
Medias abiertas	E		O
Abiertas	a_j	a	a_2

Vocales tónicas	Anteriores	Central	Posteriores
Cerradas	i_7		u_7
Medias	e_7		o_7
Medias abiertas	E_7		O_7
Abiertas	a_j_7	a_7	a_2_7

Figura 10. Inventario de alófonos vocálicos del nivel T54 (Cuétara 2004:69).

El nivel T44 tiene un inventario de 20 alófonos consonánticos más 6 cierres de las consonantes oclusivas y uno de la africada, 7 vocálicos átonos y 5 tónicos. En este nivel aparecen 5 representaciones para las codas silábicas, que son las parejas de consonantes, que se neutralizan en posición final de sílaba (*p/b, t/d, k/g, m/n r/rr*) (Pineda *et al.* 2010). En la Figura 11 se puede observar el inventario de alófonos consonánticos; en la Figura 12 el de las vocales tónicas y átonas, y en la Figura 13 el inventarios de las codas silábicas.

Consonantes	Labiales	Labiodentales	Dentales	Alveolares	Palatales	Velares
Oclusivas sordas	p/p_c		t/t_c			k/k_c
Oclusivas sonoras	b/b_c		d/d_c			g/g_c
Africada sorda					tS/tS_c	
Fricativas sordas		f		s		x
Fricativas sonoras	V		D		Z	G
Nasales	m			n	n~	
Vibrantes				r (/ r		
Lateral				l		

Figura 11. Inventario de alófonos consonánticos del nivel T44 (Pineda *et al.* 2010:365).

Vocales átonas	Anteriores	Central	Posteriores
Paravocales	j		w
Cerradas	i		u
Medias	e		o
Abiertas		a	

Vocales tónicas	Anteriores	Central	Posteriores
Cerradas	i_7		u_7
Medias	e_7		o_7
Abierta		a_7	

Figura 12. Inventario de alófonos vocálicos del nivel T44 (Pineda *et al.* 2010:365).

	Coda silábica
Labiales p/b	-B
Dentales t/d	-D
Velares k/g	-G
Nasales n/m	-N
Vibrantes r(/r	-R

Figura 13. Inventario de codas silábicas del nivel T44 (Pineda *et al.* 2010:365)

Finalmente, el nivel T22 cuenta con los fonemas básicos: 17 consonánticos y 5 vocálicos (Pineda *et al*. 2010). En la Figura 14 se puede ver el inventario de los alófonos consonánticos y en la Figura 15 el de los vocálicos. En la Figura 16 se puede observar la transcripción de la palabra *aumentar*, en cada uno de los niveles antes descritos.

Consonantes	Labiales	Labiodentales	Dentales	Alveolares	Palatales	Velares
Oclusivas sordas	p		t			k
Oclusivas sonoras	b		d			g
Africada sorda					tS	
Fricativas sordas		f		s		x
Fricativa sonora					Z	
Nasales	m			n	n~	
Vibrantes				r (/ r		
Lateral				l		

Figura 14. Inventario de alófonos consonánticos del nivel T22 (Pineda *et al.* 2010:365-366).

Vocales	Anteriores	Central	Posteriores
Cerradas	i		u
Medias	e		o
Abierta		a	

Figura 15. Inventario de alófonos vocálicos del nivel T22 (Pineda *et al.* 2010:366).

Nivel	Transcripción
T54	a_2 w m e n_[t a r(
T44	a w m e –N t a -R
T22	a u m e n t a r(

Figura 16. Ejemplo de transcripción en los tres niveles T54, T44, T22

El *Corpus DIMEx100*, primero se transcribió en el nivel T54; para este nivel se usó una herramienta de segmentación y transcripción automática, la cual hace este trabajo a un nivel básico, utilizando las reglas contextuales para cada alófono (Pineda *et al.* 2004, 2010); es decir, hace la transcripción canónica de la palabra y no la que realmente fue pronunciada. Posteriormente, fonetistas expertos revisaron esta transcripción con el propósito de alinear los límites temporales con la imagen espectrográfica y de especificar representaciones alofónicas de acuerdo con lo que fue pronunciado (Pineda *et al.* 2010).

La transcripción de los niveles T44 y T22 se hizo automáticamente a partir de la transcripción del nivel T54. En el nivel T44, una parte del etiquetado se hizo manualmente pues, a diferencia de los otros niveles, cuenta con la representación de las codas silábicas (Pineda *et al.* 2010). Hacer la transcripción del corpus en tres niveles distintos permitió construir un diccionario acústico de cada nivel, para los sistemas de reconocimiento. En el diccionario se incluyeron las pronunciaciones de cada palabra, que se obtuvieron a partir de las transcripciones. Esto ayudó a probar qué era más conveniente para los sistemas de reconocimiento, si tener más o menos muestras de las diferentes pronunciaciones de una palabra. También sirvió para demostrar que el *Corpus DIMEx100* es un recurso útil para la construcción de modelos acústicos de dichos sistemas (Pineda *et al*. 2010).

3.5. El *Speech View* del *CSLU* como herramienta para transcribir

El programa *Speech View* es una herramienta computacional que permite el análisis y estudio del habla. Esta herramienta pertenece al sistema *CSLU*, creado por el *Center for Spoken Language Understanding* del Oregon Graduate. El *CSLU* representa "an effort to make the core technology and fundamental infrastructure accessible, affordable and easy to use." (Sutton *et al*.1998:3221). Este sistema puede obtenerse e instalarse gratuitamente a través de la red electrónica mundial.

Para el estudio del habla, el *Speech View* cuenta con varias herramientas; entre ellas, un oscilograma manipulable que permite seleccionar determinada parte del audio y hacer acercamientos o alejamientos de la onda sonora; y tres espectrogramas para ver los diferentes matices de los formantes: uno de banda ancha, otro de banda estrecha y uno más a colores; estos espectrogramas posibilitan "el estudio de las transiciones y la observación de las trayectorias formánticas" (Llisterri *et al.* 1999:455). También por medio de ellos se pueden obtener datos sobre la frecuencia, la amplitud y la intensidad de los sonidos."Sobre el eje horizontal se halla el tiempo, y sobre el eje vertical, la frecuencia. La intensidad viene determinada por el mayor o menor grado de oscuridad de las secciones del documento" (Garrido *et. al* 1998:32). En la Figura 17 se puede observar el *Speech View* con sus diferentes oscilogramas y espectrogramas.

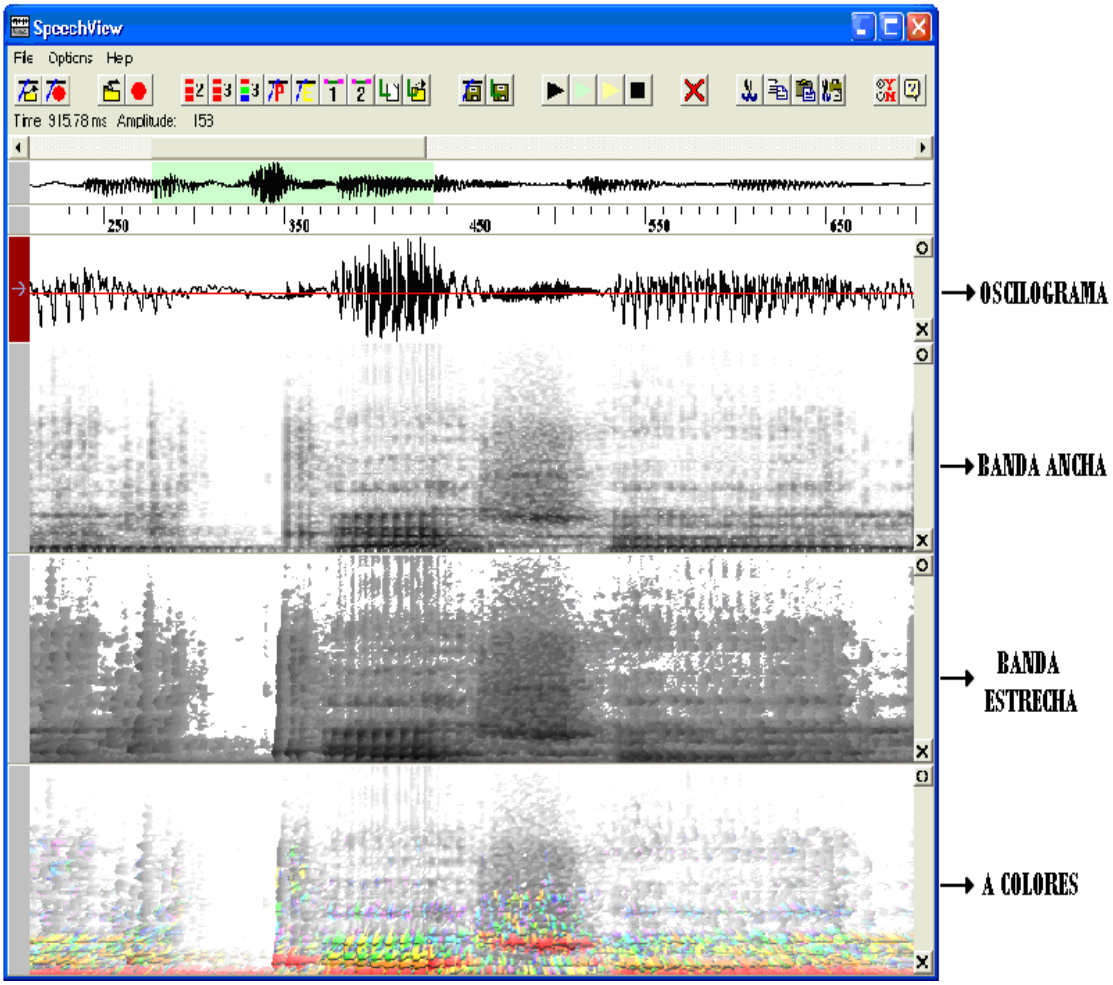

Figura 17. Herramienta computacional *Speech View* del CSLU.

Este programa también cuenta con una aplicación que permite la colocación de etiquetas para la transcripción de la señal sonora. Las etiquetas, además de describir los rasgos acústicos de la señal, delimitan tanto la imagen del alófono como la duración temporal del sonido. Para transcribir el corpus en diferentes niveles lingüísticos, la aplicación para el etiquetado puede abrirse cuantas veces se desee; por ejemplo, para el *Corpus DIMEx100* fueron necesarias cuatro barras para los niveles de transcripción T54, T44, T22 y Tp. En la Figura 18 podemos observar la oración *y qué hacer* etiquetada en los diferentes niveles.

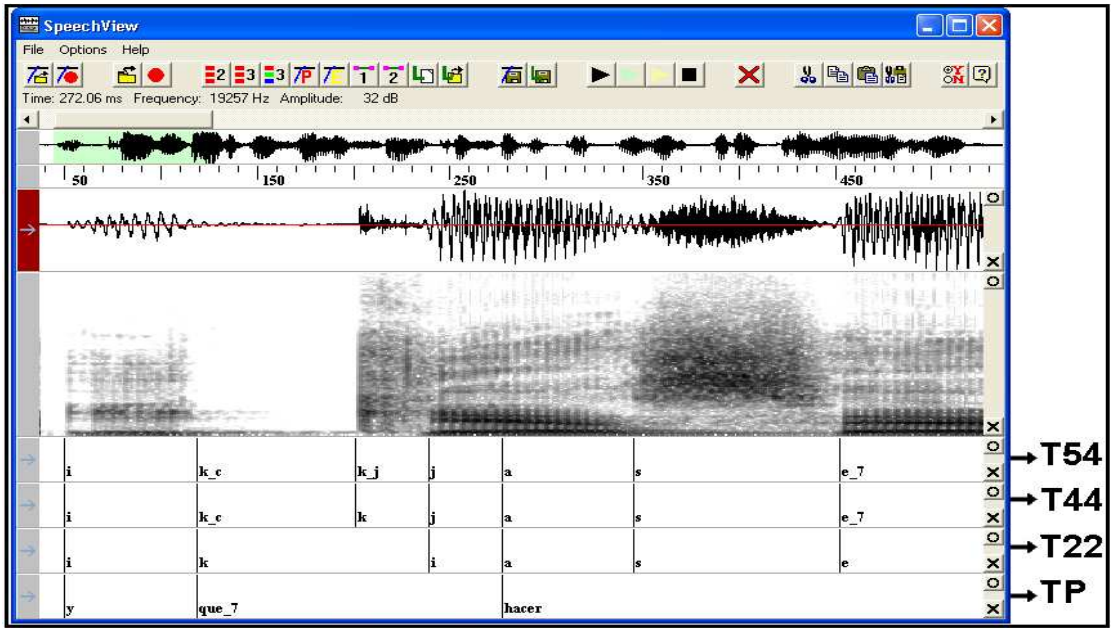

Figura 18. Etiquetado de los niveles de transcripción.

El *Speech View* es una herramienta muy práctica, ya que las diversas herramientas que contiene son opcionales y se abren en diferentes ventanas (Sutton *et al*.1998); así se pueden desplegar únicamente aquellas que se van a usar y en el orden deseado. Por ejemplo, para la transcripción de un corpus se pueden abrir simultáneamente el oscilograma, los diversos espectrogramas y las ventanas para la transcripción, lo que permite al etiquetador segmentar con más exactitud la señal acústica y, a su vez, alinear perfectamente las diversas etiquetas de transcripción. Otra de las funciones del *Speech View* es que las etiquetas de transcripción, se pueden guardar dentro de la misma carpeta del audio; así, se podrá consultar, simultáneamente, el audio con su respectiva transcripción.

En este capítulo se presentó al *Proyecto DIME*, su objetivo de trabajo y los recursos computacionales y lingüísticos que ha desarrollado. Entre los recursos computacionales están los diversos sistemas de reconocimiento del habla; en los recursos lingüísticos los corpus *DIME* y *DIMEx100.* Éste último se describió con más detenimiento, por ser el corpus que se utilizó para las imágenes del *Manual* que se presenta en esta tesis.

Posteriormente, se resaltó la importancia de la recopilación de corpus orales para los estudios lingüísticos y para los sistemas que desarrollan las tecnologías del habla. Después,

se describieron los procesos de segmentación y transcripción computacional de corpus. Ambos procesos se describieron en el contexto del *Proyecto DIME*. Finalmente, se presentaron dos de las herramientas que el proyecto ha utilizado para los procesos antes mencionados. Por un lado, se describió el alfabeto fonético computacional *Mexbet*, el cual posee un inventario alofónico para el español de la ciudad de México. Por otro lado, se describió el programa computacional *Speech View*, el cual contiene varias herramientas que permiten visualizar la señal sonora, segmentarla y hacer su transcripción.

En el siguiente capítulo se expondrán algunas de las dificultades que surgen durante la transcripción de corpus, de las cuales nacieron las necesidades e intereses para la creación de un manual de transcripción fonética computacional. Posteriormente, se describe el proceso de la elaboración de *Manual*.

4. Un Manual de etiquetado fonético, para su uso en las tecnologías del habla.

Como se ha podido ver a lo largo de los capítulos anteriores, los procesos de segmentación y transcripción de un corpus son parte fundamental en la programación de los sistemas de reconocimiento y síntesis de habla. Ambos procesos representan un gran esfuerzo de trabajo por varios motivos. Por un lado, porque para transcribir un corpus se necesita contar con un grupo numeroso de etiquetadores, suficiente presupuesto y, sobre todo, tiempo, ya que está constituido por muchas horas de grabación, lo que equivale a miles de oraciones. Por otro lado, porque cuando hay algún etiquetador principiante, es necesario enseñarle cómo hacerlo, lo que representa más tiempo en la transcripción del corpus. Aunado a lo anterior, si un corpus es transcrito por varias personas se corre el riesgo de que haya errores o falta de uniformidad en la segmentación y en la transcripción; por ello, es necesario hacer varias correcciones. Todas estas dificultades provocan que la transcripción de un corpus se haga larga, complicada y tediosa.

Por lo anterior, nació la propuesta de hacer un manual de etiquetado fonético computacional que, de alguna manera, resolviera o menguara las dificultades que se presentan durante la fase de etiquetado. Para ello, se estableció que el *Manual* tuviera como finalidades:

1) Uniformar, facilitar y agilizar el proceso de segmentación y transcripción.
2) Ser una guía accesible para los etiquetadores expertos e inexpertos.
3) Ser visualmente claro y atractivo, de manera que se pueda aprender a segmentar visualmente
4) Ser una herramienta de aprendizaje sobre fonética acústica.

Se establecieron estos objetivos por ser los que responden directamente a los problemas que surgen durante la transcripción de un corpus. Si el etiquetador aprende a reconocer las imágenes correspondientes de cada alófono, podrá segmentarlas con más exactitud y, asimismo, transcribir más rápido. El *Manual* se presenta como una herramienta de uso

sencillo, ya que debe ser entendible tanto para usuarios expertos como inexpertos; así, cuando surja alguna duda durante el proceso de transcripción, se podrá consultar rápidamente para solucionar el problema. Finalmente, el *Manual* contiene teoría sobre fonética acústica, para que el usuario adquiera conocimientos que le sirvan para llevar a cabo de una manera rápida y eficaz su trabajo; todo esto generará que, además, al final el corpus esté correctamente transcrito y uniforme. Para llevar a cabo los objetivos anteriores, se pensó que el *Manual* debía contener los siguientes parámetros:

1) Presentar la distribución complementaria de los alófonos por grupos fonemáticos.
2) Presentar las reglas de realización de cada uno de los alófonos.
3) Presentar la imagen acústica ejemplar de cada alófono.
4) Instruir en la segmentación y transcripción de los alófonos.
5) Detectar la aparición de fenómenos fonéticos recurrentes en el español.

Para la elaboración de los primeros dos puntos se recurrió al alfabeto *Mexbet* (Cuétara 2004), que posee un inventario de los fonemas y alófonos del español de la ciudad de México y sus reglas contextuales. Para el tercer punto, se tomó como objeto de estudio el *Corpus DIMEx100*. En él se hizo con la herramienta computacional *Speech View* una revisión de la imagen acústica de los 37 alófonos del español de la ciudad de México que propone *Mexbet* en su inventario. Para que esta revisión fuera ordenada, se hizo por fonemas y sus realizaciones. Por ejemplo, en el caso del fonema alveolar fricativo sordo /s/, se hizo la búsqueda de la imagen de cada uno de sus alófonos: el prototípico [s], el dental [s_[] y el sonoro[z].

Posteriormente, con base en la observación y el análisis de las distintas imágenes de cada alófono, se eligió la imagen ejemplar para cada uno de ellos; ésta debía reflejar, en la medida de lo posible, los elementos acústicos correspondientes a su alófono. Una vez seleccionado un fragmento, se alineó con sus etiquetas de la transcripción correspondientes a cada nivel (T54, T44, T22 y Tp) y se hizo una copia para la ilustración del *Manual*.

Una vez recopiladas las imágenes de los 54 alófonos y los datos obtenidos del análisis, se compararon y fundamentaron con base en la teoría, con el propósito de establecer

parámetros sobre la imagen acústica y, con base en ellos, plantear la segmentación para cada alófono.

Durante la búsqueda de las imágenes de los alófonos, se registraron dos fenómenos recurrentes en el habla: la homologación de sonidos idénticos y la elisión; de tal suerte, que resultó importante integrarlos al *Manual*. Cabe mencionar que ambos fenómenos se abordan de forma muy somera, puesto que ya se han hecho estudios sobre ellos en el *Proyecto DIME* (Ceballos 2007 y Espinosa 2007). Estos estudios han dado como resultado los contextos en los que ocurren y su recurrencia en el habla, motivo por el cual ha sido importante su implementación en los sistemas de reconocimiento.

Una vez obtenidos todos los materiales para la elaboración del *Manual*, se pensó cómo hacer su organización. Cabe mencionar que el *Manual* sufrió varios cambios durante su elaboración hasta llegar a la versión que aquí se presenta; esto ocurrió porque a cada momento se buscó la forma más didáctica para organizarlo.

Principalmente, el *Manual* se divide en tres partes: la primera está dedicada a las consonantes, la segunda a las vocales y la tercera a los fenómenos fonéticos. En la primera parte, los sonidos consonánticos aparecen clasificados por fonemas y sus realizaciones. Por ejemplo, el primer apartado se dedica a los fonemas oclusivos sordos y sonoros, el segundo al fonema palatal africado sordo, el tercero al fonema labiodental fricativo sordo, etc.

Para cada fonema se da una introducción teórica sobre sus realizaciones, basada en la teoría de la fonética acústica ya existente. Después, se presenta un cuadro con los alófonos, sus reglas de realización y sus representaciones ortográficas y fonéticas, en el alfabeto *Mexbet*. Posteriormente, se hace el análisis espectrográfico de cada alófono. Para ello, en primer lugar, se presenta la imagen acústica junto con la descripción del alófono, con el propósito de que el usuario pueda identificar visualmente la imagen de cada alófono. En seguida, se da la explicación para hacer la segmentación de la imagen espectrográfica, apoyándola con imágenes para una mejor comprensión del proceso. Finalmente, se muestra la transcripción correspondiente para las etiquetas de los tres niveles de transcripción (T22, T44, T54).

La segunda parte del *Manual* corresponde a las vocales; ésta aparece organizada de la misma manera que la de las consonantes, es decir, por fonemas. Para su mayor comprensión, en la clasificación de los sonidos vocálicos se sigue el orden del abecedario (a, e, i, o, u) y no el orden del triángulo vocálico. Sin embargo, no omite puntualizar la abertura vocálica de cada fonema, ya que ésta resulta importante para la ubicación de los formantes.

La tercera parte está dedicada a los fenómenos fonéticos. En ella se describen dos: la homologación de sonidos idénticos y la elisión. El primero se refiere a la unión de dos sonidos vocálicos o consonánticos idénticos, por ejemplo en la emisión de la frase "esa*s* *s*illas", se une la *s* final de *esas* con la inicial de *sillas*: *esa[s]illas*. El segundo es la pérdida de algún sonido de la palabra; éste se divide en tres: aféresis, síncopa y apocope. Cada uno de ellos hace referencia al lugar de la palabra donde se ha elidido algún sonido. Finalmente, en el apéndice se muestran las tablas de equivalencia de *Mexbet* con los alfabetos AFI y RFE, pues como se mencionaba, varias personas harán uso de este *Manual*, principalmente aquellos que colaboren dentro del *Proyecto DIME*. Por tal razón, es importante abarcar las necesidades para los diferentes usuarios del *Manual*.

Antes de terminar este capítulo, quiero mencionar que actualmente, varios países han comenzado a hacer el etiquetado automático, es decir, cuentan con programas computacionales que segmentan y transcriben oraciones; aunque esto representa un gran avance para las tecnologías y contrarresta casi totalmente las dificultades por las cuales se ha hecho este trabajo de tesis, se debe tomar en cuenta que dichos programas computacionales transcriben la realización esperada del alófono y no la que se pronunció realmente, como se hace en una transcripción manual (Goddijn y Binnenpoorte 2003). Además, no se debe pasar por alto que una segmentación y una transcripción manual son de suma importancia, porque ayuda al conocimiento de la lengua y, a su vez, a mejorar los sistemas de transcripción automática.

El *Manual de etiquetado fonético e imágenes acústicas de los alófonos del español de la ciudad de México, para su uso en las tecnologías del habla* aparece como apéndice de esta tesis. Se decidió ponerlo así, puesto que es un trabajo independiente por dos motivos: el primero es que este *Manual*, desde una perspectiva evolutiva de las herramientas para los estudio del habla, forma parte de nuevas herramientas que permiten estudiar el habla, ya que sus ilustraciones son muestra verídicas obtenidas a partir de un corpus y analizadas con programas computacionales; de la misma forma, pone evidencia que, con el uso de herramientas computacionales, los estudios del habla tienen mayor precisión en el análisis, pues permiten ver diferentes características del habla, como los formantes, la curva melódica, la frecuencia fundamental, etc. De esta manera se hacen evidentes los avances en los estudios que ha tenido la fonética acústica. El segundo es que el *Manual* surge en el marco del *Proyecto DIME*. Para este proyecto el *Manual* servirá como una herramienta de ayuda para instruir en el proceso transcripción de corpus, pues, como se ha venido diciendo a lo largo de este capítulo, el *Manual* contiene información teórica y práctica, para que el usuario pueda aprender a segmentar y transcribir la señal sonora; así como para adquirir conocimientos de fonética acústica. Por lo anterior, el *Manual* es un anexo de este trabajo; así, el usuario podría consultarlo y prescindir de los capítulos anteriores. No obstante, para los fines de esta tesis, los capítulos presentados son de suma importancia e indispensables, puesto que son el contexto teórico, histórico y metodológico que justifica su existencia.

5. Conclusiones

La propuesta de trabajo para esta tesis de licenciatura es la creación de un manual de etiquetado fonético computacional para el español de la ciudad México, implementado en las tecnologías del habla. Este trabajo se ha hecho en el contexto del *Proyecto DIME,* el cual ha trabajado durante años en la creación de sistemas de reconocimiento de habla, por lo tanto posee varios recursos para el estudio de la lengua. Los objetivos del *Manual* elaborado en esta tesis son ayudar y facilitar el proceso de segmentación y transcripción de corpus orales, así como crear un medio de aprendizaje para los etiquetadores principiantes y para los alumnos de fonética. Para entender la parte en donde se pondrá en uso este *Manual* se hizo un panorama sobre las tecnologías del habla. En el primer capítulo se presentó una introducción de este trabajo, en el cual se mencionaron los temas a desarrollar, así como los beneficios que tendría la construcción de un manual de etiquetado fonético.

En el segundo capítulo se habló de la lingüística computacional y de sus diferentes ramas. Entre ellas se destacaron a las tecnologías del habla y sus aplicaciones, en específico dos: los sistemas de síntesis y los sistemas de reconocimiento de habla; de estos se describió su arquitectura y los distintos módulos que los componen. Esto, con el propósito de observar que la construcción de ambos sistemas es un trabajo multidisciplinario, en el que colaboran conjuntamente varias disciplinas como la lingüística, la ingeniería y la computación.

Posteriormente, se resaltó el papel de la fonética en las tecnologías del habla pues ésta, entre otras cosas, trabaja para desarrollar elementos necesarios para que los sistemas de reconocimiento y síntesis del habla, reconozcan y reproduzcan el habla humana con más facilidad y naturalidad. Después, se abordó una de las ramas de la fonética, llamada fonética instrumental; para ésta, se hizo un recorrido a su historia para conocer los instrumentos de los que se ha valido para sus investigaciones, con el propósito de observar que de la misma manera que otras disciplinas han recurrido a la fonética para apoyarse de sus estudios, la fonética también se ha beneficiado de los avances tecnológicos de otras

disciplinas, tales como la informática y la computación. Para finalizar se abordaron dos alfabetos fonéticos para el español: el alfabeto de la *Revista de Filología Española* (RFE) y el *International Phonetic Alphabet* (IPA); se abordó su creación y sus características. El repaso de estos alfabetos fue importante, ya que éstos han servido como herramientas de estudio tanto para la fonética como para las tecnologías del habla y porque a partir de ellos se han hecho los alfabetos fonéticos computacionales.

En el tercer capítulo se presentó el *Proyecto Diálogos Inteligentes Multimodales en Español* (DIME) del Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas (IIMAS) de la Universidad Nacional Autónoma de México. De este proyecto se describieron los diversos recursos con los que cuenta para investigaciones en las áreas de lingüística y computación. Entre ellos, se trató con detenimiento el *Corpus DIMEx100*, por ser el recurso que se utilizó para esta tesis; se puntualizó el proceso de su recolección, así como los fines de su creación. Posteriormente, se enseñó cómo se llevan a cabo los procesos de recopilación de corpus orales, la segmentación y la transcripción de los mismos. Fue de gran importancia describir esos procesos, pues sirvió para mostrar dónde tendrá utilidad el *Manual*. Para finalizar, se presentaron dos herramientas importantes para la transcripción de los corpus del *Proyecto DIME*. Una de ellas es el alfabeto *Mexbet*, del cual se mencionaron sus diferentes versiones; con más detalle la última, realizada por Cuétara (2004), se expuso el trabajo que desarrolló para concretar un inventario pertinente para los alófonos del español de la ciudad de México. De esta última versión del alfabeto *Mexbet*, se tomó el inventario fonético y fonológico y las presentaciones simbólicas para el manual que se ofrece aquí. A lo largo de este capítulo, como de los otros, se recurrió frecuentemente al trabajo de Cuétara (2004). Esto porque dicho trabajado ha sido la base para la elaboración del manual, por ser el trabajo de investigación para definir el inventario alofónico del alfabeto *Mexbet* y de los corpus *DIME* y *DIMEx100*. La otra herramienta es el programa *Speech View*, que forma parte del paquete de herramientas del *CSLU* del Oregon Graduate Institute. De este programa, se describieron algunas de las herramientas que contiene para el análisis del habla, tales como los oscilogramas y los espectrogramas.

El capítulo cuarto se dedicó al *Manual de etiquetado fonético e imágenes acústicas de los alófonos del español de la ciudad de México, para su uso en las tecnologías del habla*. En éste se habló sobre las dificultades que motivaron su elaboración, y cómo a partir de ellas se buscó que el Manual tuviera dos objetivos principales: el primero, ser una herramienta que enseñe, guíe, ayude y facilite la segmentación y la transcripción de corpus a todas las personas que se deban enfrentar a ello, entre las que se destacan los estudiantes que colaboran en el *Proyecto DIME*. El segundo, ser un medio de estudio y aprendizaje sobre la fonética acústica del español de la ciudad de México, útil para los estudiantes de la asignatura de Fonética del Colegio de Lengua y Literaturas Hispánicas de la Facultad de Filosofía y Letras de la UNAM.

Como se pudo ver a lo largo de los capítulos, la segmentación y la transcripción de un corpus son trabajos sumamente laboriosos, y para llevarlos a cabo es necesario contar con un grupo numeroso de personas destinadas a estas tareas. En el *Proyecto DIME* las personas que hacen la transcripción de los corpus suelen ser estudiantes de diversas formaciones académicas, como lingüistas, ingenieros, computólogos, etc; por lo que muchos de ellos son inexpertos y, antes de iniciar la transcripción del corpus, es necesario enseñarles cómo hacerlo. Sumado a lo anterior, cuando un corpus es transcrito por varias personas se corre el riesgo de que haya inconsistencias, como la falta de uniformidad en el etiquetado, provocada por la mala segmentación entre las fronteras de los sonidos y por la confusión de las representaciones alofónicas, etc. Todas estas dificultades hacen que la transcripción del corpus demore mucho tiempo.

Por lo anterior, la propuesta para esta tesis de licenciatura fue la creación de un manual que muestre las imágenes de los alófonos del español de la ciudad de México, para su correcta segmentación y transcripción. Los alófonos que se muestran en el *Manual* están basados en el alfabeto *Mexbet* de Cuétara 2004, por lo que a lo largo de esta tesis se recurre constantemente a dicho trabajo, pues fue la base para definir el alfabeto y, por lo tanto, la base para el *Manual* de esta tesis.

Para tratar de menguar o subsanar las dificultades que se presentan en las labores antes mencionadas, se pensó que el *Manual* debía contener información pertinente y suficiente. Para eso, se presenta la distribución complementaria de los alófonos de la ciudad de México, sus reglas de realización, sus imágenes prototípicas y se instruye en su segmentación y transcripción.

Por medio de este *Manual* se espera que la segmentación y la transcripción de los corpus se agilicen y que, sobre todo, se hagan de manera más consistente. Con esos fines, en primer lugar, en el *Manual* se enseña a identificar las imágenes espectrográficas de los alófonos, así el etiquetador podrá distinguir, a simple vista, las características acústicas que lo definen. En segundo lugar, se instruye en cómo hacer la segmentación de cada una de las imágenes, con el propósito de evitar que se haga de manera incorrecta y que los segmentos alofónicos queden contaminados entre sí. En tercer lugar, se muestra la transcripción correcta para cada alófono, de acuerdo a sus reglas contextuales y a su realización.

Por lo anterior, este *Manual* será de gran utilidad en la transcripción fonética computacional de corpus orales, ya que ayudará a que dicha tarea sea más fácil, rápida y uniforme. Además, contar con este *Manual* durante la transcripción de un corpus representa una gran ventaja, ya que si el etiquetador tiene alguna duda durante el proceso, podrá consultarlo mientras transcribe, pues está organizado de forma sencilla y práctica; ya que está diseñado para dar capacitación previa a los etiquetadores principiantes, para resolver dudas sobre la segmentación y la transcripción de los alófonos, y para proveer de conocimientos de fonética acústica a sus lectores. Con ello se espera que el etiquetador, al ir transcribiendo y consultando el manual, lleve a cabo su trabajo sin errores y la transcripción del corpus sea consistente. En la descripción de estos procesos quedó explícita la necesidad de hacerlos de una manera cuidadosa, ya que con base en ellos se toma información para la construcción de los sistemas de síntesis y reconocimiento de habla.

Para finalizar, es importante mencionar que el manual presentado en este trabajo de tesis está enfocado al español de la ciudad de México; lo cual lejos de restarle mérito es una motivación a continuar con investigaciones que abarquen los diversos dialectos del español

de México, para que en un futuro tanto la lingüística como la computación cuenten con herramientas más robustas para sus estudios.

6. Referencias bibliográficas

AGUILAR, LOURDES, BEATRIZ BLECUA, MARÍA J. MACHUCA y RAFAEL MARÍN. 1993. "Phonetic reduction processes in spontaneous speech", en *3rd European Conference on Speech Communication and Technology*, Berlín: ISCAA Archive, pp. 433-436. [http://liceu.uab.es/publicacions/Aguilar_et_al_93_Phonetic_Reduction.pdf, 18 de marzo de 2009]

AGUILAR, LOURDES, JUAN MARÍA GARRIDO y JOAQUIM LLISTERRI. 1994. "Incorpusción de conocimientos fonéticos a las tecnologías del habla", comunicación presentada en el *I Congrés de Lingüística General*, Valencia: Universidad de Valencia. [http://liceu.uab.es/~joaquim/publicacions/valencia_94.html, 19 de marzo de 2009]

AGUILAR, LOURDES y MARÍA J. MACHUCA. 1995. "Pragmatic Factors Affecting the Phonetic Properties of Diphthongs", en *3rd European Conference on Speech Communication and Tecnology*, Madrid: ISCAA Archive, pp. 2251-2254. [http://liceu.uab.es/publicacions/Aguilar_Machuca_95_Pragmatics_Diphthongs.pdf, 19 de marzo de 2009]

ALARCOS LLORACH, EMILIO. 1965/1983. *Fonología española*, Madrid: Gredos.

ALBALÁ, MARÍA JOSÉ, ELENA BATTANER, MARCELA CARRANZA, JUANA GIL, JOAQUIM LLISTERRI, MARÍA J. MACHUCA, NATALIA MADRIGAL, MONTSERRAT MARQUINA, VICTORIA MARRERO, CARMEN DE LA MOTA, MONTSERRAT RIERA y ANTONIO RÍOS. 2008. "VILE: Nuevos datos acústicos sobre las vocales del español", *Language Design. Journal of Experimental and Theoretical Linguistics*, Special Issue 2, *New Trends in Experimental Phonetics, Selected Papers From the IV International Conference on Experimental Phonetics*, pp. 1-14. [http://liceu.uab.cat/~joaquim/phonetics/VILE/VILE_IVCFE08_Vocales.pdf, 13 de marzo de 2009]

ARRARTE, GERARDO. 1999. "Normas y estándares para la codificación de textos y para la ingeniería Lingüística", en *Filología e informática. Nuevas tecnologías en los estudios filológicos*, Barcelona: Milenio, pp.17-44.

AUTESSERRE, DENIS, GUY PÉRENNOU y MARÍA ROSSI. 1989. "Methodology for the transcription and labeling of a speech corpus", *Journal of the International Phonetic Association*, 19:1, pp. 2-15.

BARRY, WILLIAM J. y ADRIAN J. FOURCIN. 1992. "Levels of Labelling", *Computer Speech and Language*, 6, pp. 1-14.

BERNAL, JESÚS, JESÚS BOBADILLA y PEDRO GÓMEZ. 2000. *Reconocimiento de voz y fonética acústica*, Madrid: Ra-Ma.

BOLSHAKOV, IGOR y ALEXANDER GELBUKH. 2004. *Computational Linguistics. Models, Resources, Applications*, México: Instituto Politécnico Nacional, Universidad Nacional Autónoma de México y Fondo de Cultura Económica.

BONAVENTURA, PATRIZIA, FABIO GIULIANI, JUAN MARÍA GARRIDO e ISABEL ORTÍN. 1998. "Grapheme-to-phoneme transcription rules for Spanish, with application to automatic speech recognition and synthesis", en *Proceedings of COLING '98 Workshop*.

Partially Automated Techniques for Transcribing Naturally Occurring, Continuous Speech, Montreal: Universidad de Montreal, pp. 33-39.

CEBALLOS, ANA. 2007. *Fenómeno de pérdida en CORPUS DIME para su inclusión en un reconocedor de habla*, tesis de licenciatura inédita, México: Universidad Nacional Autónoma de México.
[http://acl.ldc.upenn.edu/W/W98/W98-0804.pdf, 20 de marzo de 2009]

CHENG-YUAN, LIN, JANG ROGER JYH-SHING y CHEN KUAN-TING. 2005. "Automatic Segmentation and Labeling for Mandarin Chinese Speech Corpus for Concatenation-based TTS", *International Journal of Computational linguistics and Chinese language Processing*, 10:2, pp. 145-166.
[http://www.aclclp.org.tw/clclp/v10n2/v10n2a1.pdf, 13 de abril de 2009]

CROOT, KAREN y BELINDA TAYLOR. 1995. *Criteria for Acoustic-Phonetic Segmentation and Word Labelling in the Australian National Database of Spoken Language*, Sidney: Universidad de Macquarie.
[http://andosl.anu.edu.au/andosl/general_info/aue_criteria.html, 21 de marzo de 2009]

CRYSTAL, DAVID. 1980/2003. *A dictionary of linguistics & Phonetics*, Malden: Blackwell.

CUÉTARA, JAVIER. 2004. *Fonética de la ciudad de México. Aportaciones desde las tecnologías del habla*, tesis de maestría inédita, México: Universidad Nacional Autónoma de México.

ESPINOZA, PRECIOSA. 2007. *Sistematización del fenómeno de silabicación en el CORPUS DIME para su aplicación en las tecnologías del habla*, tesis de licenciatura inédita, México: Universidad Nacional Autónoma de México.

GARRIDO, JUAN MARÍA, MARÍA J. MACHUCA y CARME DE LA MOTA. 1998. *Prácticas de fonética. Lengua española I*, Bellaterra: Universidad Autónoma de Barcelona.

GIBBON, DAFYDD, ROGER MOORE y RICHARD WINSKI. 1997. "SL corpus representation", *Handbook of Standards and Resources for Spoken Language Systems*, Berlin: Mouton De Gruyter, I, pp. 146-170.
[http://www.spectrum.unibielefeld.de/~gibbon/gibbon_handbook_1997/node128.html #SECTION06400000000000000000, 16 de abril de 2009]

GIL, JUANA. 1990. *Los sonidos del lenguaje*, Madrid: Síntesis.

GILI GAYA, SAMUEL. 1950/1975. *Elementos de fonética general*, Madrid: Gredos.

GODDIJN, SIMO y DIANA BINNENPOORTE. 2003. "Assessing Manually Corrected Broad Phonetic Transcriptions", en *Proceedings of 15th Congress of Phonetic Sciences*, Barcelona: ISCAA, pp. 1361-1364.
[http://lands.let.kun.nl/literature/goddijn.2003.1.pdf, 17 de abril de 2009]

GUSSENHOVEN, CARLOS y HAIKE JACOBS. 1998. *Understanding Phonology*, Londres: Hodder Arnold.

HELMHOLTZ, HERMANN. 1877/1964. *On the sensations of tone as physiological basis for the theory of music*, New York: Dover.

HERRERA, ESTHER. 2002. "La asimilación de las nasales en español. Un estudio instrumental", *Nueva Revista de Filología Hispánica*, 1:1, pp. 1-14.

HIDALGO NAVARRO, ANTONIO y MERCEDES QUILIS MERÍN. 2004. *Fonética y fonología españolas*, Valencia: Tirant lo Blanch.

HIERONYMUS, JAMES. 1994 (Ms.). "ASCII phonetic symbols for the world's languages: Worldbet", Nueva Jersey.

INTERNATIONAL PHONETIC ASSOCIATION. 1999. *Handbook of the International Phonetic Association. A guide to the use of the International phonetic Alphabet*, Cambridge: Cambridge University Press.

IRIBARREN, MARY. 2005. *Fonética y fonología españolas*, Madrid: Síntesis.

JAKOBSON, ROMAN y LINDA R. WAUGH. 1979/1987. *La forma sonora de la lengua*, México: Fondo de Cultura Económica.

LADEFOGED, PETER. 1962. *Elements of Acoustic Phonetics*, Chicago: The University of Chicago Press.

————————. 2000. *A course in Phonetics*, Boston: Thomson Wadsworth.

————————. 2001/2005. *Vowels and consonants*, Massachussets: Blackwell.

LANDER, TERRY. 1997. *The CSLU Labeling Guide*, Portland: Oregon Graduate Institute. [http://cslu.cse.ogi.edu/corpus/docs/labeling.pdf, 15 de febrero de 2009]

LASS, ROGER. 1984. *Phonology. An introduction to basic concepts*, Cambridge: Cambridge University Press.

LLISTERRI, JOAQUIM. 1991. *Introducción a la fonética: El método experimental*, Barcelona: Anthropos.

————————. 1997. "Transcripción, etiquetado y codificación de corpus orales", en *Etiquetación y extracción de información de grandes corpus textuales*, curso impartido en *Seminario de Industrias de la Lengua*, Soria. [http://liceu.uab.es/~joaquim/publicacions/FDS97.html, 03 de marzo de 2009]

————————. 2003a. "Las tecnologías del habla", en *Tecnologías del lenguaje*, M. A. Martí (Coord.), Barcelona: UOC, pp. 249-282.

————————. 2003b. "Lingüística y tecnologías del lenguaje", en *Lynx. Panorama de estudios lingüísticos*, 2, pp. 9-71.
[http://liceu.uab.es/~joaquim/publicacions/TecnoLing_Lynx02.pdf, 17 de marzo de 2009]

————————. 2007. "El papel de la fonética en las tecnologías del habla", en *Actas do 3º Internacional de Fonética Experimental*, Santiago de Compostela: Junta de Galicia, pp. 23-37.
[http://liceu.uab.es/~joaquim/publicacions/Llisterri_05_Fonetica_Teconologias_Habla_pdf, 13 de marzo de 2009]

LLISTERRI, JOAQUIM y DOLORS POCH. 1994. "Proyecto de una base de datos acústicos de la lengua española", en *Actas del Congreso de la lengua Española*, Madrid: Instituto Cervantes, pp. 278-292.
[http://cvc.cervantes.es/obref/congresos/sevilla/tecnologias/ponenc_llisterripoch.htm, 13 de marzo de 2009]

LLISTERRI, JOAQUIM, LOURDES AGUILAR, JUAN MARÍA GARRIDO, MARÍA J. MACHUCA, RAFAEL MARÍN, CARMEN DE LA MOTA y ANTONIO RÍOS. 1999. "Fonética y tecnologías del habla", en *Filología e informática. Nuevas tecnologías en los estudios filológicos*, Barcelona: Milenio, pp. 449-479.
[http://liceu.uab.es/~joaquim/publicacions/Fonetica_TecnolHabla.pdf, 11 de marzo de 2009]

LLISTERRI, JOAQUIM y MARÍA A. MARTÍ. 2002. "Las tecnologías lingüísticas en la Sociedad de la Información", en *Tratamiento del lenguaje natural. Tecnología de la lengua oral y escrita*, M. A. Martí y J. Llisterri (Eds.), Barcelona: Fundación Duques de Soria y Edicions Universidad de Barcelona, pp. 13-28.

LOPE BLANCH, JUAN. 1963-1964/1983. “En torno a las vocales caedizas del español mexicano”, en *Estudios sobre el español de México*, México: UNAM, pp.57-77.

LÓPEZ, FERNANDA. 2004. *El estudio de los diptongos del español de México para su aplicación en un reconocedor de habla*, tesis de licenciatura inédita, México. Universidad Nacional Autónoma de México.

MACHUCA, MARÍA J., MONTSERRAT RIERA y ANTONIO RÍOS. 1999. (MS.). *Criteris de segmentació i etiquetatge del corpus prosòdic del CREL per a l'estudi de la durada segmental*, Informe del Seminari de Filologia i Informàtica: Universitat Autònoma de Barcelona.

MARTENS, JEAN PIERRE, DIANA BINNENPOORTE, KRIS DEMUYNCK, RUBEN VAN PARYS, TOM LAUREYS, WIM GOEDERTIER y JACQUES DUCHATEAU. 2002. “Word segmentation in the spoken Dutch corpus”, en *Proceedings 3rd International Conference on Language Resources and Evaluation*, 5, pp. 1432-1437. [http://www.esat.kuleuven.ac.be/~spch/cgi-bin/get_file.cgi?/tlaureys/lrec02/cgn_all/lrec_oplijning.pdf, 19 de abril de 2009]

MARTÍNEZ CELDRÁN, EUGENIO. 1998. *Análisis espectrográfico de los sonidos del habla*, Barcelona: Ariel.

MATLUCK, JOSEPH. 1951. *La pronunciación en el español del Valle de México*, México: Edición de autor.

MORENO, ANTONIO. 1998. *Lingüística computacional*, Madrid: Síntesis.

MORENO DE ALBA, JOSÉ. 1988. *El español en América*, México: Fondo de Cultura Económica.

————————. 1994. *La pronunciación del español en México*, México: Colegio de México.

MOTA, CARME DE LA y ANTONIO RÍOS. 1995. “Problemas en torno a la transcripción fonética del español: los alfabetos fonéticos propuestos por IPA y RFE y su aplicación a un sistema automático”, en *Acta Universitatis Wratislaviensis*, Wroclaw: Universidad de Wroclaw, pp. 97-109.

NAVARRO TOMÁS, TOMÁS. 1918/2004. *Manual de pronunciación española*. Madrid: Consejo Superior de Investigaciones Científicas.

PÉREZ, ELIA. 2006. *Construcción de un reconocedor de voz utilizando Sphinx y el corpus DIMEx100*, tesis de licenciatura inédita, México: Universidad Nacional Autónoma de México.

PERISSINOTTO, GIORGIO. 1975. *Fonología del español hablado en la ciudad de México. Ensayo de un método sociolingüístico*, México: Colegio de México.

PINEDA, LUIS A. 2008. *El proyecto DIME y el robot conversacional Golem: una experiencia multidisciplinaria entre la computación y la lingüística*, México: Universidad Nacional Autónoma de México. [http://leibniz.iimas.unam.mx/~luis/papers/DIME-Golem.pdf, 5 de marzo de 2009]

PINEDA, LUIS A., ANTONIO MASSÉ, IVAN MEZA, MIGUEL SALAS, ERIK SCHWARZ, ESMERALDA URAGA y LUIS VILLASEÑOR. 2002. “The Dime Project”, *Lectures Notes in Artificial Intelligence*, 2313, Berlín: Springer-Verlarg, pp. 166-175. [http://leibniz.iimas.unam.mx/~luis/DIME/, 3 de marzo de 2009]

PINEDA, LUIS A., LUIS VILLASEÑOR, JAVIER CUÉTARA, HAYDE CASTELLANOS e IVONNE LÓPEZ. 2004. “DIMEx100: A New Phonetic and Speech Corpus for Mexican Spanish”, *Advances in Artificial Intelligence*, pp. 974-983.

[http://leibniz.iimas.unam.mx/~luis/DIME/publicaciones/papers/DIMEx100-LNAI3315.pdf, 9 de marzo de 2009]

PINEDA, LUIS A., HAYDE CASTELLANOS, JAVIER CUÉTARA, LUCIAN GALESCU, JANET JUÁREZ, JOAQUIM LLISTERRI, PATRICIA PÉREZ Y LUIS VILLASEÑOR. 2010. "The Corpus DIMEx100: Transcription and Evaluation", *Language and Resources and Evaluation*, 44, Berlín: Springer, pp. 347-370. DOI: 10.1007/s10579-009-9109-9.

QUILIS, ANTONIO. 1981/1988. *Fonética acústica de la lengua española*, Madrid: Gredos.

————————. 1999. *Tratado de fonología y fonética españolas*, Madrid: Gredos.

RAFEL, JOAQUIM y JOAN SOLER. 2003. "El procesamiento de corpus. La lingüística empírica", en *Tecnologías del lenguaje*, M. A. Martí (Coord.), Barcelona: UOC, pp. 41-74.

ROACH, PETER, HELEN ROACH, ANDREA DEW y PAUL ROWLANDS. 1990. "Phonetic analysis and the automatic segmentation and labeling of speech sounds", *Journal of the International Phonetic Association*, 20:1, pp. 15-21.

SUTTON, STHEPEN, RONALD COLE, JACQUES DE VILLIERS, JOHAN SCHALKWYK, PIETER VERMEULEN, MIKE MACON, YONGHONG YAN, ED KAISER, BRIAN RUNDLE, KHALDOUN SHOBAKI, PAUL HOSOM, ALEX KAIN, JOHAN WOUTERS, DOMINIC MASSARO y MICHAEL COHEN. 1998. "Universal Speech Tools: The CSLU toolkit", *5th International Conference on Spoken Language Processsing*, 7, pp. 3221-3224. [http://cslu.cse.ogi.edu/publications/ps/sutton_ICSLP_98_CSLUToolkit.pdf, 03 de abril de 2009]

TAPIAS, DANIEL. 2002. "Interfaces de voz con lenguaje natural", en *Tratamiento del lenguaje natural. Tecnología de la lengua oral y escrita*, M. A. Martí y J. Llisterri (Eds.), Barcelona: Fundación Duques de Soria y Edicions Universidad de Barcelona, pp. 189-207.

TORRUELLA, JOAN y JOAQUIM LLISTERRI. 1999. "Diseños de corpus textuales y orales", en *Filología e informática. Nuevas tecnologías en los estudios filológicos*, Barcelona: Milenio, pp. 45-77.

VILLASEÑOR, LUIS, ANTONIO MASSÉ y LUIS A. PINEDA. 2001. "The DIME Corpus", en *Tercer Encuentro Internacional de Ciencias de la Computación ENC-01*, Aguascalientes: SMCC-INEGI, pp.591-600. [http://leibniz.iimas.unam.mx/~luis/DIME/publicaciones/papers/DIMECorpus.PDF, 9 de marzo de 2009]

7. Apéndice

MANUAL DE ETIQUETADO FONÉTICO E IMÁGENES ACÚSTICAS DE LOS ALÓFONOS DEL ESPAÑOL DE LA CIUDAD DE MÉXICO, PARA SU USO EN LAS TECNOLOGÍAS DEL HABLA

MONTSERRAT ALEJANDRA CHAVARRÍA AMEZCUA

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

México, Diciembre 2010

Índice

Presentación

Las tecnologías del habla son una rama de la lingüística computacional, la cual se dedica a la construcción de sistemas de síntesis y reconocimiento de habla. Para esta disciplina es ineludible la participación de la fonética, ya que para mejorar los sistemas de síntesis y reconocimiento del habla se necesita tener una descripción detallada de los sonidos de la lengua. Para ello, es necesario recopilar corpus orales, los cuales son analizados por medio de herramientas computacionales que permiten abrir las grabaciones, ver la imagen acústica, segmentarla en unidades lingüísticas –en este caso en alófonos– y hacer su transcripción en etiquetas.

La segmentación y la transcripción de corpus se pueden hacer automática o manualmente. A simple vista, hacer estas labores puede parecer un trabajo sencillo; sin embargo no lo es, y en menor medida cuando se trata de transcribir todo un corpus, ya que está conformado por horas de grabaciones, lo que da como resultado miles de oraciones, que se deben de segmentar y transcribir en palabras y, a su vez, en alófonos. Generalmente, esta empresa es supervisada y realizada por personas expertas, pero en ocasiones puede suceder que se haga por personas principiantes, a quienes es necesario enseñarles cómo hacerla, lo que ocasiona que el proceso de transcripción demore mucho tiempo y que se corra el riesgo de que sea inconsistente.

Por lo anterior, se pensó en la elaboración de este *Manual* como una herramienta que ayude y muestre cómo hacer la segmentación y la transcripción fonética computacionales, con el objetivo de facilitar ambas labores y que, a su vez, se hagan de manera precisa y uniforme. Con este fin, el *Manual* se enfoca en el reconocimiento visual de la imagen acústica de los alófonos del español de la ciudad de México.

Este manual está dirigido, principalmente, a los estudiantes que participan en el *Proyecto DIME*; por esta razón, la transcripción de los alófonos está hecha con el alfabeto fonético computacional *Mexbet* (Cuétara 2004), el cual se hizo en el marco del *Proyecto*. A lo largo del *Manual* se recurre constantemente al trabajo de Cuétara (2004), ya que dicho trabajo se

tomó como base para la elaboración de este *Manual*, por ser la base para definir al alfabeto *Mexbet*, con el cual se han transcrito los corpus *DIME* y *DIMEx100* del mismo proyecto.
El manual también puede ser útil para los estudiantes que cursan la asignatura de Fonética, ya que se fundamenta en datos teóricos, que enseñarán al alumno, así como al etiquetador, sobre fonética acústica y articulatoria; y en general a todas las personas que estén interesadas en estas ramas de la Fonética.

Para fines prácticos este *Manual* está organizado en dos apartados: el primero está dedicado a los fonemas consonánticos y el segundo a los vocálicos. Cada apartado está dividido en grupos fonemáticos y estos, a su vez, se subdividen en fonemas y sus alófonos. Al principio de cada fonema se presenta información teórica sobre su articulación y sus realizaciones alofónicas. Posteriormente, se enseña cómo segmentar y transcribir cada alófono, estos procesos se apoyan con imágenes que muestran paso a paso cómo hacerlos. Con esta organización se busca, por un lado, que si el etiquetador tiene una duda durante la transcripción del corpus pueda encontrar la información que desee de forma rápida y sencilla. Por otro lado, que al consultar el *Manual* pueda conocer y ubicar a los sonidos por su grupo fonemático.

Este *Manual* es una guía de lo que el etiquetador puede encontrar en cuanto a las imágenes alofónicas; sin embargo, puede suceder que las imágenes salgan fuera de lo esperado, entonces, se debe transcribir el alófono que realmente fue articulado. Cabe mencionar que las imágenes que se muestran aquí son una reproducción, pero cuando el etiquetador comience a hacer la transcripción tendrá una versión electrónica que le permitirán ver las imágenes con mucho más detalle, ya que son de alta calidad. Esto le ayudará a identificar de la mejor manera todos los rasgos acústicos que se refieren en el *Manual*.

Para finalizar esta presentación, quiero agradecer al Maestro Javier Cuétara Priede de la Facultad de Filosofía y Letras de la UNAM, por ser un excelente asesor en la elaboración de este manual. También agradezco al Doctor Luis Pineda, Jefe del Departamento de Ciencias de la Computación del IIMAS, por permitirme el uso del *Corpus DIMEx100*, para obtener las imágenes que ilustran este manual.

I. FONEMAS CONSONÁNTICOS

1. **Fonemas oclusivos /p/, /t/, /k/, /b/, /d/ /g/**

El español de la ciudad de México tiene seis fonemas oclusivos; tres sordos /p/, /t/, k/ y tres sonoros /b/, /d/, /g/. Estos seis fonemas oclusivos se pueden distinguir porque se articulan en distintos puntos de la cavidad bucal: son bilabiales /p/ y /b/, dentales /t/ y /d/, y velares /k/ y /g/. El modo de articulación es parecido para todos, y se lleva a cabo con dos movimientos: en el primero los órganos de la cavidad bucal se cierran de tal manera que impide la salida del aire. Posteriormente, en el segundo movimiento se abren y el aire contenido sale haciendo una explosión. Respecto a la articulación de los fonemas oclusivos, Gili Gaya (1950:124) menciona que las consonantes oclusivas se distinguen por producirse en tres tiempos: "implosión, oclusión y explosión. En el primero, los órganos se mueven para adquirir la posición articulatoria propia de la consonante; durante la oclusión mantiene cerrada la salida del aire; la explosión es un movimiento de abertura que deshace el contacto para pasar al sonido siguiente", con él coinciden Hidalgo y Quilis (2004) e Irribarren (2005).

A diferencia de esos autores, Quilis (1981:190) no menciona el momento de implosión; únicamente menciona como características de la articulación oclusiva el *cierre* y la *explosión*, y añade otra: "Tres son las características que distinguen fundamentalmente estas consonantes del resto: *a)* la interrupción total en la emisión del sonido (esta interrupción se produce durante la tensión de la consonante); *b)* la explosión que sigue a esta interrupción (explosión que se manifiesta en forma de sonido turbulento, breve e intenso); *c)* la rapidez de las transiciones de los formantes de las vocales precedentes o siguientes"; con él coincide Alarcos (1965), quien menciona que las oclusivas se reconocen en el espectrograma por la ausencia de energía, por la barra de explosión y por la rapidez de las transiciones de los formantes de la vocal precedente. Como se puede observar los autores coinciden en las características de *cierre* y *explosión*, por lo que se puede concluir que son las principales características de las consonantes oclusivas, tal como señala Jacokbson (1979:101) "Las explosiones [Burst] y su duración y energía relativas

desempeñan un gran papel en la identificación y la distinción de oclusivas, especialmente cuando las oclusivas no están junto a vocales; sobre todo la /k/ y la /g/ compactas, ya sea que estén junto a vocales o no, requieren una explosión para su identificación".

Por su parte, Navarro Tomás (1918:78) señala que cuando las oclusivas /p/ y /k/ están al final de sílaba y precedidas de otra consonante oclusiva "se reducen a articulaciones meramente implosivas"; es decir, carecen de explosión. Pone de ejemplo la palabra *apto* y explica "mientras los labios están cerrados, forma la lengua la oclusión de dicha *t* sin dar tiempo a la salida del aire para la explosión de *p*" (1918:83).

Dichas características principales de las consonantes oclusivas pueden verse y ubicarse fácilmente en el espectrograma. Lander (1997:16) menciona que "When a word begins with any of these sounds, look for spectral evidence to signal the beginning of the closure". Efectivamente, en las imágenes espectrográficas de las oclusivas primero aparece el *cierre* que se caracteriza por ser un espacio en blanco, pues como se dijo, es un momento en el que no se emite sonido alguno a causa de la obstrucción del aire. Posteriormente, aparece la *explosión* que se ve como una barra vertical oscura, dicha barra corresponde al estallido que emite el aire al ser expulsado (ver Figura 4.).

A pesar de que la imagen espectrográfica de este grupo de consonantes es semejante, se puede diferenciar entre sordas y sonoras, ya que en el cierre de las consonantes oclusivas sonoras las cuerdas vocales comienzan a vibrar hasta el momento de la explosión. Esta vibración se refleja en el segmento espectrográfico del cierre con una barra llamada *barra de sonoridad*. Dicha barra es semejante a un formante y aparece en la parte inferior del espectrograma e incluso puede presentar ligeras estrías formánticas en la parte superior (ver Figura 21). Quilis (1981:192) menciona acerca de esta diferencia que "Los espectrogramas de las explosivas sordas se caracterizan por la ausencia total de zonas de frecuencia; en las sonoras, esta ausencia también es patente, pero una barra de sonoridad en la parte inferior de su espectro las diferencia de las anteriores; esta barra de sonoridad se origina por la vibración de las cuerdas".

Si bien se pude diferenciar la imagen acústica entre oclusivas sonoras y sordas, lo que no se puede hacer –o cuando menos es difícil– es tratar de diferenciar entre mismos grupos de oclusivas; es decir, entre mismas sordas o sonoras. Acerca de esto, Quilis (1981:192) señala que "en sí el espectro de las explosivas sordas y sonoras no proporciona ningún dato que las caracterice y que pueda explicar por qué percibimos [p] como diferente de [k], o [b] como diferente de [g]". Sin embargo, hay autores que afirman que se puede diferenciar estas consonantes por medio de la barra de explosión, como Hidalgo y Quilis (2004), quienes mencionan que la barra de explosión de las oclusivas labiales es breve, la de las dentales visible y la de las velares muy amplia; con ellos coinciden Bernal *et al*. (2000:46), quienes explican que la diferencia entre oclusivas sordas se basa por el nivel de la energía de la barra de explosión: "Los sonidos sordos presentan una zona de silencio seguida por una breve barra de explosión vertical, que tiene mayor duración temporal en el sonido [k]. La barra de explosión contiene más energía en la zona baja del espectro en el caso bilabial, en la zona media cuando se trata del sonido [t] y en la parte alta para [k]". Como se puede ver, los autores proponen distintas formas de diferenciar espectrográficamente entre mismas oclusivas sordas o sonoras. Sin embargo, para este manual he preferido no tomar en cuenta este rasgo diferenciador, pues resultaría conveniente hacer un trabajo en el que se documenten estos casos, con el propósito de comprobar qué tan importante resulta la barra de explosión como un rasgo diferenciador entre dichas consonantes.

Otra característica espectrográfica de las consonantes oclusivas son las transiciones de sus formantes con respecto a los formantes de los fonemas vocálicos. Quilis (1981:138) menciona que los formantes se crean por "la excitación producida por la glotis se extiende a las cavidades supraglóticas, que, al actuar como filtros, estructuran la señal acústica. Desde la glotis hasta los labios se pueden considerar una serie de cavidades resonantes que, a través de la función de transferencia o, lo que es lo mismo, de la función del filtrado del conducto vocal, originan los llamados *formantes*. Un formante de la onda acústica del lenguaje es, por tanto, un máximo de la función de transferencia del conducto vocal". En el espectrograma un formante se puede distinguir por ser una banda horizontal más oscura que al resto de la imagen espectrográfica (Ladefoged 2001).

Las transiciones de los formantes son producto de un cambio en la frecuencia producido por el movimiento articulatorio; estas transiciones se pueden observar en el cambio de un formante a otro y pueden ser positivas, negativas o neutras, dependiendo de la consonante o vocal con la que estén en contacto. Así, en los alófonos oclusivos la transición del primer formante (F1) normalmente será negativa, mientras que las transiciones del segundo formante (F2) varían, ya que este segundo formante refleja el lugar de articulación de las oclusivas (Quilis 1999). En la Figura 1 se pueden ver las transiciones del F2, según la consonante y vocal que estén en contacto.[8]

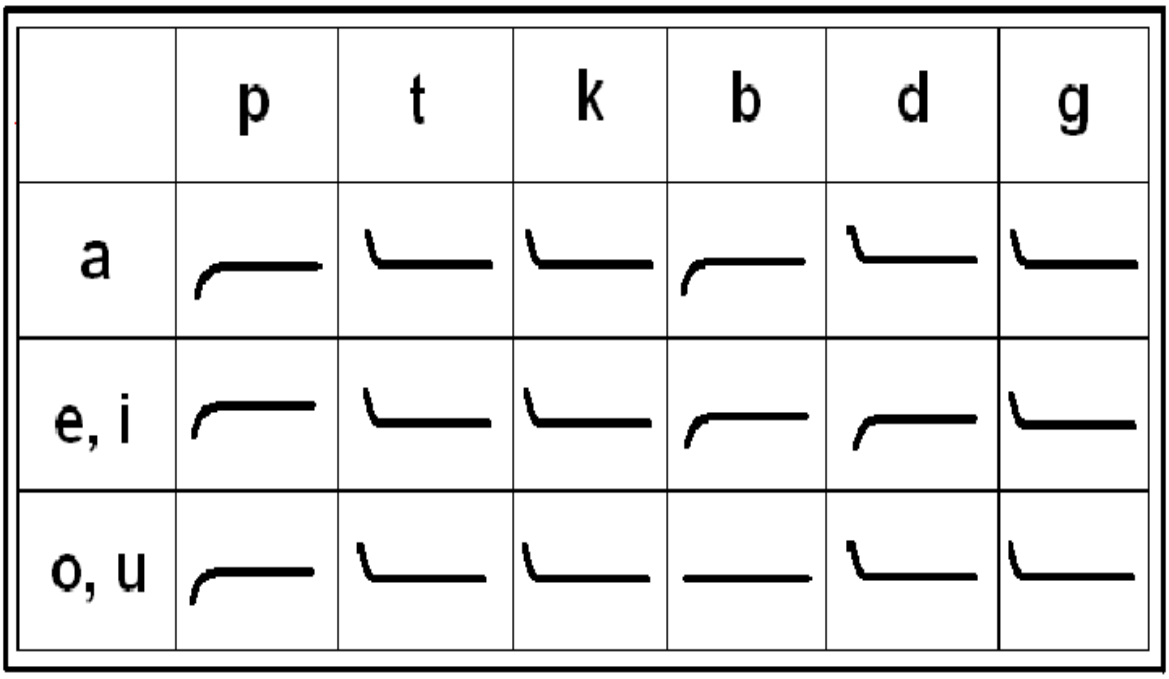

Figura 1. Transiciones del F2 vocálico en contacto con consonantes oclusivas

Cada fonema oclusivo tiene sus respectivos alófonos en distribución complementaria. Los fonemas oclusivos sordos, /p/ y /t/, tienen un sólo alófono oclusivo sordo para cada uno, los cuales ocurren cuando están en posición inicial de sílaba. El fonema /k/ tiene dos alófonos: uno oclusivo y otro palatalizado, que toma está realización cuando está en contacto con los alófonos vocálicos palatales [e], [i] y [j] (Cuétara 2004:95).

[8] Estas transiciones formánticas las he tomado del *Tratado de fonología y fonética* de Antonio Quilis.

Los fonemas oclusivos sonoros tiene dos alófonos en distribución complementaria: uno oclusivo que se da en inicio absoluto de sílaba y posterior a [m, n]; y otro aproximante que se da en todos los demás contextos.

1.1. Alófonos aproximantes [V], [D], [G]

Los alófonos aproximantes [V], [D] y [G] son formas alofónicas de los fonemas oclusivos sonoros /b/, /d/ y /g/. Se diferencian de los alófonos oclusivos en primer lugar por su articulación y en segundo por el contexto en el que se realizan.

Estos alófonos se ocurren en contexto interior y final de sílaba. Reciben el nombre de *aproximantes* porque los órganos de la cavidad bucal con los que son articulados se aproximan para emitir el sonido, de la misma manera que en los sonidos africados, aunque la constricción de los órganos en los aproximantes, es menor que en los africados (Gil 1990, Hidalgo y Quilis 2004).

La imagen espectrográfica de los alófono aproximantes se caracteriza por tener formantes parecidos a los vocálicos, pero de baja frecuencia. Martínez Celdrán (1998:71) menciona que dichos formantes “son meras transiciones entre los formantes vocálicos y su frecuencia determina su distinto punto de articulación, sobre todo de las transiciones de F2 y de la frecuencia de ese mismo F2”.

1.2. Codas silábicas [-B], [-D], [-G]

El término *coda silábica* se refiere a la posición final de sílaba en la que puede aparecer un alófono, en específico [p], [t], [k], [V], [D], [G] [m], [n_[], [N], [r(] y [r]. Cuando estos alófonos se encuentran en posición de *coda silábica* pueden alternar su realización, pero sólo con aquellos que tienen rasgos en común, o bien, que se oponen por sonoridad. Al conjunto de alófonos que presentan en común y opuestos se les ha neutralizado y se le ha llamado *codas*. Estos son representados bajo un símbolo en específico.

En los alófonos oclusivos la neutralización se da por sonoridad y por punto de articulación, así se oponen los bilabiales: [p] / [V], los dentales: [t] / [D] y los velares: [k] / [G] (Hidalgo y Quilis 2004); por ejemplo *apto ~ abto, atmósfera ~ admósfera, agnóstico ~ acnóstico,* etc. A estos pares se les ha representado fonéticamente en el alfabeto *Mexbet* con las *codas* [-B], [-D], [-G] respectivamente. Las codas se transcriben únicamente en el nivel T44. En la Figura 2 aparecen los alófonos oclusivos que se oponen y el símbolo correspondiente de coda silábica en *Mexbet*.

Oposición de alófonos	Representación en *Mexbet*	Ejemplo
[p] / [V]	-B	*abrupto* [abru-Bto]
[t] / [D]	-D	*ciudad* [siuda-D]
[k] / [G]	-G	*acto* [a-Gto]

Figura 2. Representación en *Mexbet* de las codas de oclusivas sordas y aproximantes.

Cabe mencionar que no siempre se presentará la alternancia entre los alófonos que se oponen, eso depende en gran medida de la pronunciación del hablante. Quilis (1999:205) señala que “depende tanto de lo hábitos o del énfasis del hablante, como de la norma regional: puede aparecer desde el mantenimiento como explosiva sorda o sonora, hasta su desaparición.”. Los hablantes del *Corpus DIMEx100* con frecuencia debilitaron o perdieron la [d] final de sílaba, esta pérdida también ha sido documentada en el español de México por Moreno de Alba (1994).

En el siguiente apartado se presentan las reglas combinatorias de los alófonos de los fonemas oclusivos, conforme a Cuétara 2004, junto con el proceso de segmentación y transcripción para cada uno de ellos.

1.4. Fonemas oclusivos sordos /p/, /t/, /k/

1.4.1. Fonema bilabial oclusivo sordo /p/

El fonema /p/ cuenta con un solo alófono, [p], el cual se realiza en todos los contextos; en otras palabras, el fonema /p/ siempre se pronunciará como [p], sin importar el lugar que ocupe en la palabra o junto a qué consonantes o vocales se emita. Perissinotto (1975:44) reporta a /p/ como un fonema "muy estable y resiste tanto la sonorización como el relajamiento en cualquier posición". En la Figura 3 se muestra un cuadro con la regla distribucional de /p/, de acuerdo con Cuétara 2004.

Fonema	Alófono	Contexto	Grafía
bilabial oclusivo sordo /p/	Bilabial oclusivo sordo [p_c] [p]	Inicio de sílaba	*p*

Figura 3. Reglas distribucionales de los alófonos del fonema /p/.

Como se dijo anteriormente (§1) para localizar un alófono oclusivo, en este caso [p], en una imagen espectrográfica, es necesario primero ubicar el *cierre* u *oclusión*, el cual se distingue por un espacio en blanco. Posterior al cierre, se encontrará la *barra explosión*. En la Figura 4 se muestra una imagen espectrográfica, en la cual se pueden observar ambos segmentos, cada uno de ellos señalado por su nombre.

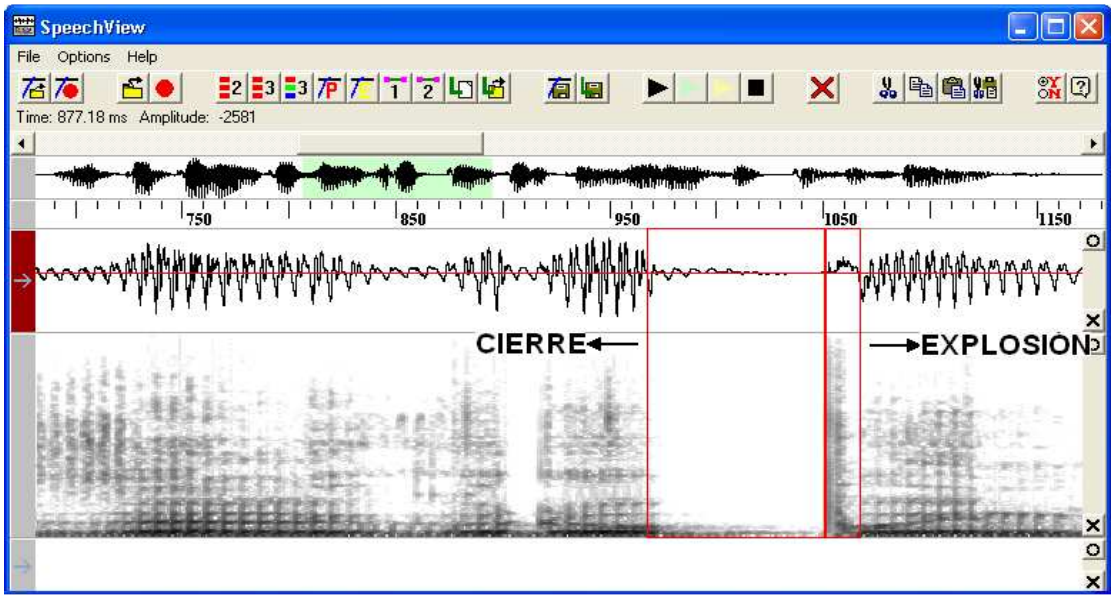

Figura 4. Imagen espectrográfica del cierre y la explosión de [p].

A continuación se mostrará el proceso de segmentación y transcripción de [p]. Cabe mencionar que la segmentación y la transcripción de los alófonos sólo se hace manualmente en el nivel T54, ya que en los niveles T44 y T22 se hacen automáticamente, por medio de un segmentador y un transcriptor. Sin embargo no se dejarán de mencionar aquí ambos procesos correspondiente para dichos niveles.

Para delimitar [p], es necesario hacer tres segmentaciones. A continuación, se muestra paso a paso cada una de ellas junto con la transcripción.

1. La primera segmentación se hace donde comienza el *cierre*. En el oscilograma se puede notar porque la frecuencia fundamental (F0) cesa o hay un cambio brusco en la onda. En el espectrograma se puede notar por el comienzo del cierre (el espacio en blanco); también se puede tomar como referencia el final de la energía del alófono anterior (Lander 1997:46). En la Figura 5 se muestra dónde ha sido hecha la primera segmentación en los tres niveles de transcripción.

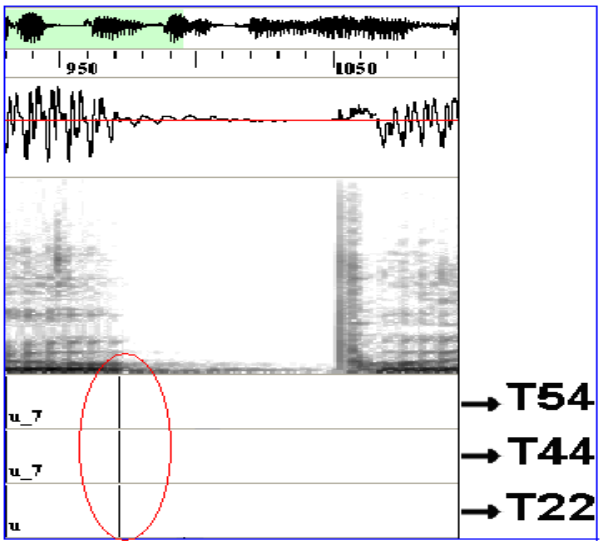

Figura 5. Imagen espectrográfica y segmentación de [p].

2. La segunda delimitación se hace donde termina el cierre, o bien, al comienzo de la barra de explosión. A propósito de esta segmentación, Lander (1997:45) señala que "When labelers see movement in the waveform begin they should set the left boundary of the burst label". Esta misma segmentación la propone para las consonantes africadas y vibrantes, las cuales también constan de intervalos de alta y baja frecuencia, como se verá más adelante (§ 2.1; §10.2; §10.3). En la Figura 6 se puede observar que la segmentación ha sido hecha justo al inicio de la barra de explosión. Si se observa el oscilograma, la segmentación se hizo al comienzo de la onda acústica, tal como señala Lander.

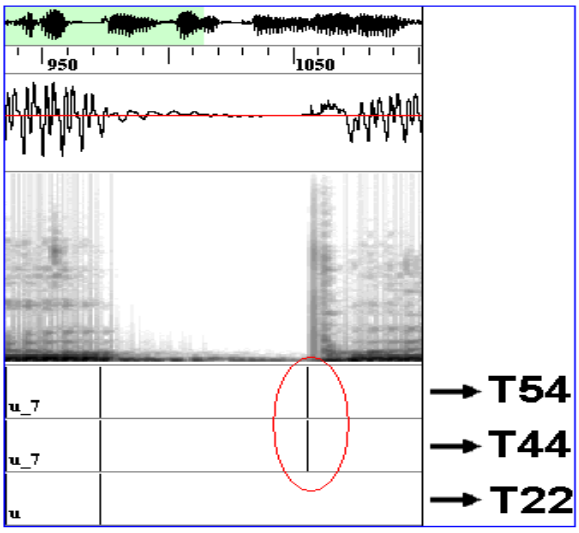

Figura 6. Imagen espectrográfica y segmentación de [p].

3. Las etiquetas que han sido creadas, con las segmentaciones anteriores en los niveles T54 y T44, corresponden al segmento del cierre. En ellas se debe transcribir el símbolo [p_c], que es el símbolo para marcar el cierre de [p]. En la Figura 7 se puede observar la transcripción.

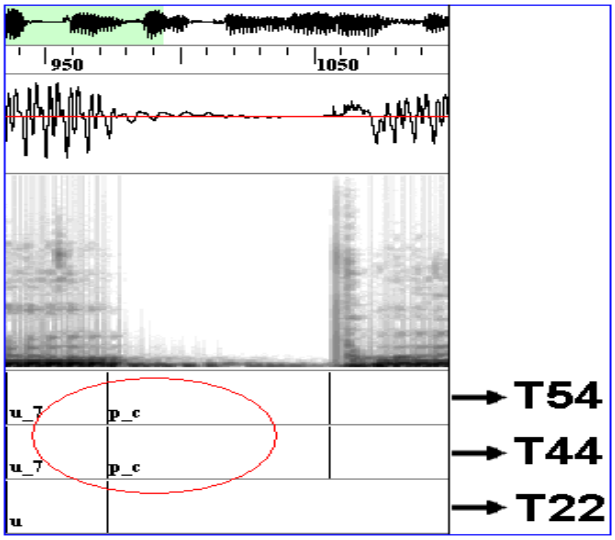

Figura 7. Imagen espectrográfica y transcripción de [p].

4. La tercera y última delimitación del alófono [p], se hace al final de la barra de explosión, o bien, donde comienza la vibración, sonoridad o frecuencia del siguiente alófono. Croot y Taylor (1995:§3.1) señalan que esta segmentación "is determined by the initial boundary of the following segment". En la Figura 8 se pude observar dónde ha sido hecha la delimitación de la barra de explosión para [p].

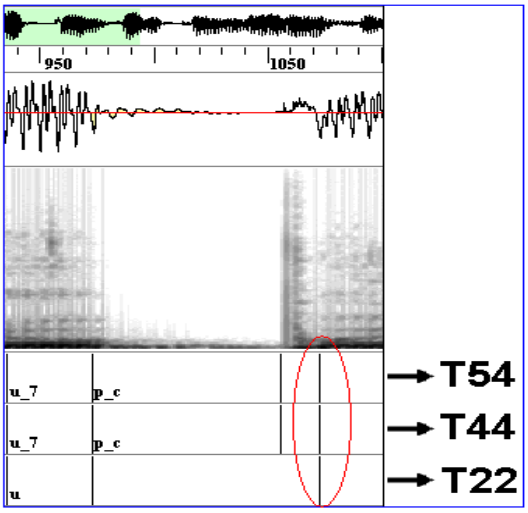

Figura 8. Imagen espectrográfica y segmentación de [p].

5. Finalmente, en las etiquetas creadas con la última segmentación se coloca el símbolo [p], que corresponden a las segundas etiquetas de los niveles T54 y T44. Para la segmentación de [p] en nivel T22, se llevará a cabo el primer y último paso; y en la etiqueta se transcribirá el símbolo [p], ya que este nivel cuenta únicamente con los fonemas básicos. En la Figura 9 se muestra la imagen espectrográfica de [p], en donde se resalta en la transcripción en las etiquetas de los tres niveles.

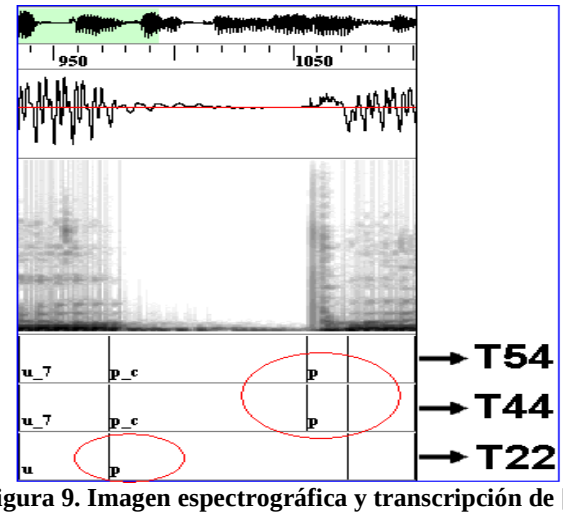

Figura 9. Imagen espectrográfica y transcripción de [p].

Como se dijo anteriormente (§1.2.1), cuando [p] está en posición final de sílaba o de coda silábica, ocasionalmente, puede tomar la realización del alófono [V], como por ejemplo *apto* ~ abto, [apto] ~ [aVto]. De la misma manera, [V] puede realizarse como [p], en

posición de coda silábica; por ejemplo *objeto* ~ opjeto, [oVxeto] ~ [opxeto]. Cuando [p] se encuentre en este contexto, suceda la alternancia o no, la segmentación en el nivel T44 se hará de la misma manera que en el nivel T22 y se transcribirá en la etiqueta el símbolo [-B]. La segmentación y la transcripción para los niveles T54 y T22 se hará de la misma forma que en [p] al inicio de sílaba. En la Figura 10 se puede observar el espectrograma y transcripción de la palabra *adoptar*, en ella se ha resaltado la segmentación y transcripción de la coda [-B] en el nivel T44.

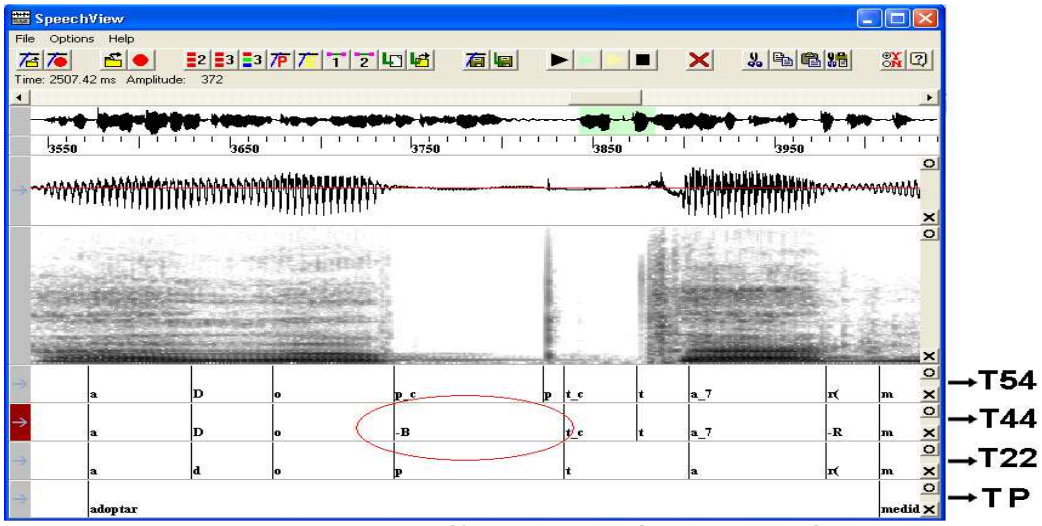

Figura 10. Imagen espectrográfica, segmentación y transcripción de [-B].

1.4.2. Fonema dental oclusivo sordo /t/

En la Figura 11 se muestra la regla de realización del fonema /t/, conforme a Cuétara 2004.

Fonema	Alófono	Contexto	Grafía
Dental oclusivo sordo /t/	Dental oclusivo sordo [t_c] [t]	Inicio de sílaba	*t*

Figura 11. Reglas distribucionales de los alófonos del fonema /t/.

La imagen espectrográfica de este alófono, al igual que [p], tiene un segmento de *cierre* y otro de *explosión*. Para la segmentación de este alófono remitirse a [p]. Las etiquetas correspondientes para los segmentos de [t] son [t_c] y [t] respectivamente. En la Figura 12 se resalta la imagen espectrográfica de este alófono junto con su transcripción.

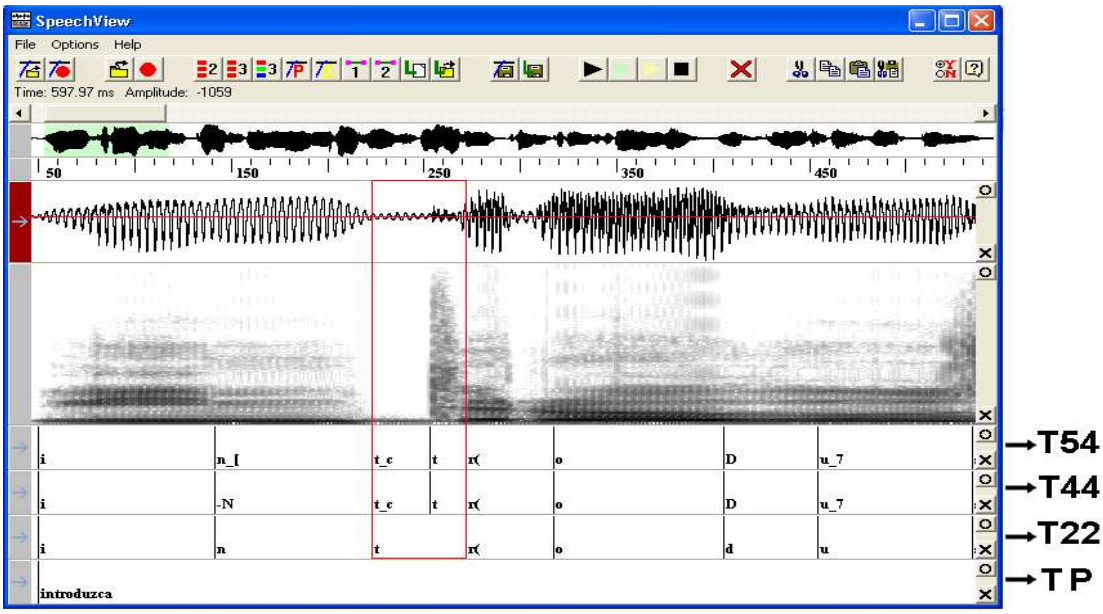

Figura 12. Imagen espectrográfica, segmentación y transcripción de [t_c] [t].

Cuando [t] está en posición de coda silábica puede alternar su realización con [D]. Por ejemplo *atmósfera* > *admósfera*, [atmO_7sfer(a] > [aDm_O_7sfer(a]; y viceversa, [D] puede alternar su realización con [t]. Este fenómeno se trasncribe con el símbolo [-D], en el nivel T44. Para la segmentación y transcripción ver [-B] en [p] en posición de coda silábica. En la Figura 13 se resalta la transcripción de [-D], en la palabra *salud*.

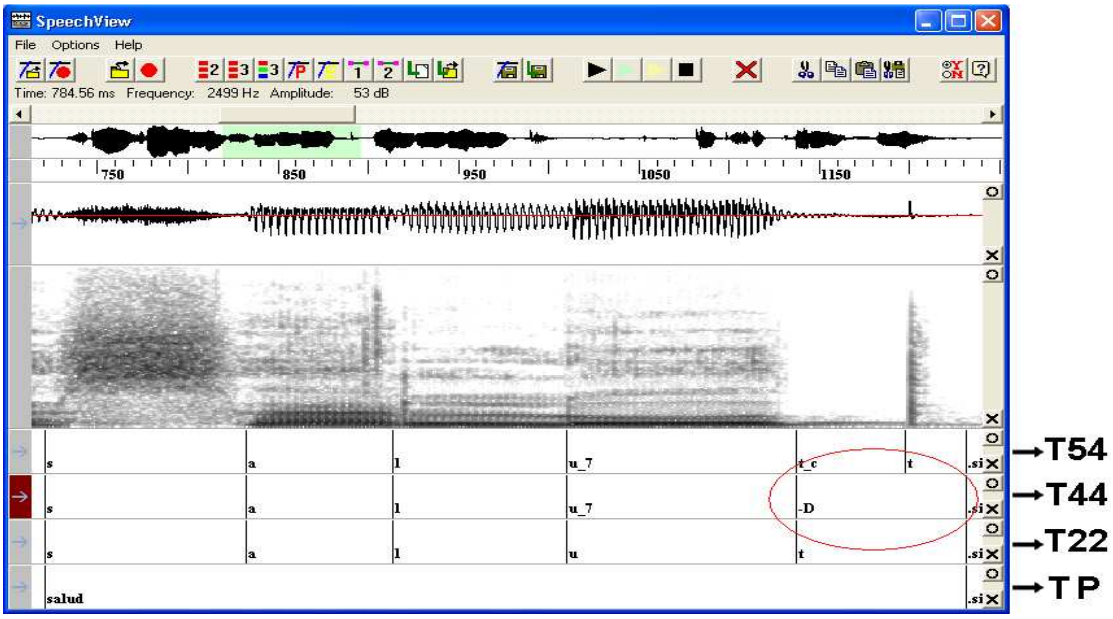

Figura 13. Imagen espectrográfica, segmentación y transcripción de [-D].

1.4.3. Fonema velar oclusivo sordo /k/

El fonema velar oclusivo sordo tiene dos alófonos: uno velar oclusivo sordo [k], que ocurre al inicio de sílaba; y el palatal oclusivo sordo, que se realiza cuando se encuentra en contexto anterior a [e, i, j] (Cuétara 2004:95). Articulatoriamente la diferencia entre estas dos formas alofónicas de /k/ es que [k] se articula en el velo de paladar y [k_j] en el paladar duro. En la Figura 14 se muestra las reglas distribucionales del fonema /k/, conforme a Cuétara 2004.

Fonema	Alófono	Contexto	Grafía
Velar oclusivo sordo /k/	Velar oclusivo sordo [k_c] [k]	Inicio de sílaba	*k* *c + a, o, u*
	Palatal oclusivo sordo [k_c] [k_j]	Anterior a [e, i, j]	*k* *qu + e, i,*

Figura 14. Reglas distribucionales de los alófonos del fonema /k/.

1.4.3.1. Alófono velar oclusivo sordo [k]

El proceso de segmentación de la imagen espectrográfica para los dos alófonos de /k/ es el mismo que se hizo para [p] (ver [p]). Las etiquetas correspondientes para el alófono velar oclusivo [k], en los niveles de transcripción T54 y T44 son [k_c] y [k]; mientras que para el nivel T22 es [k]. En la Figura 15 se puede ver la imagen espectrográfica, la segmentación y la trascripción de este alófono, en los tres niveles de transcripción.

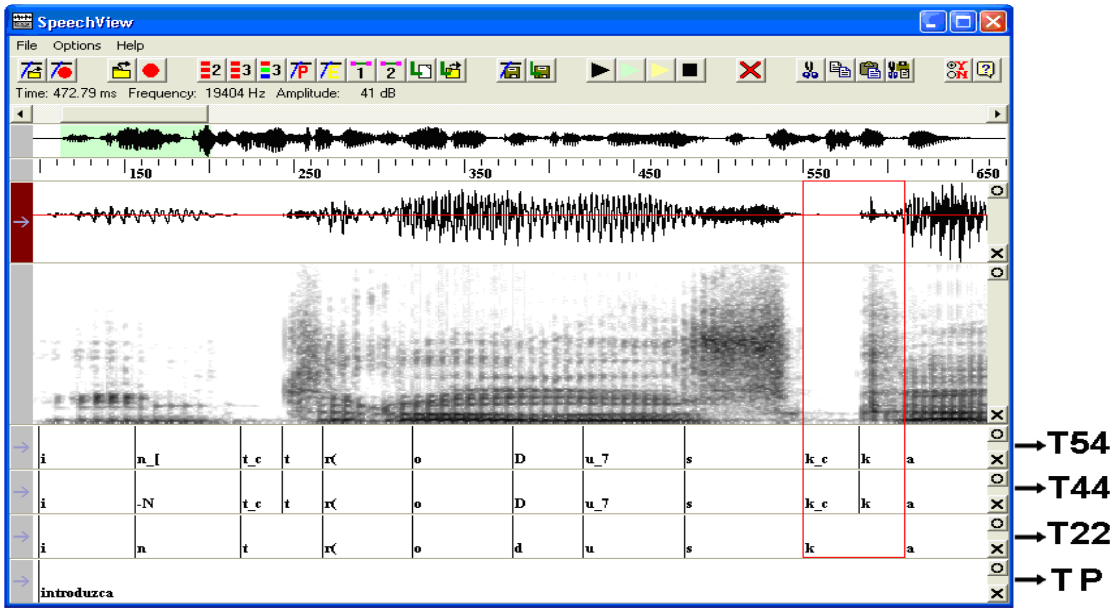

Figura 15. Imagen espectrográfica, segmentación y transcripción de [k_c] [k].

1.4.3.2. Alófono palatal oclusivo [k_j]

Las etiquetas para el alófono palatal oclusivo [k_j] son diferentes para cada nivel de transcripción: en el nivel T54 se transcriben en las etiquetas [k_c] y [k_j] respectivamente; en el nivel T44 [k_c] y [k]; y para el nivel T22 es [k]. En la Figura 16 se puede observar la transcripción de [k_j] en los tres niveles de etiquetado. Es preciso recordar que /k/ se palataliza por influencia de los alófonos vocálicos palatales [e, i, j], únicamente cuando está

en posición anterior a ellos, como es el caso de la palabra *quiero* [k_jje_7r(o] que se muestra en la Figura 16.

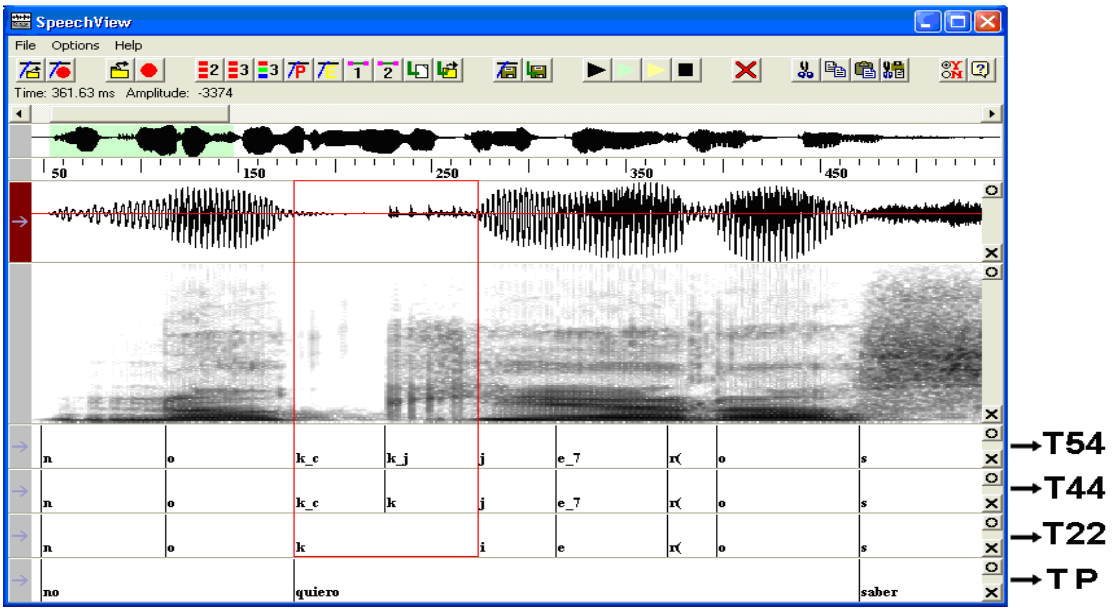

Figura 16. Imagen espectrográfica, segmentación y transcripción de [k_c] [k_j].

Espectrográficamente, podemos identificar y diferenciar al alófono [k_j] de las demás oclusivas sordas, porque la transición de su segundo formante es ascendente con respecto al de la vocal que lo precede: "Este alófono se refleja en el espectrograma con una elevación en la transición del segundo formante de la vocal, señalando la palatalización en la realización fonética de /k/ ante vocales palatales" (Cuétara 2004:96). En la Figura 17 se muestra la imagen espectrográfica de [k_j], en la cual se señala dicha transición formántica. En la Figura 18 se puede observar un dibujo de los formantes, con el propósito de mostrar con más precisión dicha transición.

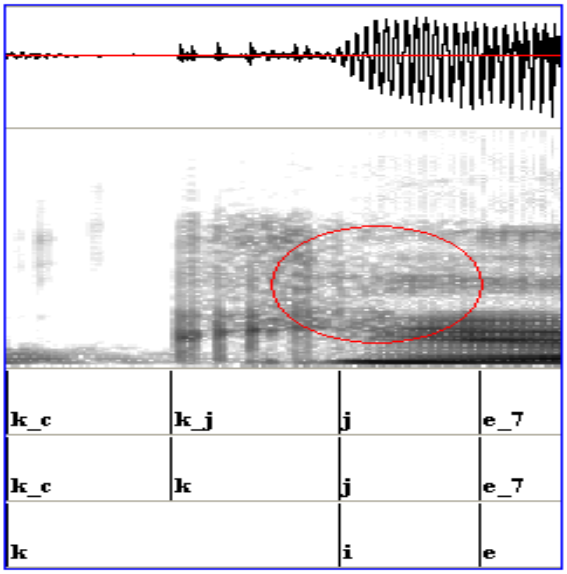

Figura 17. Imagen espectrográfica de la transición del segundo formante vocálico en contacto con la consonante [k_j].

Figura 18. Dibujo de la transición del formante vocálico en contacto con [k_j].

En posición de coda silábica, [k] tiende a alternar su realización con [G], y viceversa, cuando [G] está en dicha posición también tiende alternar con [k]. Esta alternancia en el nivel T44 se transcribe con el símbolo [-G]. Para la segmentación de [k] en posición final de sílaba ver [-B]. En la Figura 19 se señala la transcripción de [k] en el nivel T44 con la palabra *conflicto* [konfli_7kto].

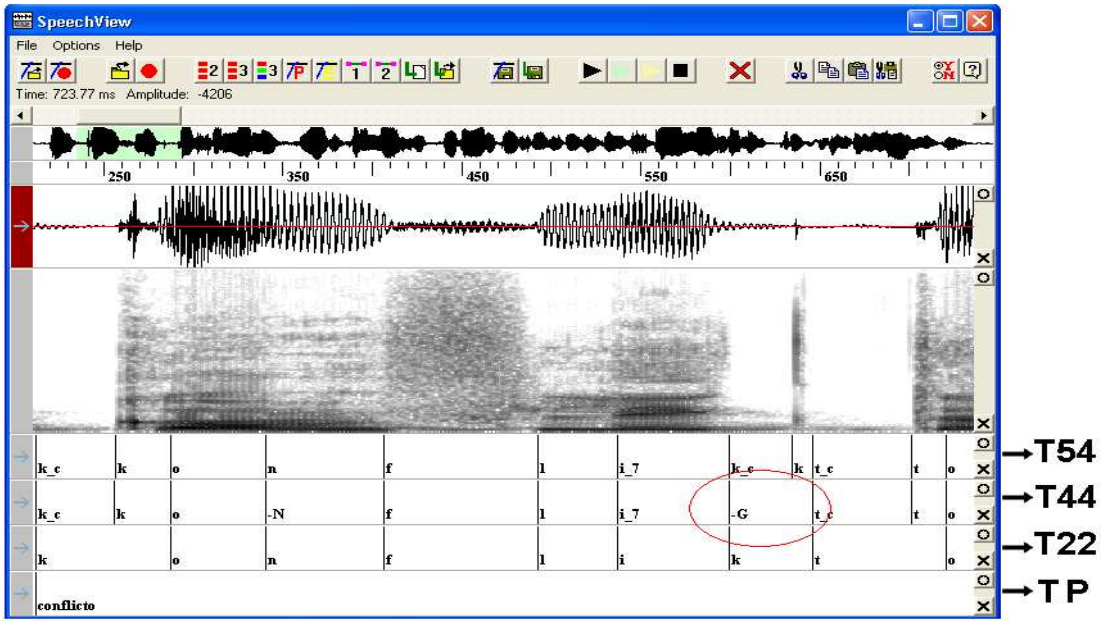

Figura 19. Imagen espectrográfica, segmentación y transcripción de [-G].

1.5. Fonemas oclusivos sonoros /b/, /d/, /g/

1.5.1. Fonema bilabial oclusivo sonoro /b/

El fonema bilabial oclusivo sonoro tiene dos alófonos: uno bilabial oclusivo sonoro, [b], que ocurre en inicio absoluto de sílaba y en contexto posterior nasal; y otro aproximante bilabial fricativo sonoro, [V], que ocurre al interior y final de sílaba. En la Figura 20 se muestra el cuadro con las reglas combinatorias del fonema bilabial, conforme a Cuétara 2004. Posteriormente, se muestra el proceso de segmentación y transcripción de cada uno de sus alófonos.

Fonema	Alófono	Contexto	Grafía
Bilabial oclusivo sonoro /b/	Bilabial oclusivo sonoro[b_c] [b]	En inicio absoluto y posterior a nasal [m, n]	*b, v*
	Bilabial fricativo sonoro [V]	En interior y final de sílaba	*b, v*

Figura 20. Reglas distribucionales de los alófonos del fonema /b/.

El modo de articulación de los alófonos oclusivos sonoros se realiza como el de los oclusivos sordos, por lo tanto la imagen acústica será similar. Sin embargo, como se dijo anteriormente (§1), los oclusivos sonoros se pueden diferenciar de los sordos, porque en el segmento del cierre presentan una *barra de sonoridad*; es decir, la frecuencia fundamental no desaparece como en los sordos. En la Figura 21 se señala la *barra de sonoridad* de una oclusiva sonora en posición inicial de sílaba.

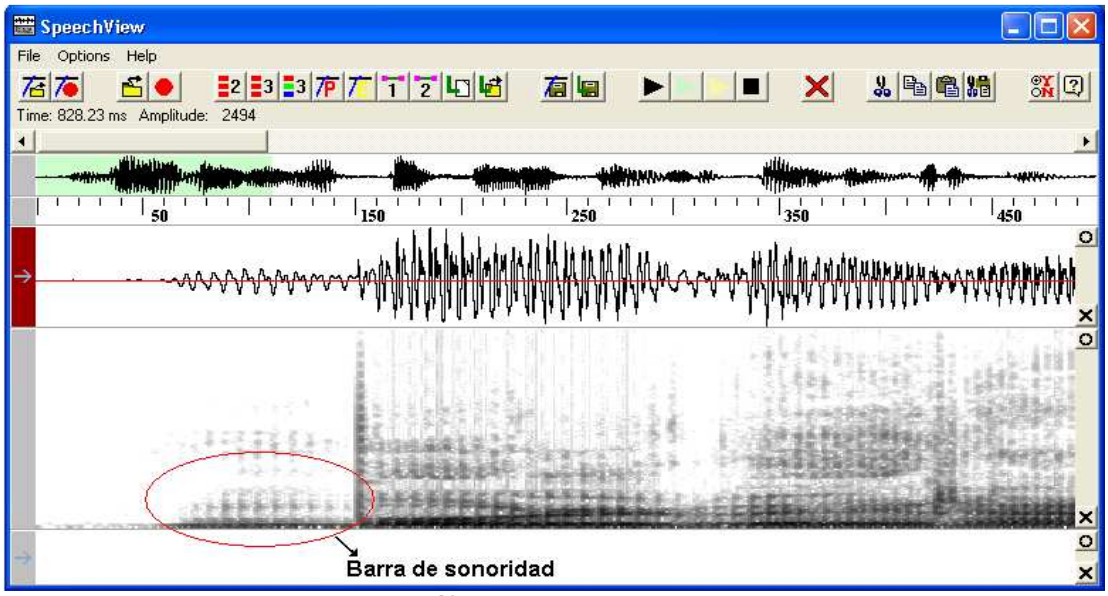

Figura 21. Imagen espectrográfica de la barra de sonoridad en las oclusivas sonoras.

1.5.1.1. Alófono bilabial oclusivo sonoro [b]

Para el proceso de segmentación del alófono bilabial sonoro remitirse a [p]. La transcripción para las etiquetas del *cierre* y la *explosión* son [b_c] y [b] en los niveles T54 y T44. Para el nivel T22 es [b]. En la Figura 22 se pueden observar la imagen y la transcripción de la palabra *volver*, en la cual se resalta el segmento correspondiente a [b].

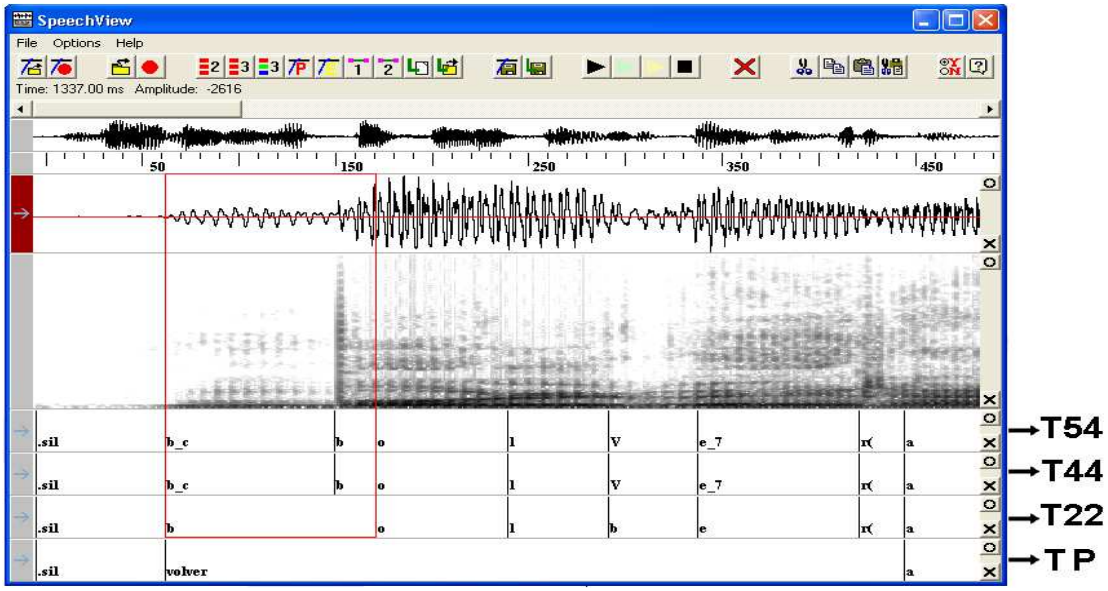

Figura 22. Imagen espectrográfica, de segmentación y etiquetado de [b] en posición inicial de sílaba.

Cuando [b] –o cualquier otro alófono oclusivo sonoro– se encuentra en contexto posterior a las nasales [m, n], mantiene su realización; sin embargo, la imagen de espectrográfica de [b] mostrará un *cierre* muy pequeño, o bien, no aparecerá. Lander menciona que en este caso el cierre no se etiqueta, porque no aparece o es difícil distinguirlo de la nasal. Esto sucede "When the place of articulation is the same for the nasal and the closure, part of the nasal acts as the perceived closure. The velum is closed just before the burst to allow the pressure to build up for the burst." (Lander 1997:47). Por este motivo, la segmentación y la transcripción de las oclusivas sonoras en este contexto serán diferentes. A continuación, se muestra cómo segmentar y transcribir a [b] en contexto posterior a [m], [n].

1. La primera segmentación se hace justo al comienzo de la barra de explosión. En la Figura 23 se puede observar la barra de explosión de una oclusiva, en contexto posterior a un alófono nasal. En esta figura se resalta la primera segmentación de la barra de explosión del alófono oclusivo bilabial.

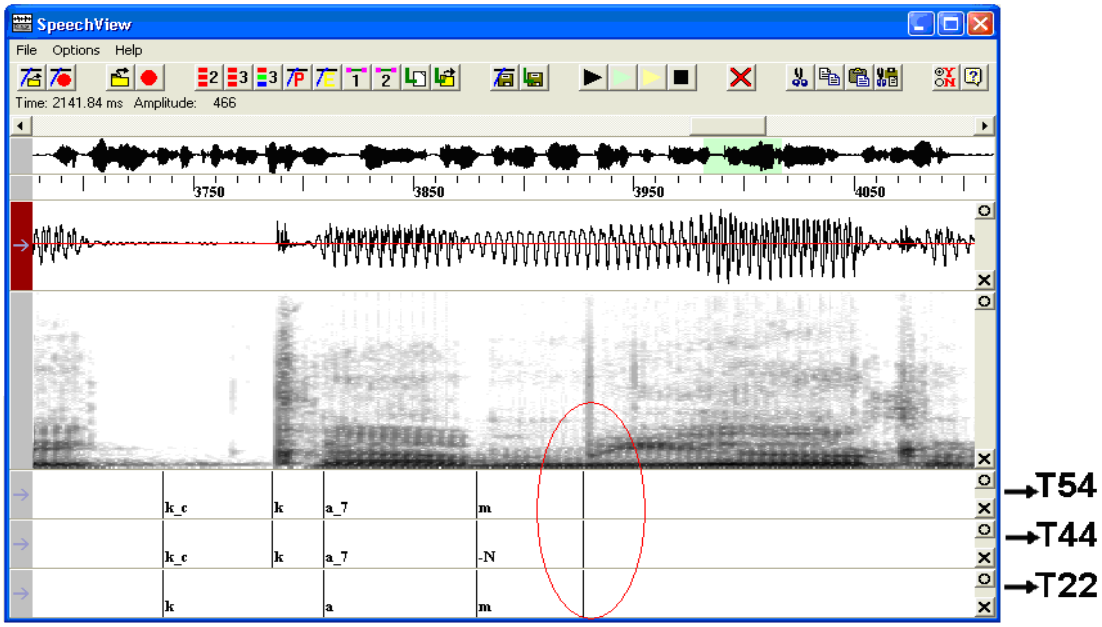

Figura 23. Imagen espectrográfica, segmentación y transcripción de [b] en contexto posterior a [m, n].

2. La segunda segmentación se hace después de la barra de explosión. Finalmente, a las etiquetas se les asigna el símbolo [b], en los tres niveles de transcripción. En la Figura 24 se puede observar la segmentación y transcripción de la oclusiva bilabial en contexto posterior a un alófono nasal.

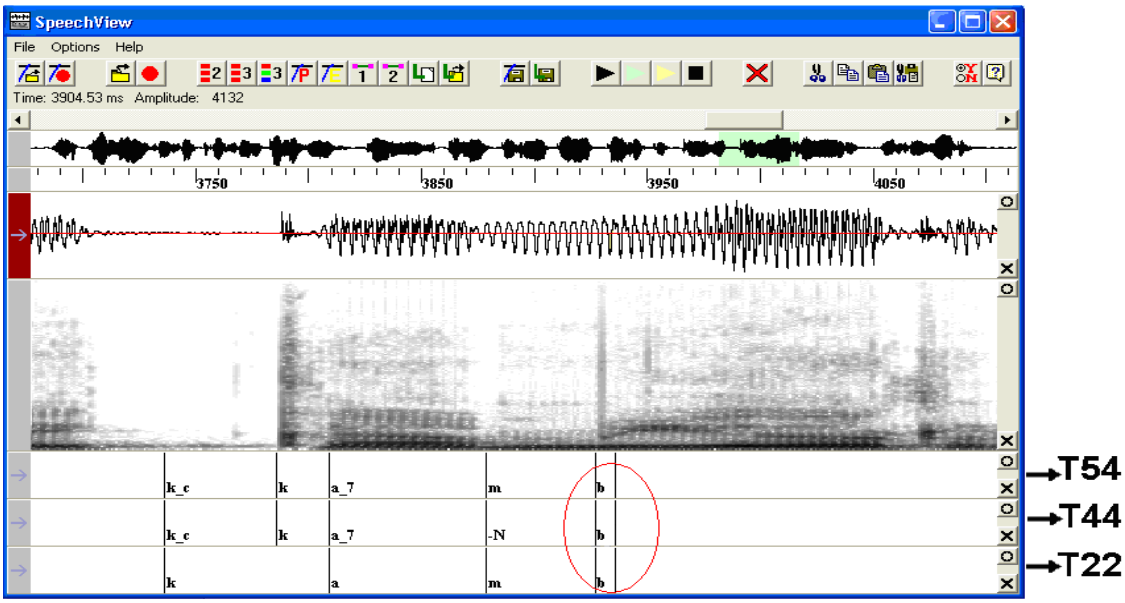

Figura 24. Imagen espectrográfica, segmentación y transcripción de [b] en contexto posterior a [m, n].

1.5.1.2. Alófono bilabial aproximante sonoro [V]

El alófono bilabial aproximante sonoro, [V], ocurre al interior y final de sílaba. Este alófono se identifica en el espectrograma porque sus formantes se perciben más bajos de energía con respecto a los alófonos que lo preceden o anteceden. En el oscilograma, se puede identificar porque la amplitud de la onda baja. En la Figura 25 se señala con un círculo la imagen espectrográfica de [V].

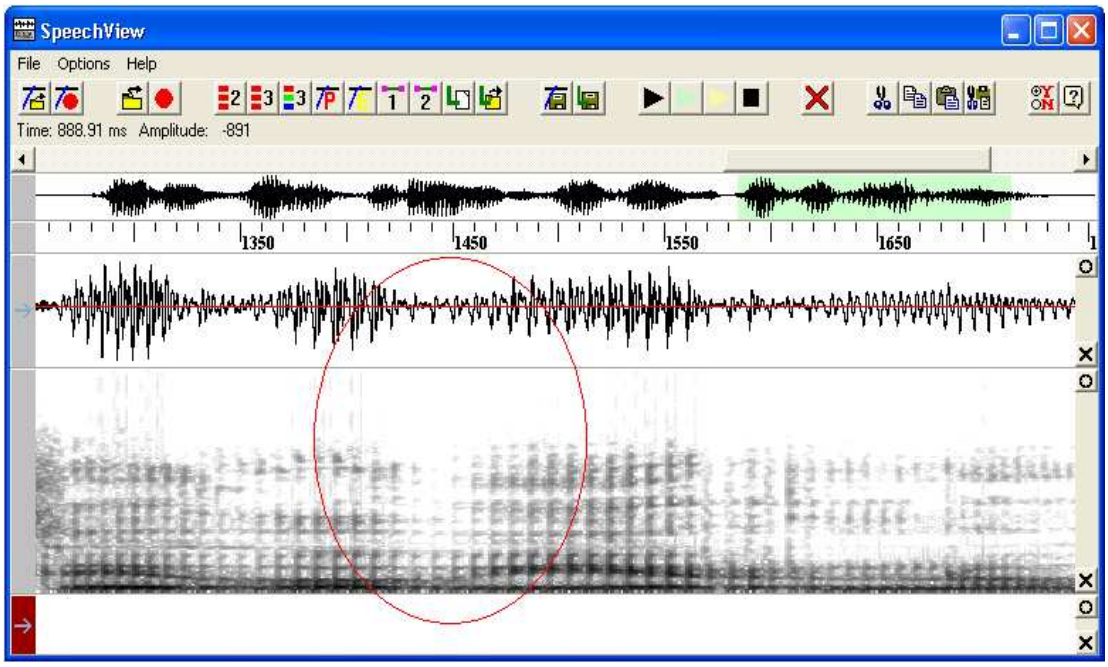

Figura 25. Imagen espectrográfica de [V].

Para la delimitación de este alófono se hacen dos segmentaciones: la primera donde hay un cambio en la frecuencia fundamental con respecto al alófono anterior. Machuca *et al.* (1999:23) sugieren que "la marca es col·loca on es produeix un canvi en l'amplitud o en la forma de l'ona; si no s'observa un canvi en l'amplitud o en la forma de l'ona, la marca es col·loca al mig de la transició". En la Figura 26 se muestra dónde se ha hecho esta primera segmentación de [V]. En este caso, la segmentación se ha hecho en donde los formantes se desvanecen.

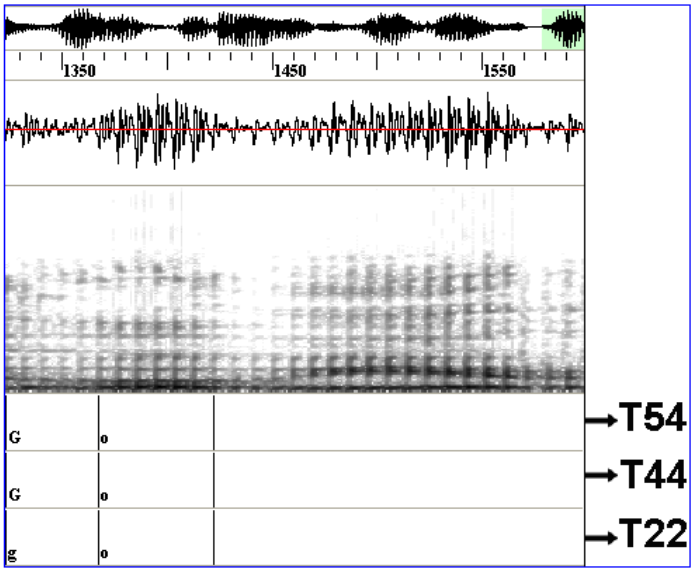

Figura 26. Imagen espectrográfica y segmentación de [V].

El segundo límite se realiza donde la frecuencia fundamental comienza a ascender nuevamente. Croot y Taylor (1995: §3.2) señalan que si el alófono posterior es sonoro, la frecuencia fundamental del segundo formante disminuirá y justo ahí se hará la segmentación: "If the preceding segment is a sonorant, the boundary is placed where energy above 500 Hz or F2 decreases". En la Figura 27 se muestra la segunda delimitación para el segmento de [V].

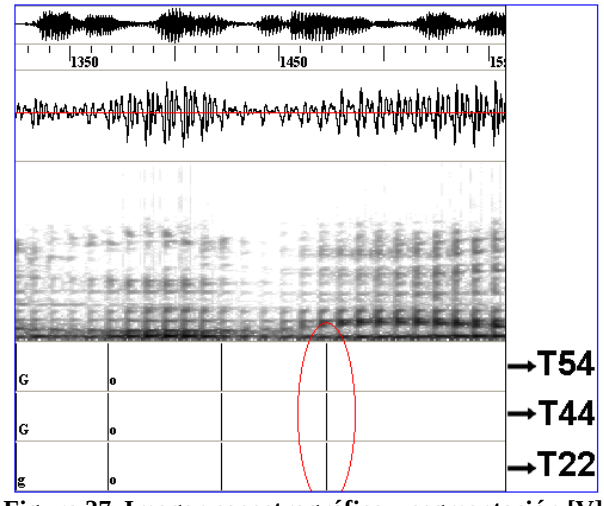

Figura 27. Imagen espectrográfica y segmentación [V].

Finalmente, la transcripción de este alófono para las etiquetas de los niveles T54 y T44 es [V]; y [b] para el nivel T22. En la Figura 28 se muestra la imagen espectrográfica de la palabra *gobierno*, en la cual se resalta la transcripción del alófono [V] en los tres niveles de transcripción.

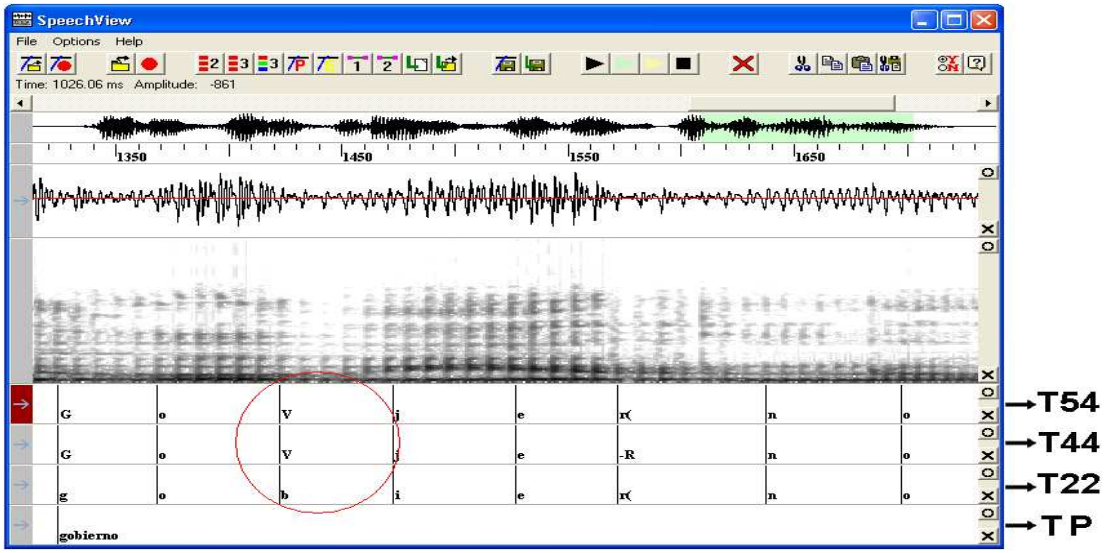

Figura 28. Imagen espectrográfica y transcripción de [V].

Como se dijo anteriormente (§1), cuando [V] está en posición de coda silábica puede alternar en su realización con [p], como en la palabra objetivo [oVxetiVo] ~ [opxetiVo]. En este caso, se transcribirá con el símbolo [-B] en la etiqueta correspondiente al nivel T44. En los otros dos niveles se colocará la etiqueta del alófono que ha sido emitido. Para el proceso segmentación remitirse a [-B] en [p] final de sílaba. En la Figura 29 se puede observar la transcripción de [V] en posición de coda silábica en la palabra *objetivo*.

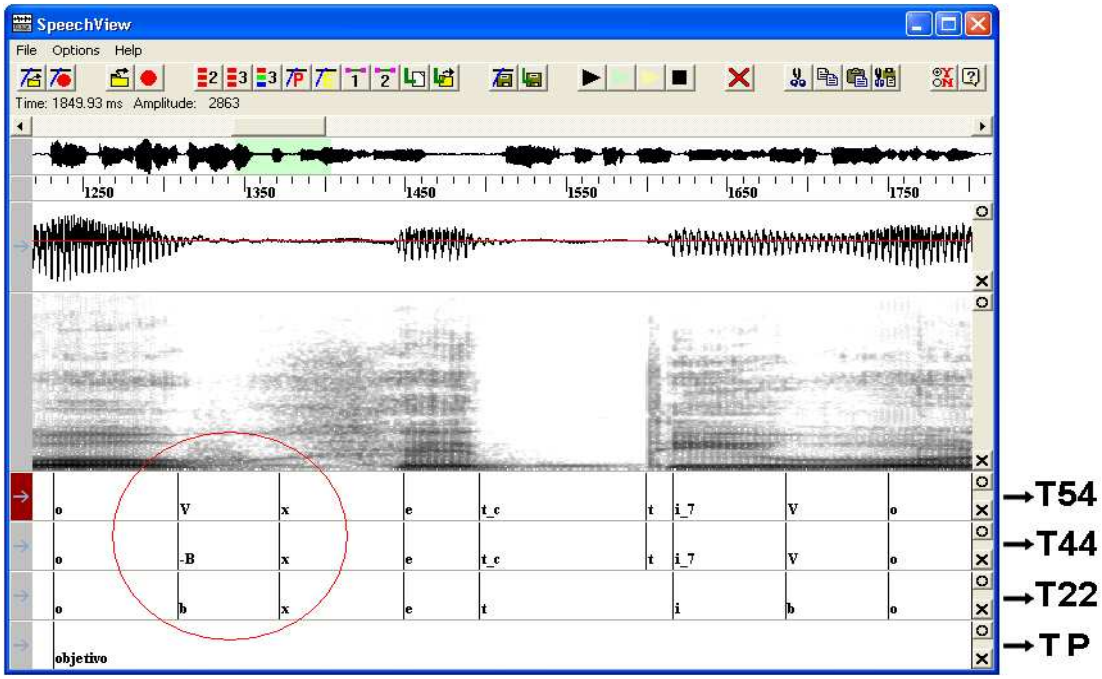

Figura 29. Imagen espectrográfica, segmentación y transcripción de [-B].

1.5.2. Fonema dental oclusivo sonoro /d/

El fonema dental oclusivo sonoro tiene dos alófonos: uno dental oclusivo sonoro que ocurre en inicio absoluto de sílaba y en contexto posterior a [m, n, l]; y otro aproximante dental fricativo sonoro que se realiza al interior y final de sílaba. En la Figura 30 aparecen las reglas combinatorias del fonema oclusivo dental, conforme a Cuétara 2004.

Fonema	Alófono	Contexto	Grafía
Dental oclusivo sonoro /d/	Dental oclusivo sonoro[d_c] [d]	En inicio absoluto, posterior a [m, n, l]	*d*
	Dental fricativo sonoro [D]	En posición interior y final de sílaba	*d*

Figura 30. Reglas distribucionales de los alófonos del fonema /d/.

1.5.2.1. Alófono dental oclusivo sonoro [d]

La imagen acústica del alófono dental oclusivo sonoro se segmenta de igual manera que [p] (ver [p]). Las etiquetas para cada segmento de [d] son [d_c] y [d] en los niveles T54 y T44, y [d] para el nivel T22. En la Figura 31 se puede observar la segmentación y transcripción de la imagen espectrográfica de [d] en contexto inicial de sílaba, en la palabra *detalle*.

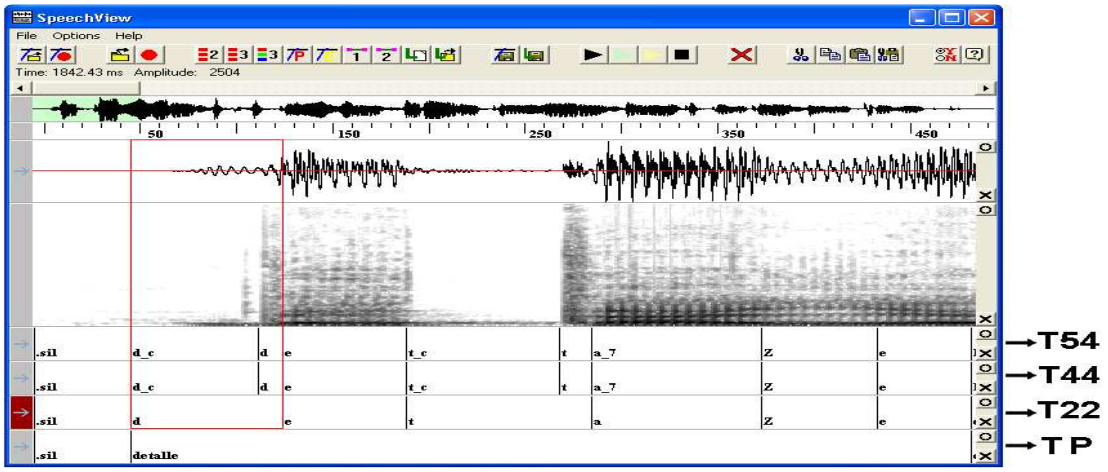

Figura 31. Imagen espectrográfica, segmentación y transcripción [d] en posición inicial de sílaba.

Cuando [d] está en contexto posterior a las nasales [m, n] y a la lateral [l], al igual que [b], no presenta cierre; por lo que la transcripción para los tres niveles será únicamente [d], como se puede observar en la Figura 32. Para el proceso de segmentación remitirse a [b] en contexto posterior a nasal.

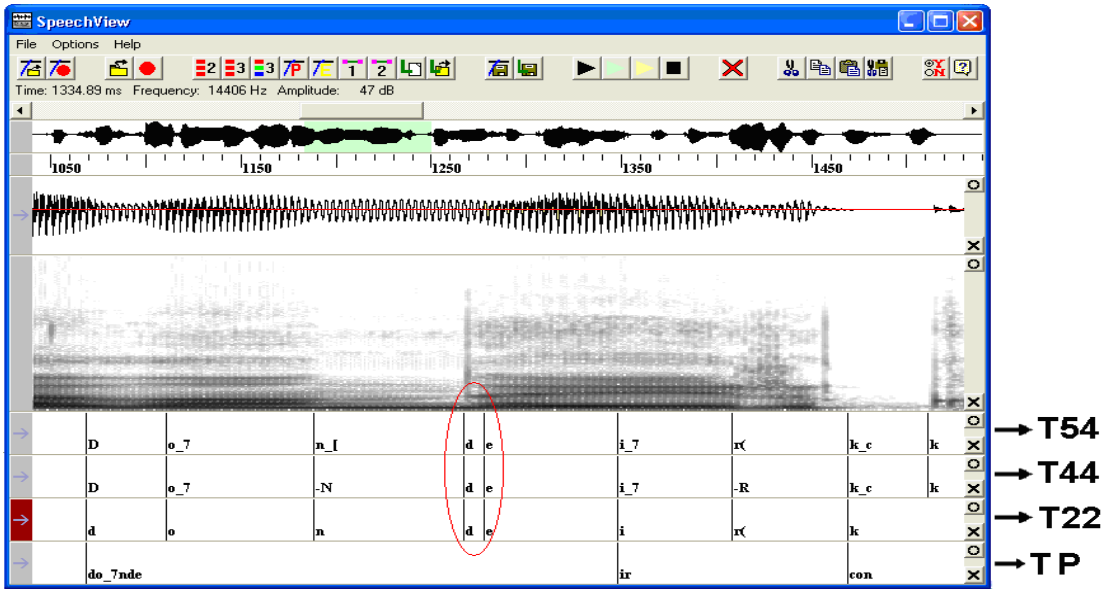

Figura 32. Imagen espectrográfica, segmentación y transcripción de [d] en contexto posterior a [m, n].

1.5.2.2. Alófono dental aproximante sonoro [D]

El alófono dental aproximante sonoro [D] ocurre en todos los demás casos en los que no se realiza el alófono oclusivo [d]. El proceso de segmentación de este alófono se hace de la misma manera que [V]. La transcripción para las etiquetas de los niveles T54 y T44 es [D]; mientras que para el nivel T22 es [d]. En la Figura 33 se puede observar la segmentación y transcripción de este alófono, en la palabra *todo*.

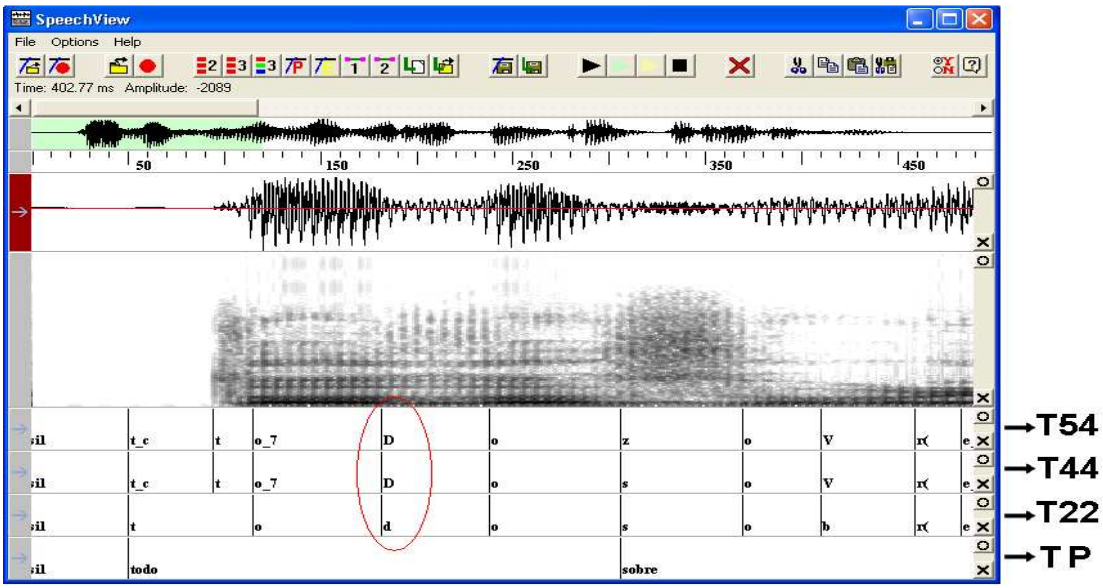

Figura 33. Imagen espectrográfica, segmentación y transcripción de [D].

Cuando [D] está en posición de coda silábica, se transcribirá con el símbolo [-D] en el nivel T44; de igual forma que [t] cuando aparece en esta posición. Para lo niveles T54 y T22 se transcribirá la grafía correspondiente al alófono que se emitió. En la Figura 34 se observa la segmentación y transcripción de [-D] en la palabra *tranquilidad*.

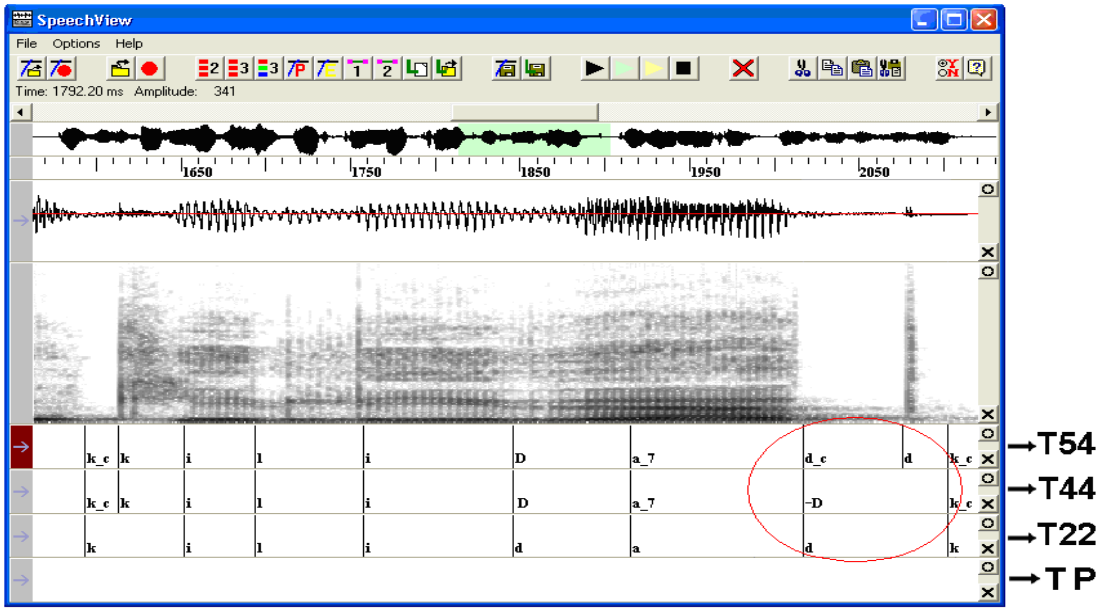

Figura 34. Imagen espectrográfica, segmentación y transcripción de [-D].

1.5.3. Fonema velar oclusivo sonoro /g/

El fonema velar oclusivo sonoro tiene dos alófonos: el velar oclusivo sonoro [g], que ocurre en los contextos de inicio absoluto de sílaba y en posición posterior a [m, n]; y el aproximante velar fricativo sonoro [G], que se realiza en cualquier posición de sílaba. En la Figura 35 se muestran las reglas distribucionales del fonema /g/, de acuerdo con Cuétara 2004. Posteriormente, se muestra la segmentación y transcripción de cada uno de sus alófonos.

Fonema	Alófono	Contexto	Grafía
Velar oclusivo sonoro /g/	Velar oclusivo sonoro [g_c] [g]	En inicio absoluto y posterior a nasal [m, n]	*g*
	Velar fricativo sonoro [G]	En cualquier posición de sílaba	*g*

Figura 35. Reglas distribucionales de los alófonos del fonema /g/.

1.5.3.1. Alófono velar oclusivo sonoro [g]

Para la segmentación del alófono velar oclusivo sonoro, [g], seguir el mismo procedimiento que en [p]. Las etiquetas para cada segmento de éste son [g_c] y [g] en lo niveles T54 y T44, para el nivel T22 es [g]. En la Figura 36 se observa la imagen espectrográfica de la palabra *gozando*, en la cual se muestra la segmentación y la transcripción del alófono velar, en los tres niveles de transcripción.

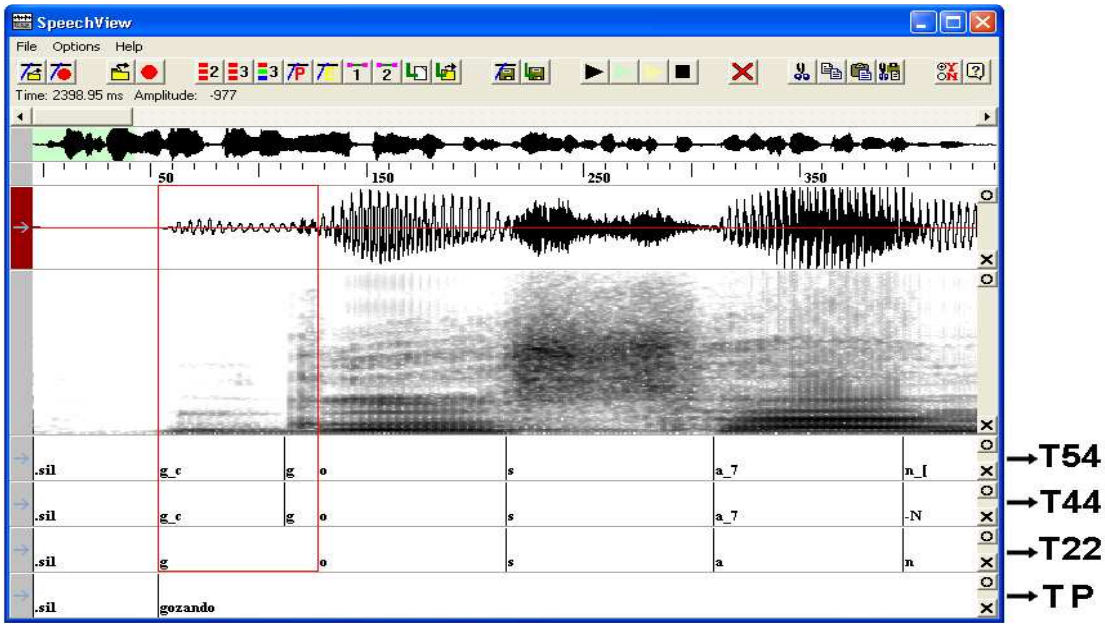

Figura 36. Imagen espectrográfica, segmentación y transcripción de [g] en posición inicial de sílaba.

El alófono velar oclusivo sonoro [g], en contexto posterior a nasal, tampoco presenta cierre; por lo que su segmentación se hace de igual forma que en [b] (ver [b] en contexto posterior a [m, n]). La transcripción es [g] para los tres niveles de etiquetado. En la Figura 37 se puede observar la segmentación y transcripción de la imagen de [g] en este contexto.

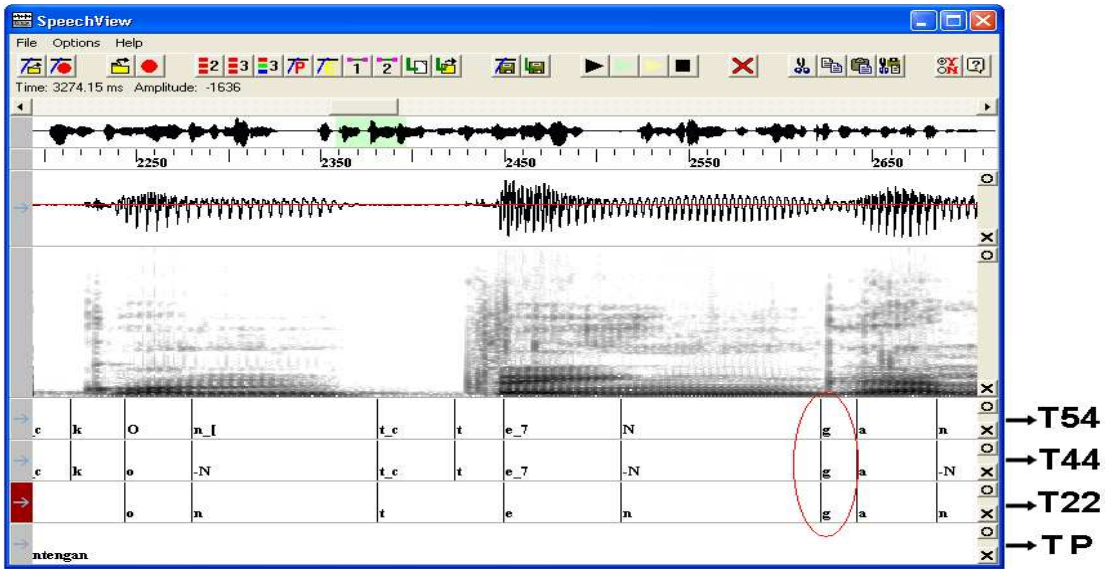

Figura 37. Imagen espectrográfica, segmentación y transcripción de [g] en contexto posterior a [m, n].

1.5.3.2. Alófono velar aproximante sonoro [G]

La imagen espectrográfica del alófono aproximante velar fricativo sonoro se reconoce por la baja energía de sus formantes. Para la segmentación de este alófono se debe seguir el procedimiento de [V]. La transcripción de este alófono es [G] para los niveles T54 y T44, y [g] para T22. En la Figura 38 se observa su transcripción en la palabra *gobierno*.

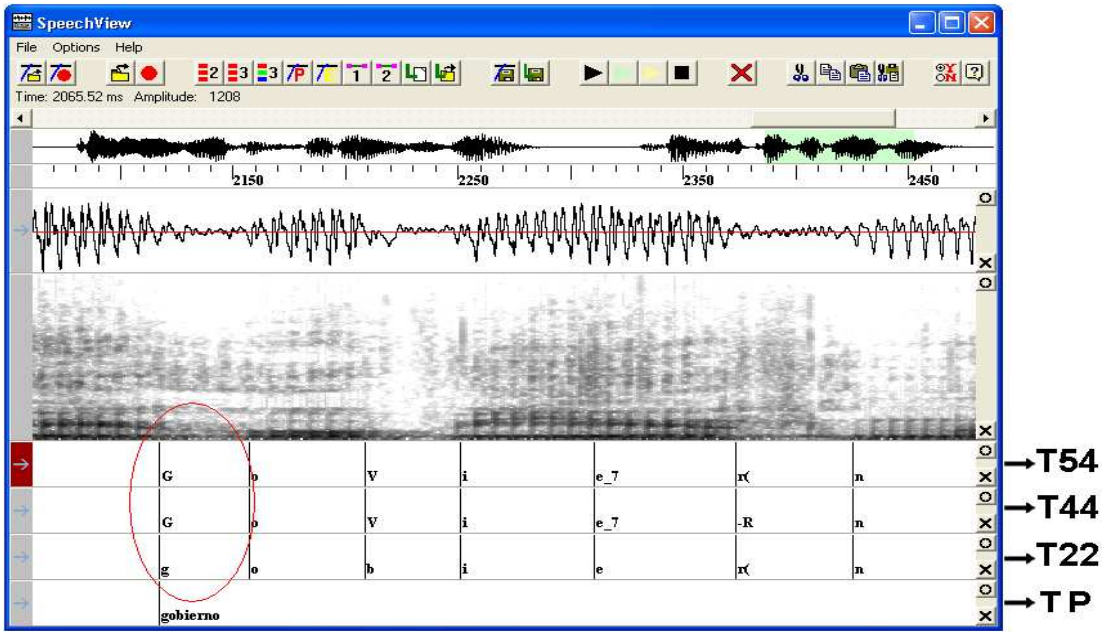

Figura 38. Imagen espectrográfica, segmentación y transcripción de [G].

Cuando [G] está en posición de coda silábica se segmenta de misma manera que [G]. La transcripción para la etiqueta del nivel T44 es [-G]; mientras que para los niveles T54 y T22 se transcribe el alófono que se emitió. En la Figura 39 se puede observar la segmentación y transcripción de [-G] en la palabra *magnetismo*.

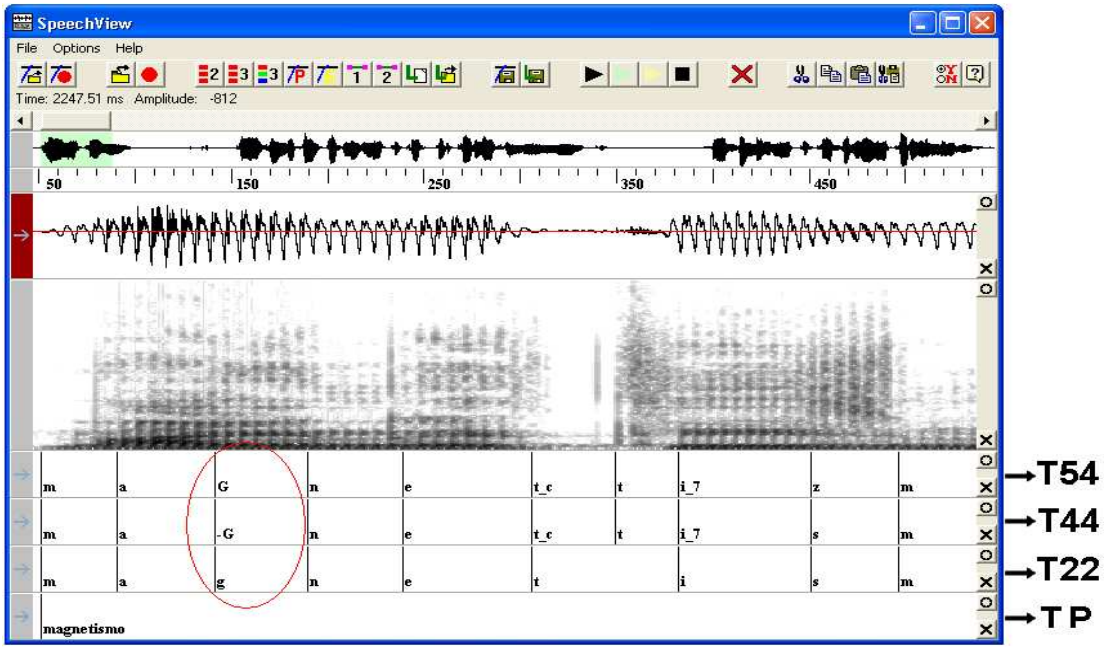

Figura 39. Imagen espectrográfica, segmentación y transcripción de [-G].

2. Fonema palatal africado sordo /tS/

En el español de la ciudad de México el fonema Africado sordo /tS/ tiene un solo alófono: el palatal africado sordo, [tS], que ocurre en posición inicial de sílaba. En cuanto a su articulación, Irribarren (2055:170) menciona que "Las africadas son resultado de la combinación de una fase oclusiva y una fase fricativa"; es decir, tiene un primer momento en el cual se cierran los órganos que lo articulan impidiendo la salida del aire, como sucede en los alófonos oclusivos. Posteriormente, dichos órganos se abren dejando salir el aire contenido por medio de una fricción. Acerca de esto, Quilis (1981:257) señala que los alófonos africados "se caracterizan porque en su emisión intervienen dos momentos: uno interrupto, similar al de las explosivas, seguido de otro constrictivo. Estos dos momentos se realizan en el mismo lugar articulatorio y, además, durante el momento de su tensión".

La articulación de este alófono se refleja en el espectrograma en dos segmentos: el primero por un cierre, como el de las oclusivas, que corresponde al momento de la obstrucción del aire; y el segundo por una fricción, que es una mancha turbulenta que corresponde al momento de la constricción. Dicha mancha es producto del aire al pasar "por el estrechamiento producido entre los órganos, en el mismo lugar donde previamente se habían adherido" (Hidalgo y Quilis 2004: 83-84).

La realización de los alófonos africados ha sido motivo de discusión, en la cual se han tomado dos posturas. Por un lado están los autores que plantean que las africadas son la unión de dos consonantes: una oclusiva más una fricativa. Ladefoged (2000:53) comenta que en inglés "An affricate is simply a sequence of a stop followed by a homorganic fricative. Some such sequences, for example the dental affricate [tθ] as in "eighth" or the alveolar affricate [ts] as in "cats", have been given no special status in English phonology". Por otro lado, están los autores que sostiene que las consonantes africadas son un sólo segmento. Quilis en su *Tratado de fonología y fonética españolas* (1999:289) comenta acerca de esta discusión y da 6 puntos por los cuales considerar a las africadas como

sonidos simples; entre ellos, retomo el primero, en cual dice que los dos segmentos de las africadas se articulan en el mismo lugar; y en el punto quinto, cita a Dauzat y Chlumsky: “El hablante que posee africadas en su lengua materna las siente, en sus emisión y en su percepción, como si fuesen consonantes simples y no compuestas”. Navarro Tomás (1918:20) también menciona que la articulación de las consonantes africadas se hace en el mismo punto de la cavidad bucal: “prodúcese en el canal vocal un contacto que interrumpe momentáneamente, como en las oclusivas, la salida del aire; después este contacto se resulte suavemente, sin transición brusca, en una estrechez; la oclusión y la estrechez se verifican en el mismo punto y entre los mismos órganos”.

En la Figura 40 se muestra la regla combinatoria del fonema /tS/, seguido del proceso de segmentación y transcripción de su alófono.

Fonema	Alófono	Contexto	Grafía
Palatal Africado sordo /tS/	Palatal Africado sordo [tS_c] [tS]	Inicio de sílaba	*ch*

Figura 40. Reglas distribucionales de los alófonos del fonema /tS/.

2.1. Alófono palatal africado sordo [tS]

Como se mencionó (§2) este alófono consta de dos momentos, el primero donde se obstruye el aire y el segundo donde es expulsado; por lo tanto, su imagen espectrográfica se caracteriza por tener dos segmentos: en el primero se muestra un cierre y en el segundo aparece una mancha turbulenta correspondiente a la fricción. Martínez Celdrán (1998:79) señala que el segmento turbulento es “un ruido tan largo o más que la zona de silencio correspondiente a la oclusión”. En la Figura 41 se señala la imagen espectrográfica que pertenece a [tS], en la cuál se puede observar ambos segmentos del alófono: el cierre y la fricción.

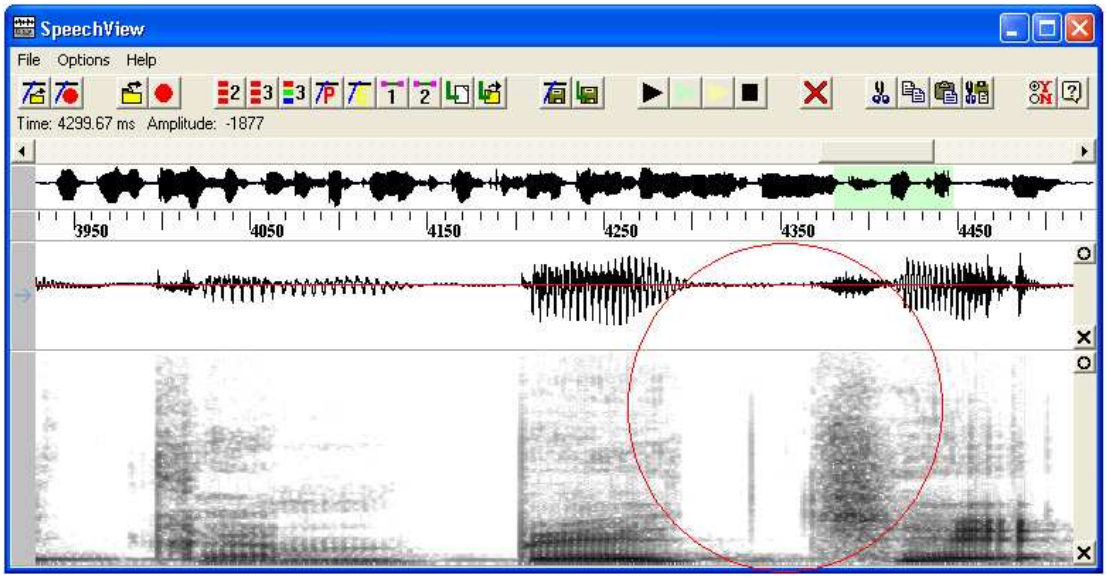

Figura 41. Imagen espectrográfica de [tS].

La segmentación de este alófono es parecida a la de las oclusivas, primero se segmenta el *cierre* (para más detalles ver [p]), y en las etiquetas de los niveles T54 y T44 se transcribe el símbolo [tS_c]. En el nivel T22 únicamente se hace una segmentación al inicio del cierre. En la Figura 42 se observa la segmentación y la transcripción del cierre del alófono palatal.

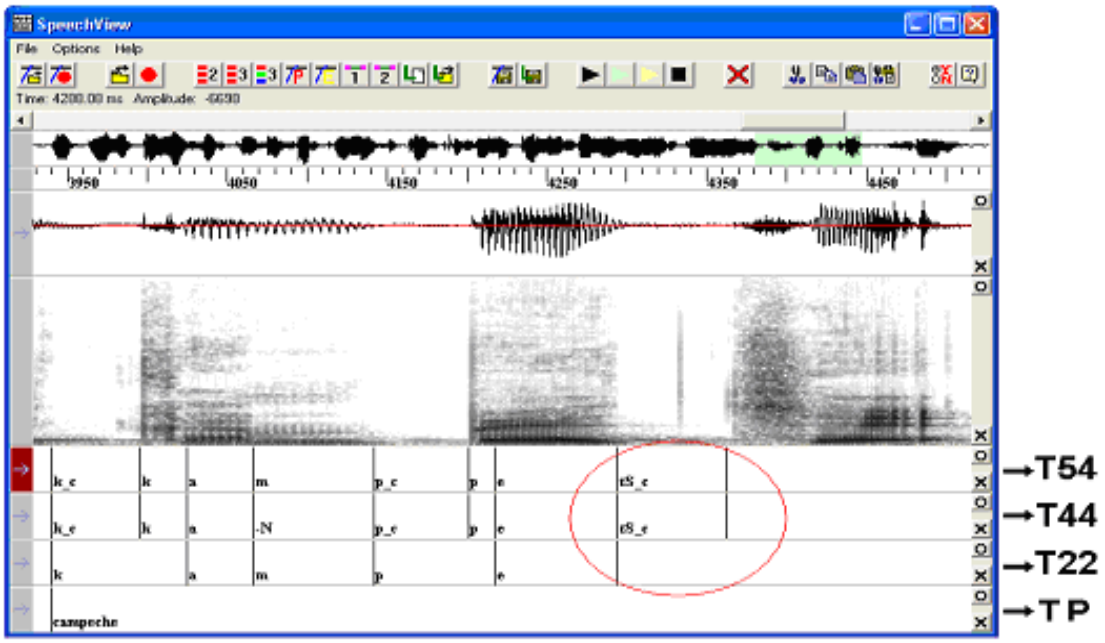

Figura 42. Imagen espectrográfica, segmentación y transcripción de [tS_c].

Posteriormente, se hace una tercera segmentación justo al final de la fricción, en los tres niveles de transcripción; y en las etiquetas se transcribe el símbolo [tS]. En la Figura 43 se muestra la imagen espectrográfica de la palabra *Campeche*, en ella se resalta la transcripción del alófono [tS].

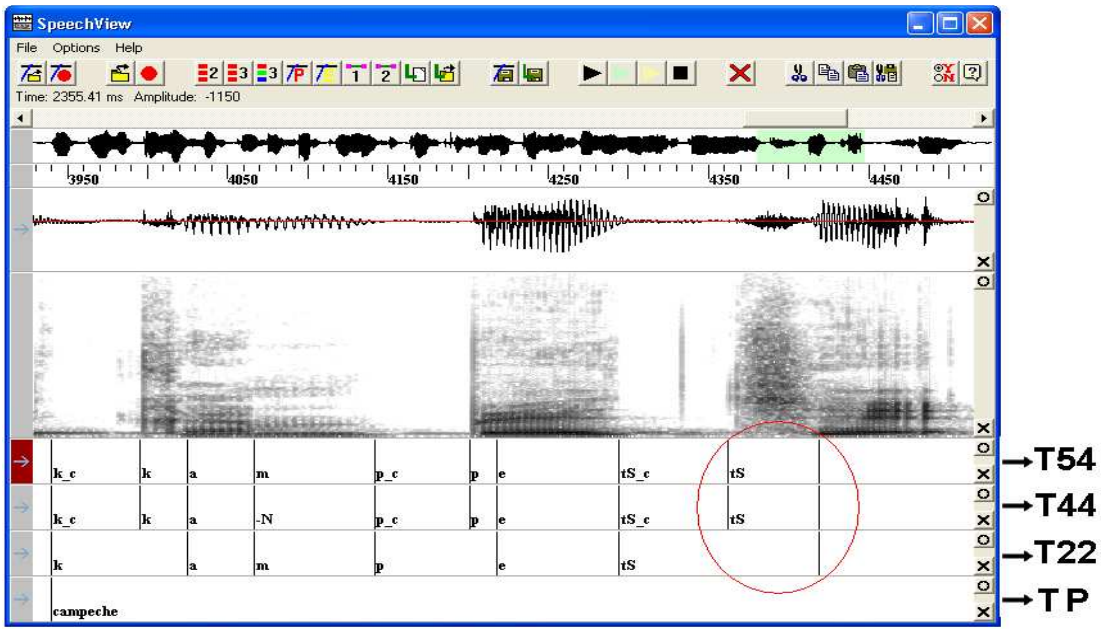

Figura 43. Imagen espectrográfica, segmentación y transcripción de [tS].

3. Fonemas fricativos /f/, /s/, /x/, /Z/

3.1 Fonema labiodental fricativo sordo /f/

El fonema fricativo sordo /f/ tiene un solo alófono en distribución complementaria, el labiodental fricativo sordo, [f], que ocurre en inicio de sílaba. Ladefoged (2001:56) explica que, a los sonidos fricativos sordos, como /f/, se les ha llamado así porque "the vocal folds are not vibrating, and fricative to indicate that the noise is produced by the friction, the resistance to the air as it rushes through a narrow gap".

Este alófono se articula con los dientes superiores apoyados en el labio inferior, provocando que el aire salga dificultosamente haciendo una fricción (Quilis 1999). Dicha fricción se refleja en el espectrograma como una mancha turbulenta. Acerca de esto, Lander (1997:49) menciona que "sibilants are the easiest to isolate because of the high energy they produce: the onset can be determined by a sudden, heavy increase of random energy in the spectrogram. Other types of fricatives are also marked by random energy in the spectrogram, but the amplitude (and visibility) may be very low in the waveform".

Los alófonos fricativos se dividen en dos grupos: los de *resonancias bajas*, que se definen así porque la mayor parte de su energía está en la parte baja del espectrograma (en este grupo aparece únicamente [Z]); y los de *resonancias altas*, que se caracterizan porque su energía se centra en la parte alta del espectrograma y por lo regular llegan a ocuparlo todo (a este grupo pertenecen [f]. [s], [x]((Quilis 1988:221). El alófono labiodental fricativo forma parte del grupo de las fricativas de *resonancias altas*, ya que en el espectrograma su imagen suele ocupar casi todo el espectrograma y se distingue por ser una gran mancha turbulenta de fricción (Hidalgo y Quilis 2004).

Cuétara (2004) reportó este alófono en el *Corpus DIME* como un alófono de baja frecuencia, el cuál no presentó muchas modificaciones; por lo que propone a [f] como único alófono para el modelado computacional.

En la Figura 44 se muestran las reglas distribucionales de /f/, de acuerdo con Cuétara 2004. Posteriormente, se mostrará el proceso de segmentación y transcripción de su alófono.

Fonema	Alófono	Contexto	Grafía
Labiodental fricativo sordo /f/	Labiodental fricativo sordo [f]	En inicio de sílaba	*f*

Figura 44. Reglas distribucionales de los alófonos del fonema /f/.

3.1.1. Alófono labiodental fricativo sordo [f]

El alófono labiodental fricativo sordo se caracteriza en el oscilograma por tener una onda débil, y en el espectrograma por ser una mancha de sonido turbulento o de fricción, la cual carece de una estructura formántica. En la Figura 45 se pude observar la imagen acústica de [f].

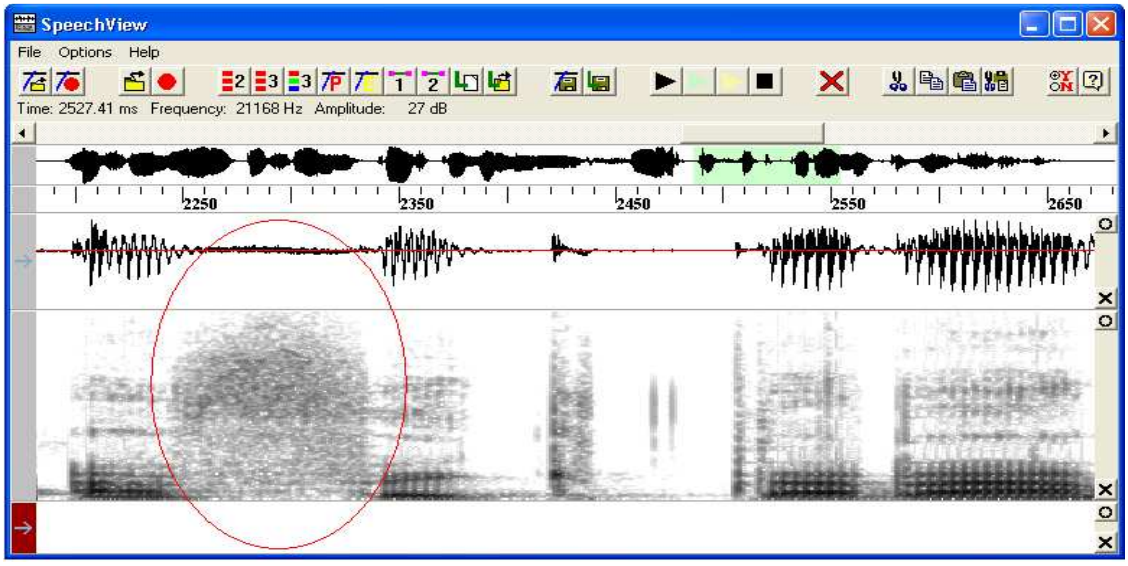

Figura 45. Imagen espectrográfica de [f].

La delimitación de este alófono se hace con dos segmentaciones. La primera donde la frecuencia fundamental decrece bruscamente, o bien, si se ve el espectrograma, se hace donde comienza la mancha de fricción. El segundo límite se coloca donde termina la fricción o donde hay un aumento en la amplitud de la onda. En la Figura 46 se muestra la imagen espectrográfica de [f]; en esta imagen se puede observar dónde han sido hechas ambas segmentaciones del alófono.

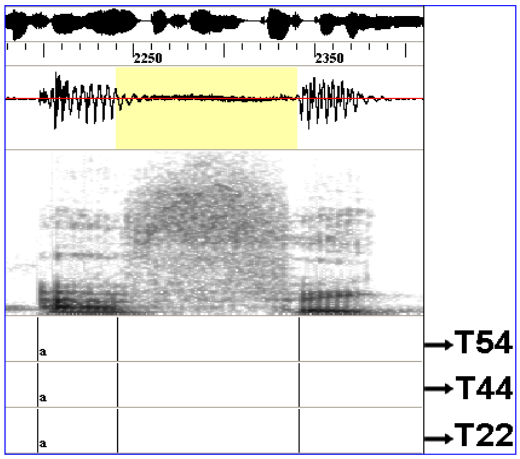

Figura 46. Imagen espectrográfica y segmentación [f].

La transcripción de este alófono es [f] para las etiquetas de los tres niveles. En la Figura 47 se muestra la imagen acústica de la palabra *afectarán*, en la cual se señala la transcripción del alófono labiodental fricativo.

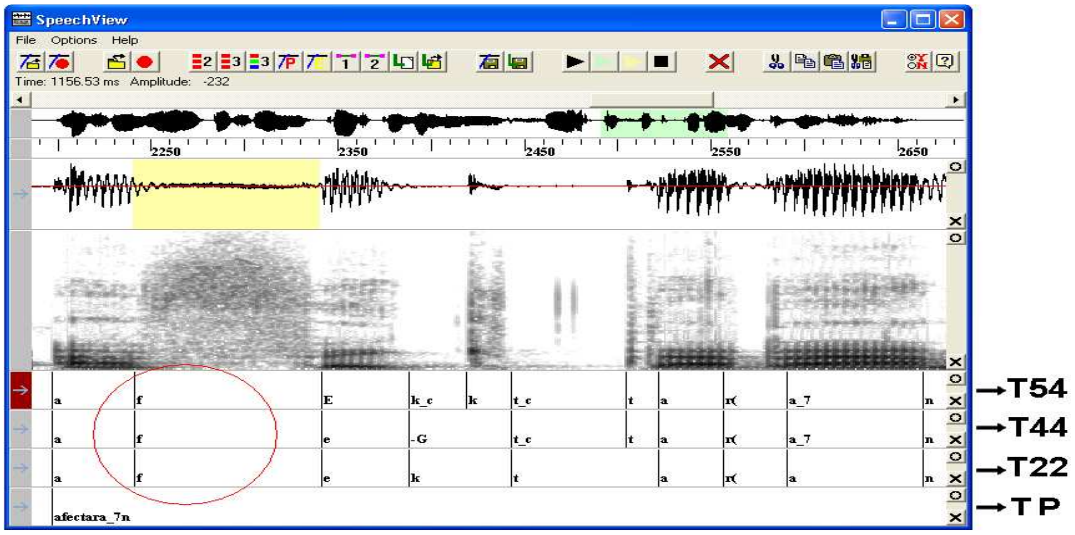

Figura 47. Imagen espectrográfica y transcripción de [f].

3.2. Fonema alveolar fricativo sordo /s/

El fonema alveolar fricativo sordo /s/ tiene alta frecuencia de aparición en el español, razón por la cual se han registrado varios cambios en él (Cuétara 2004). Moreno de Alba (1988) reportó este alófono con una alta frecuencia en América y con uso general en varios lugares de México, como el Valle de México, Jalisco, Yucatán y Oaxaca. Específicamente, en el español de la ciudad de México tiene tres alófonos en distribución complementaria: uno sordo, otro dentalizado y uno más sonoro.

El alófono alveolar fricativo sordo [s] se realiza al inicio de sílaba. Los órganos articuladores de este alófono presentan una aproximación más cercana, haciendo que la fricación sea más grande. Por ello, [s] forma parte del grupo de las consonantes fricativas de *resonancias altas*, y de ellas es la que presenta mayor energía (Quilis 1981). Lo anterior puede ser visto en su imagen acústica, la cual se caracteriza por ser una mancha turbulenta o una fricción que ocupa todo el espectrograma. Matluck (1951:72-73) menciona acerca de la pronunciación de [s], específicamente del valle de México, que "es un sonido predorso-álveodental convexo fricativo sordo, de tensión media, de timbre muy agudo y de larga duración".

El alófono dental fricativo sordo [s_[] siempre ocurre en contexto final de sílaba y precedido de [t]; razón por la cual cambia su punto de articulación a dental. Quilis (1999) sostiene que este alófono no existe en el español y no es más que la asimilación de /s/ al alófono dental. Sin embargo, para el español de la ciudad de México, Cuétara (2004) lo mantiene dentro de las realizaciones del fonema /s/, respaldándose en la diferencia de la estridencia entre la /s/ castellana y la /s/ mexicana, la cual documentan varios autores (Matluck 1951, Lope Blanch 1963-1964 y Perissinotto 1975); y en los datos acústicos que arroja la unión entre /s/ y /t/, sobre todo en los cambios de la onda sonora de /s/.

Finalmente, el fonema /s/ tiende a sonorizarse cuando está en contexto intervocálico y ante las consonantes sonoras [b, d, g, Z, m, n, n~, l, r, r(], dando como resultado la realización del alófono alveolar fricativo sonoro [z]. Para la articulación de este alófono los órganos se

juntan ligeramente, de tal manera que el aire sale casi libremente. Durante su emisión las cuerdas vocales vibran por influencia de la sonoridad de las vocales o de los alófonos sonoros que lo preceden. Este alófono pertenece al grupo de las fricativas de *resonancias bajas* porque su imagen espectrográfica se concentra en la parte baja del espectrograma y se caracteriza por tener ligeros formantes (Quilis 1999).

En la Figura 48 se muestran las reglas distribucionales del fonema /s/ y su representación en *Mexbet*, conforme a Cuétara 2004. Posteriormente, se explica el proceso de segmentación y transcripción de cada uno de sus alófonos.

Fonema	Alófono	Contexto	Grafía
Alveolar fricativo sordo /s/	Alveolar fricativo sonoro [z]	En contexto intervocálico y ante [b, d, g, Z, m, n, n~, l, r, r(]	*s, z, x*
	Dental fricativo sordo [s_[ ]	En contexto anterior a [t]	*s, z,*
	Alveolar fricativo sordo [s]	En contexto inicial de sílaba.	*s, z,x*

Figura 48. Reglas distribucionales de los alófonos del fonema /s/.

3.2.1. Alófono alveolar fricativo sordo [s]

El alófono alveolar fricativo sordo ocurre en contexto inicial de sílaba. Éste se identifica en el espectrograma por ser una gran mancha oscura, que carece de formantes y posee alta frecuencia, por lo que regularmente abarca todo el espectrograma de manera vertical. En la Figura 49 se puede observar la imagen espectrográfica de este alófono.

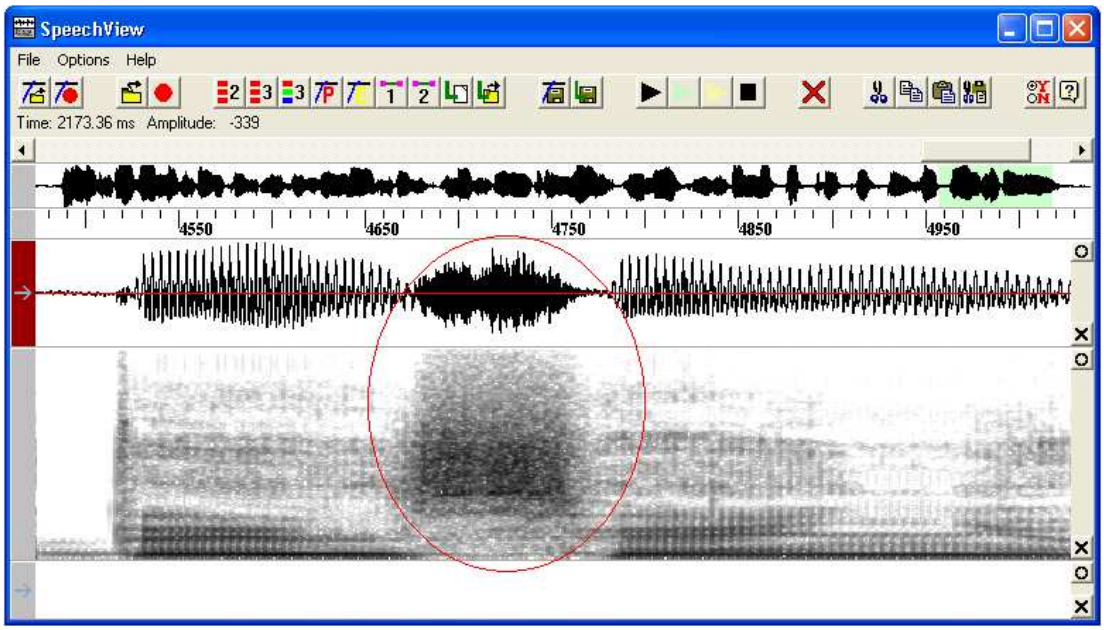

Figura 49. Imagen espectrográfica de [s].

Para segmentar este alófono se hace el mismo procedimiento que para [f]. La transcripción es [s] para las etiquetas de los tres niveles. En la Figura 50 se observa la imagen acústica de la palabra *enseño*, en la cual se resalta la segmentación y transcripción de [s].

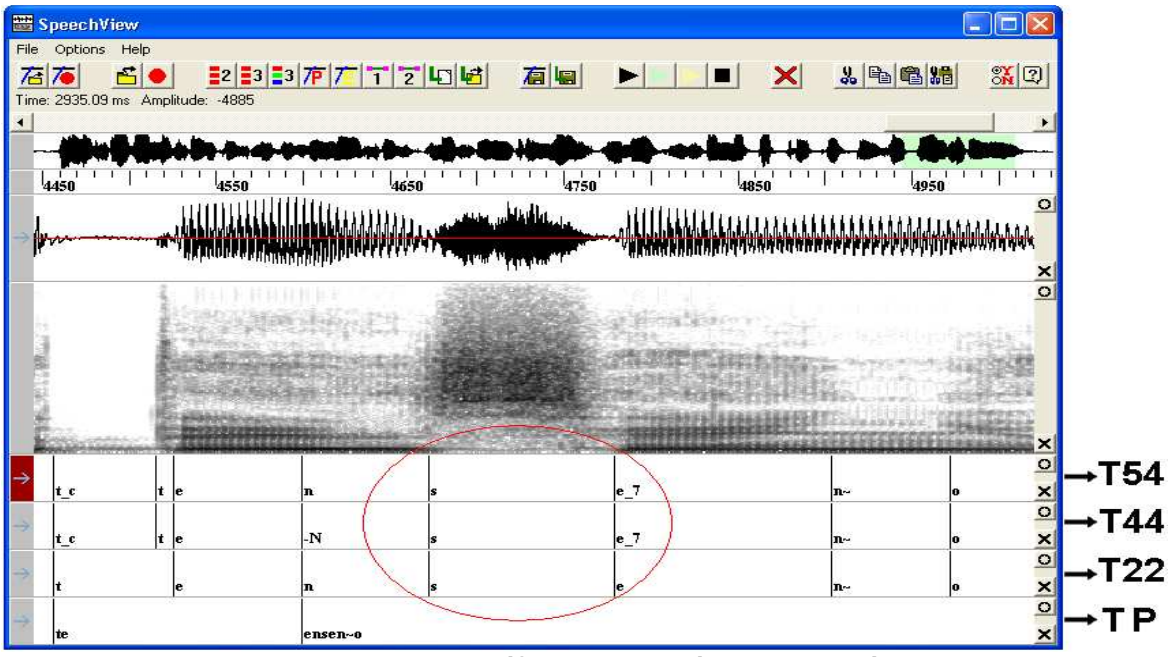

Figura 50. Imagen espectrográfica, segmentación y transcripción de [s].

3.2.2. Alófono dental fricativo sordo [s_[]

El fonema alveolar fricativo sordo /s/ se dentaliza cuando está en contexto anterior a [t], como es el caso en las palabras *azteca*, *hasta* o *está*. Este alófono, por su punto de articulación, recibe el nombre de dental fricativo sordo, [s_[].

En el espectrograma puede verse como una mancha turbulenta, la cual suele ocupar todo el espectrograma; pero se puede diferenciar de la imagen del alófono alveolar, porque es mucho más angosta. Cabe señalar que la imagen de [s_[] siempre estará siempre precedida del cierre de [t]. En la Figura 51 se ha señalado la imagen espectrográfica del alófono dental, [s_[]. Posterior a ella, también puede observarse el segmento espectrográfico de [t].

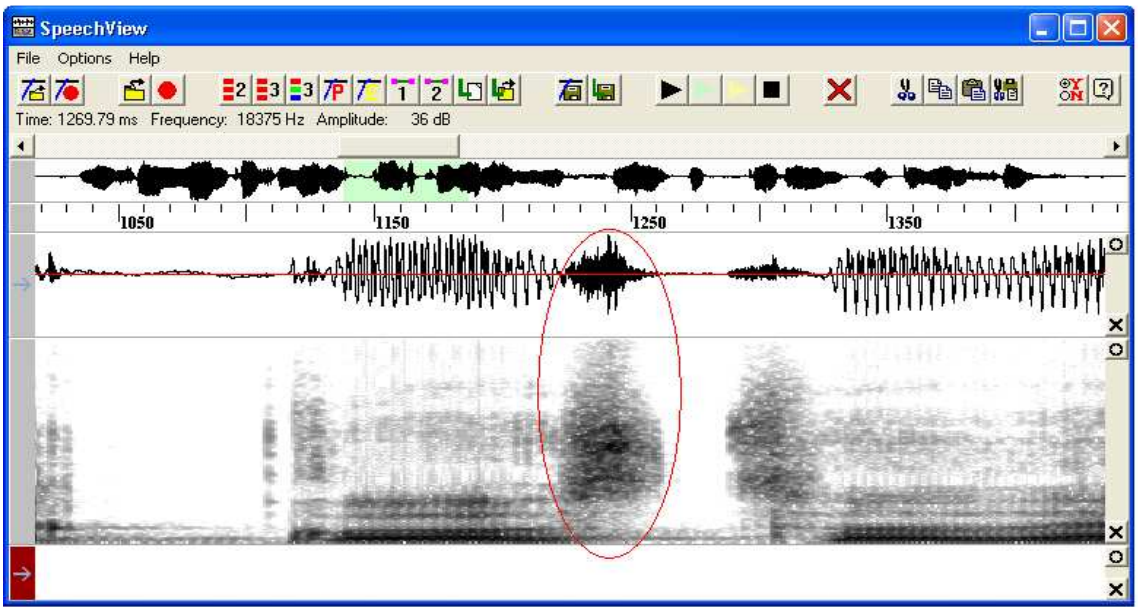

Figura 51. Imagen espectrográfica de [s_[ ].

La segmentación de este alófono se hace del inicio de la mancha turbulenta hasta el final de la misma, o bien hasta el inicio del cierre de [t]. La transcripción para este alófono en el nivel T54 es [s_[]; mientras que para los otros dos niveles (T44 y T22) es [s]. En la Figura 52 se puede observar la imagen espectrográfica de la palabra *costo*, en la cual se ha resaltado la transcripción del alófono dental.

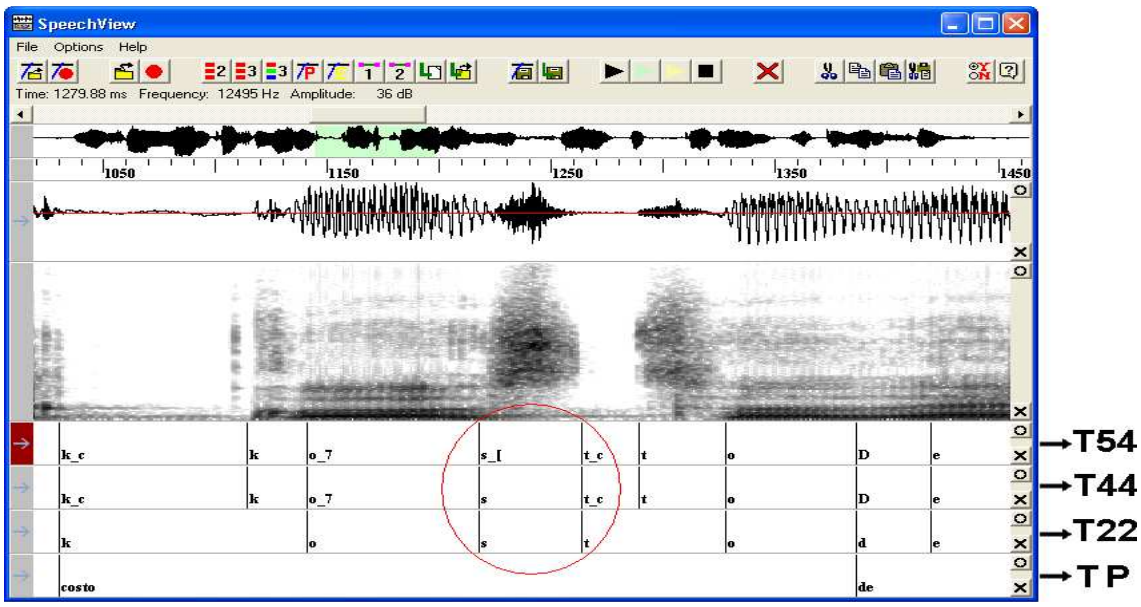

Figura 52. Imagen espectrográfica, segmentación y transcripción de [s_[ ].

3.2.3. Alófono alveolar fricativo sonoro [z]

El fonema alveolar fricativo sordo se sonoriza cuando está entre vocales o en contexto anterior a [b, d, g, Z, m, n, n~, l, r, r(]. Esta realización da lugar al alófono alveolar fricativo sonoro [z]. Éste se distingue en el oscilograma por tener una frecuencia fundamental baja, y en el espectrograma por ser una mancha que se asemeja a un círculo, formado por líneas verticales oscuras. En la Figura 53 se pude observar la imagen espectrográfica de [z], señalada en un óvalo.

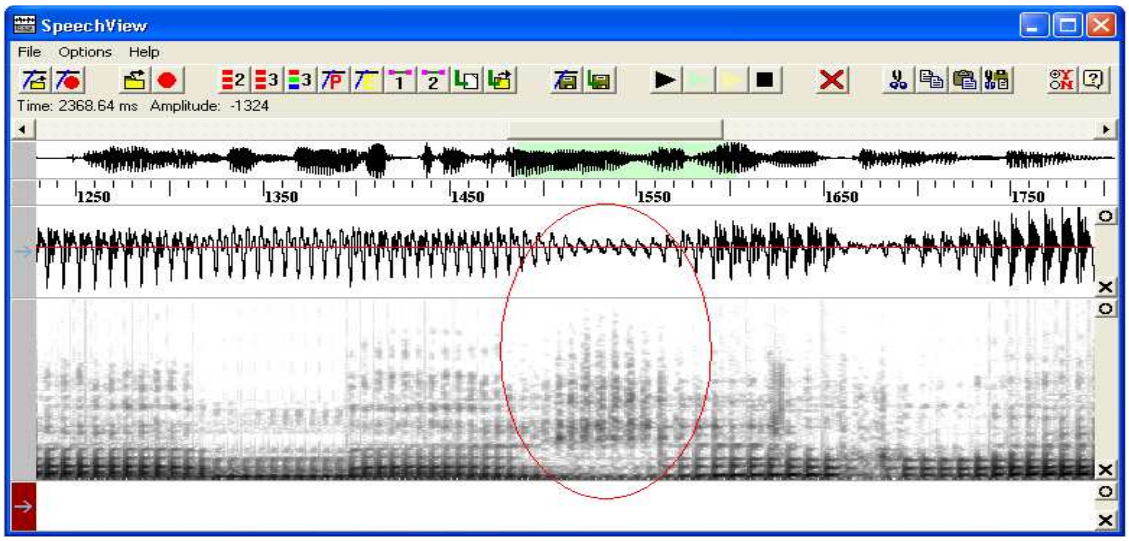

Figura 53. Imagen espectrográfica de [z].

La imagen espectrográfica de este alófono se segmenta a partir de donde comienza la mancha de fricación hasta el final de la misma; o bien, de donde la frecuencia fundamental baja hasta donde vuelve a ascender. En la Figura 54 se puede observar la segmentación de la imagen espectrográfica de [z].

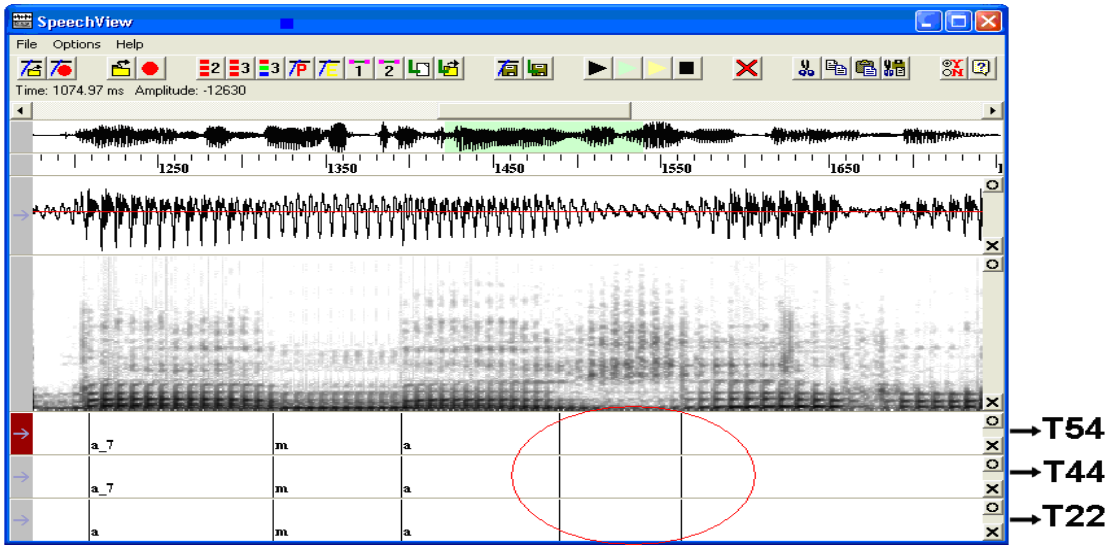

Figura 54. Imagen espectrográfica y segmentación [z].

La transcripción de este alófono para el nivel T54 es [z]; mientras que para los niveles T44 y T22 es [s]. En la Figura 55 se observa la transcripción de [z] en los tres niveles de transcripción.

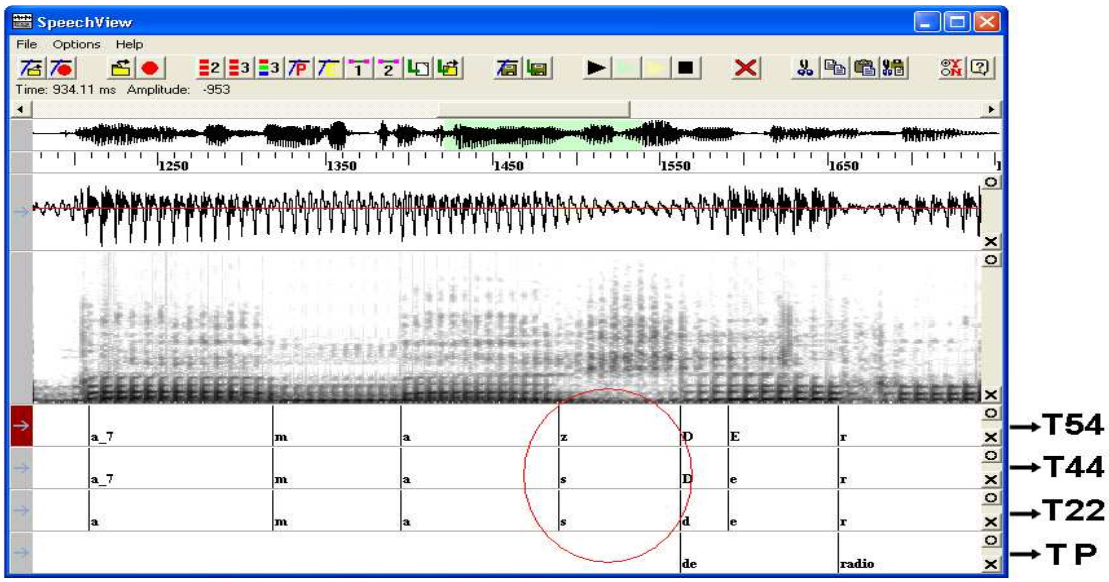

Figura 55. Imagen espectrográfica y transcripción de [z].

En ocasiones, el alófono alveolar fricativo sonoro no sonoriza y se realiza como el alófono fricativo sordo. Cabe señalar que este caso es una excepción y que, cuando sucede, se transcribe en las etiquetas de los tres niveles el símbolo [s]; es decir, se transcribe cómo ha sido su realización. En la Figura 56 se muestra un ejemplo con la oración *no se puede*. A pesar de que por contexto el fonema /s/ debió sonorizarse, ya que está entre vocales, se realizó como sordo, por lo tanto se transcribió como fue emitido.

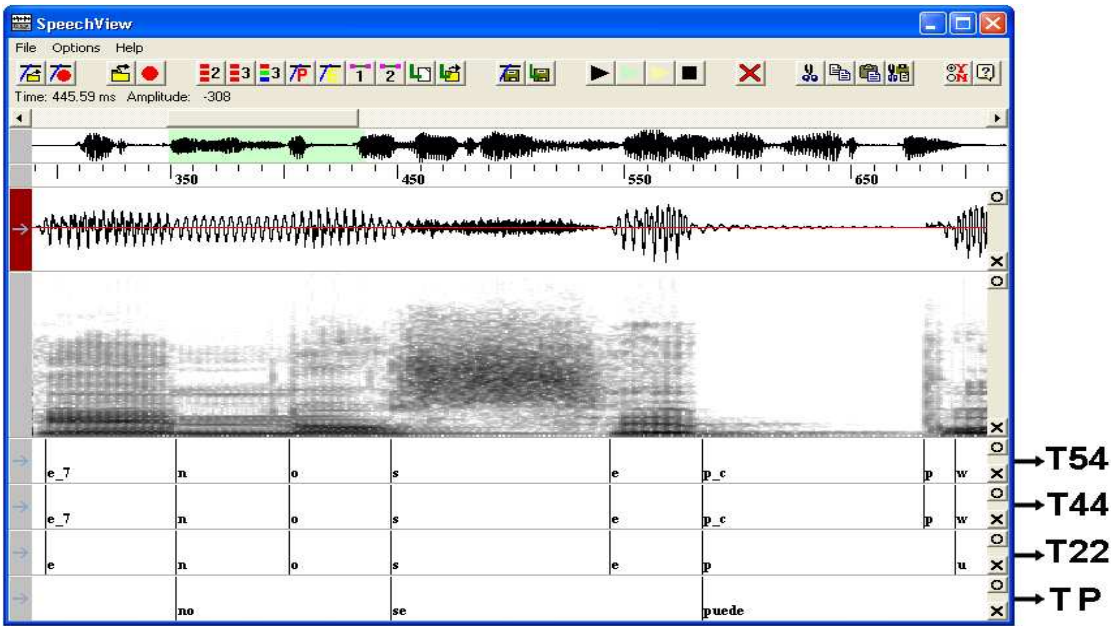

Figura 56. Imagen espectrográfica y transcripción de [z] no sonorizada.

3.3. Fonema velar fricativo sordo /x/

El fonema velar fricativo sordo tiene un sólo alófono en distribución complementaria, el velar fricativo sordo [x], que se realiza al inicio y final de sílaba. Este alófono se articula en el velo del paladar y las cuerdas vocales no vibran durante su emisión (Quilis 1999). También forma parte de las consonantes de *altas frecuencias*, pero su imagen acústica posee menor energía que la imagen de [s] (Quilis 1981). Cuétara (2004) lo documentó como un alófono de baja frecuencia de aparición, en el *Corpus DIME*. En la Figura 57 se muestra la regla de distribución complementaria, junto con su grafía en *Mexbet*, conforme a Cuétara 2004, y su equivalencia ortográfica.

Fonema	Alófono	Contexto	Grafía
Velar fricativo sordo /x/	Velar fricativo sordo [x]	Inicio y final de sílaba	*j* *g* + *a*, *e*, *i*

Figura 57. Reglas distribucionales de los alófonos del fonema /x/.

3.3.1. Alófono velar fricativo sordo [x]

El alófono velar fricativo sordo, [x], como casi todas las fricativas, se distingue en el espectrograma por ser una mancha turbulenta que lo ocupa casi completamente. En la Figura 58 se puede observar la imagen espectrográfica de este alófono, delimitada en un círculo.

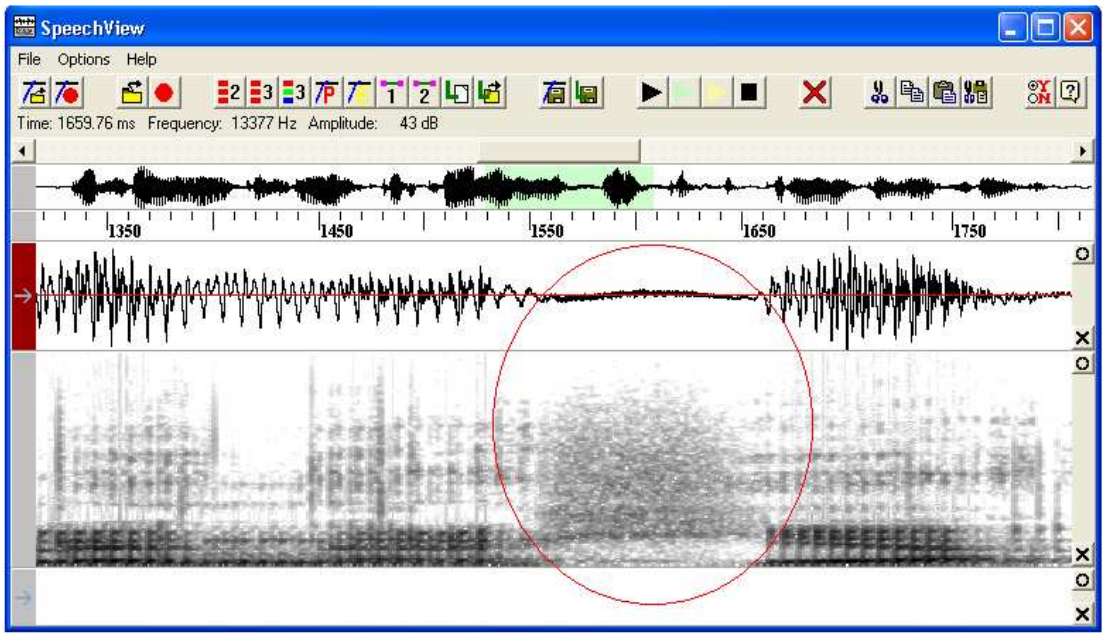

Figura 58. Imagen espectrográfica de [x]

Para el proceso de segmentación de este alófono remitirse a [f]. La trascripción es [x] para las etiquetas de los tres niveles. En la Figura 59 se puede observar la imagen espectrográfica de la palabra *energía*, en la cual se resalta la segmentación y la transcripción del alófono velar fricativo sordo.

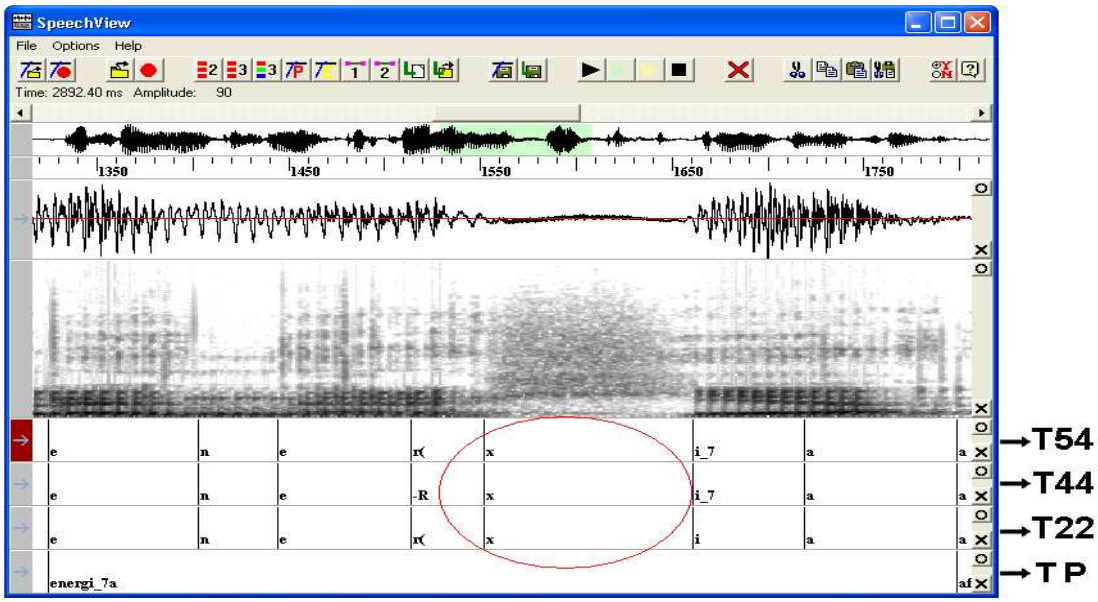

Figura 59. Imagen espectrográfica, segmentación y transcripción de [x]

3.4. Fonema palatal fricativo sonoro /Z/

El fonema palatal fricativo sonoro /Z/ tiene dos alófonos: uno prototípico palatal fricativo sonoro [Z], que ocurre al interior de sílaba; y el palatal africado sonoro [dZ], que ocurre en posición de inicio absoluto y posterior a las nasales [m, n].

La articulación del alófono palatal fricativo sonoro "se realiza con el predorso de la lengua contra la región palatal" (Quilis 1999:252). Espectrográficamente, este alófono, a diferencia de las fricativas sordas, muestra más armónicos que ruido y su energía se concentra en la parte inferior del espectrograma; razón por la cual pertenece al grupo de las fricativas de *resonancias bajas* (Hidalgo y Quilis 2004).

El alófono africado se articula del mismo modo que [tS], por lo que de igual manera, consta de dos momentos: uno interrupto y otro constrictivo. Esto dos momentos se reflejan en el espectrograma por medio de dos segmentos: el primero como un cierre y el segundo como una fricación.

En la Figura 60 se presentan las reglas distribucionales de los alófonos de /Z/, conforme a Cuétara 2004. Posteriormente, se lleva a cabo el proceso de segmentación y transcripción para cada uno de los alófonos.

Fonema	Alófono	Contexto	Grafía
Palatal fricativo sonoro /Z/	Palatal fricativo sonoro [Z]	En contexto interior de sílaba	*ll, y*
	Palatal africado sonoro [dZ_c] [dZ]	En contexto inicial de sílaba y posterior a [m, n, l]	*ll, y*

Figura 60. Reglas distribucionales de los alófonos del fonema /Z/.

3.4.1. Alófono palatal fricativo sonoro [Z]

El fonema palatal fricativo sonoro, [Z], ocurre al interior de sílaba. Este alófono se identifica en el espectrograma por ser un segmento con formantes parecidos a los vocálicos pero de menor energía, razón por la cual es menos oscuro que los segmentos que lo preceden o anteceden. En la Figura 61 se resalta la imagen correspondiente a este alófono.

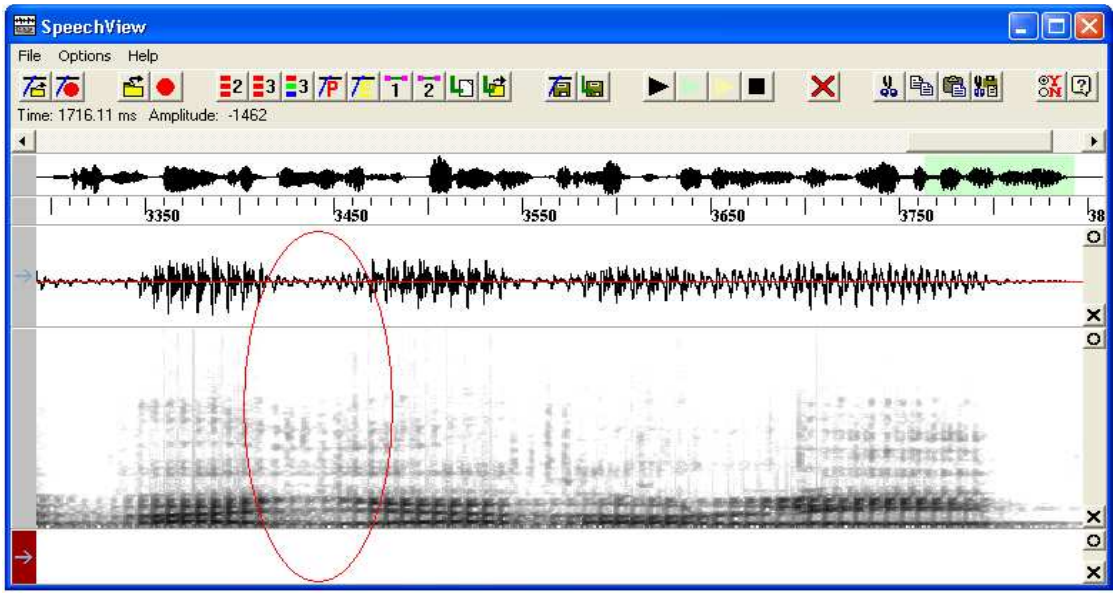

Figura 61. Imagen espectrográfica de [Z].

La segmentación del alófono fricativo sonoro se hace a partir de donde baja la frecuencia fundamental hasta donde vuelve a ascender, o bien, de donde los formantes se desvanecen hasta donde vuelven a tener energía. La transcripción para las etiquetas de los tres niveles es [Z]. En la Figura 62 se puede observar la imagen espectrográfica de la palabra *Valladolid*; en la cual se ha resaltado la segmentación y transcripción de [Z].

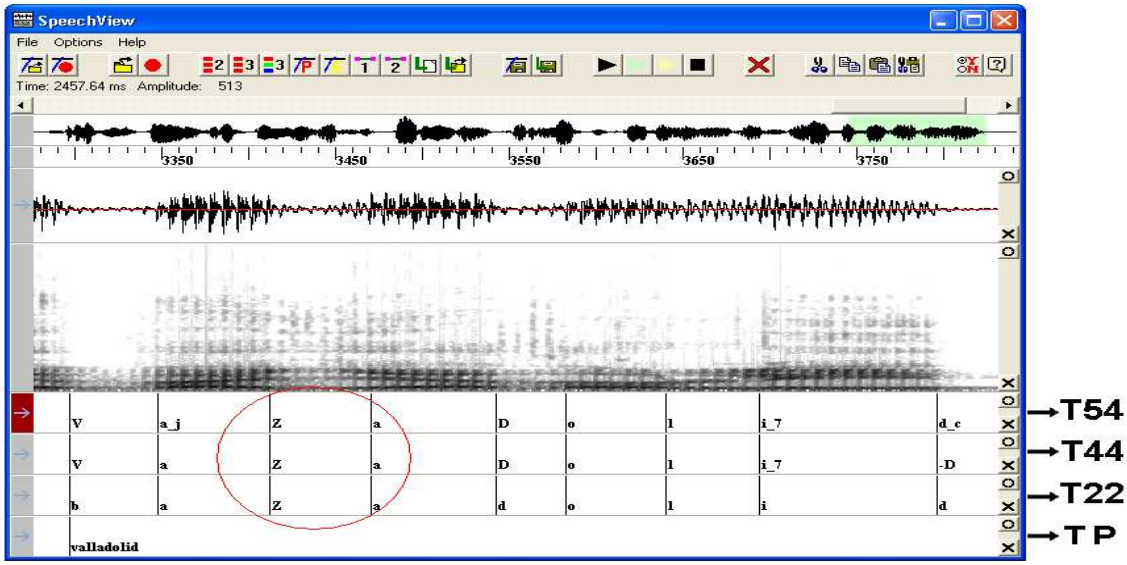

Figura 62. Imagen espectrográfica, segmentación y transcripción de [Z].

3.4.2. Alófono palatal africado sonoro [dZ]

La imagen espectrográfica del alófono palatal africado sonoro, [dZ], se distingue en el espectrograma por tener un cierre y una barra de explosión como los oclusivos, pero se diferencia de ellos porque inmediatamente a la explosión aparece una mancha turbulenta, que pertenece al momento de fricación. En la Figura 63 se puede apreciar la imagen espectrográfica del alófono africado en contexto inicial.

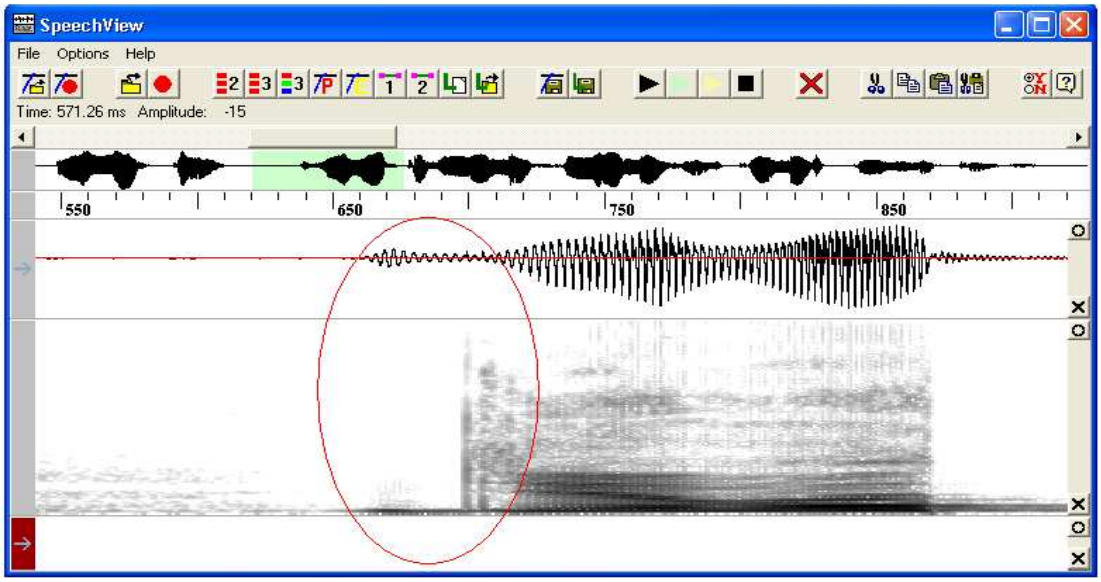

Figura 63. Imagen del espectrográfica de [dZ] en contexto inicial.

En el nivel T54, la delimitación de este alófono se hace en dos segmentos. El primero corresponde al cierre, esta segmentación se hace al comienzo de [dZ] hasta el inicio de la barra de explosión. El segundo se hace a partir del comienzo de la barra hasta el final de la fricación.

En los niveles T44 y T22 [dZ] se delimita en un sólo segmento, que va desde el comienzo del cierre hasta el final de la turbulencia, o bien, de la fricación. En la Figura 64 se puede observar la segmentación del alófono palatal africado en los tres niveles de transcripción.

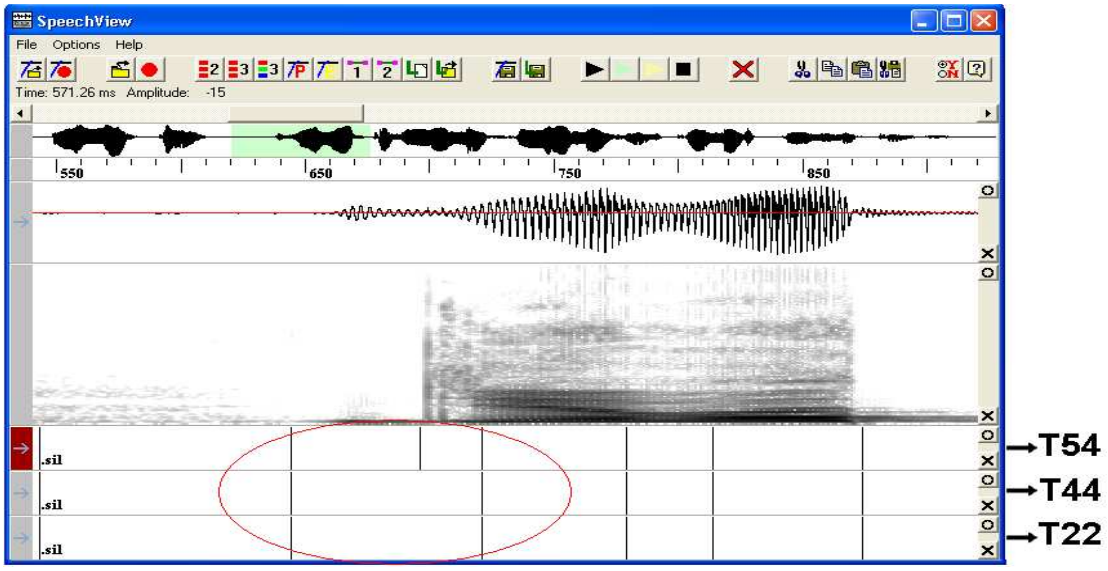

Figura 64. Imagen espectrográfica y segmentación de [dZ] en contexto inicial.

La transcripción para las etiquetas del nivel T54 son [dZ_c] para el cierre, y [dZ] para el segmento fricativo. La transcripción para las etiquetas de los niveles T44 y T22 es [Z]. En la Figura 65 se pude observar la imagen espectrográfica de la palabra *lleva*; en ella se ha resaltado la transcripción de [dZ] en contexto inicial absoluto, en los tres niveles.

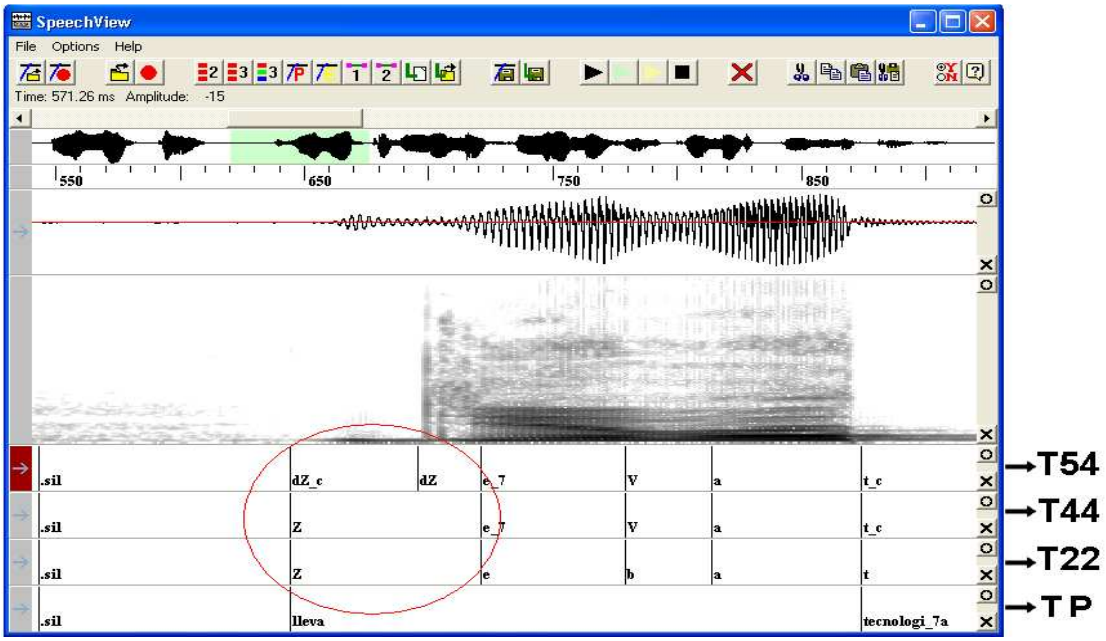

Figura 65. Imagen espectrográfica y transcripción de [dZ] en contexto inicial.

Cuando el alófono palatal africado sonoro [dZ] ocurre en contexto posterior a las nasales [m, n], en la imagen acústica no aparece el cierre, tal como ocurre con las oclusivas sonoras (§1.5). En la Figura 66 se ha señalado la imagen acústica de [dZ] en contexto posterior a nasal. En dicha imagen se puede observar que no hay cierre, sino que posterior al segmento nasal aparece la barra de explosión seguida de la fricación.

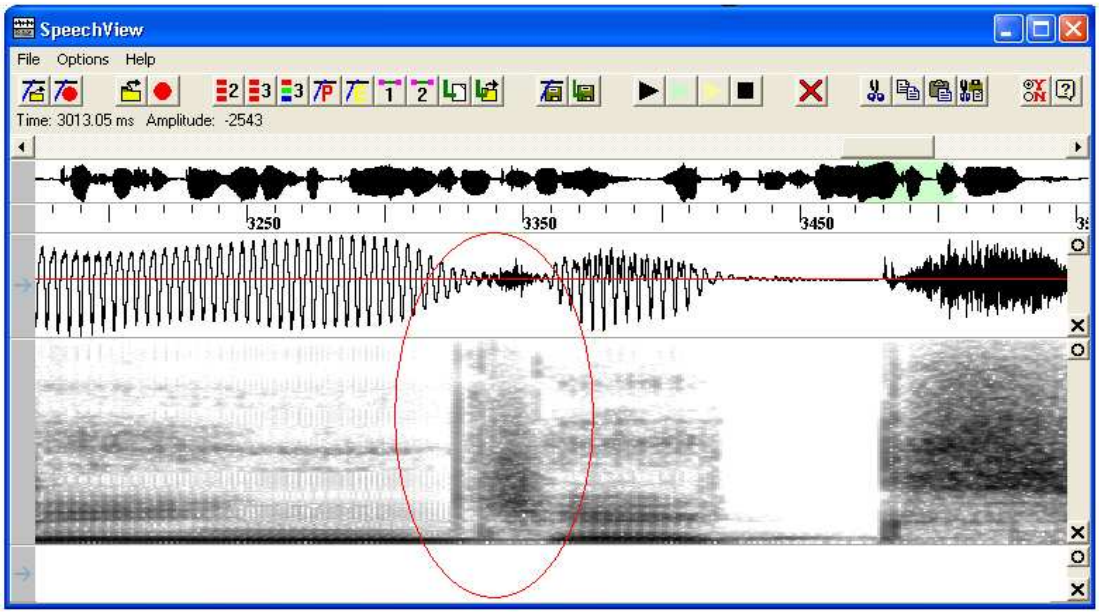

Figura 66. Imagen espectrográfica de [dZ] en posición posterior a nasal.

La segmentación para esta imagen se hace a partir de la barra explosión hasta el final de la turbulencia. La transcripción para las etiquetas es [dZ] en el nivel T54 y [Z] para los niveles T44 y T22. En la Figura 67 se puede ver la segmentación y transcripción de este alófono.

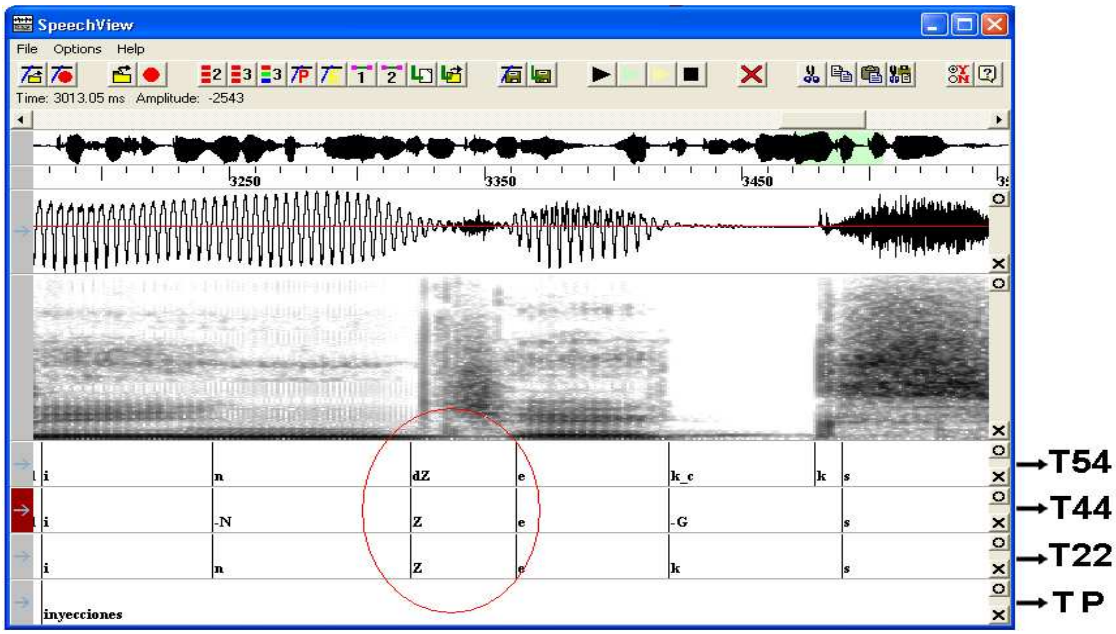

Figura 67. Imagen espectrográfica, segmentación y transcripción de [dZ] en posición posterior a nasal.

4. Fonemas nasales /m/, /m/, /n~/

4.1. Fonema labial nasal /m/

El fonema labial nasal /m/ tiene un solo alófono labial nasal [m], que ocurre en posición inicial de sílaba. En la articulación de este alófono ambos labios se juntan provocando que el aire salga a través de las fosas nasales. “El velo del paladar se separa de la pared faríngea; las cuerdas vocales son vibran” (Quilis 1999:226).

Lander (1997:50) menciona que la imagen espectrográfica de las nasales es fácil de distinguir porque “the waveform rises or drops into a highly periodic, low amplitude signal. The nasal usually carries the same formants of the preceding vowel or other phone, but is lighter in color or intensity”. Herrera (2002) señala que en ellas se puede identificar un formante de muy baja frecuencia en toda la resonancia nasal; por tal motivo, la transición de los formantes vocálicos adyacentes es negativa. Quilis (1981:209) menciona que las nasales “comparten con las explosivas orales la forma y la dirección de las transiciones del segundo y tercer formante de las vocales contiguas”.

En cuanto a su realización, Matluck (1951:105) observa que [m] “tiende a relajarse ligeramente en posición intervocálica: *amo*, *cama*, etc. El relajamiento nunca llega, como en Nuevo México, a hacerla desaparecer”. Perissinotto (1975) documenta una realización fricativa que ocurre cuando está en contexto anterior a /f/, como en la palabra *infantil*. Sin embargo, en *Mexbet* únicamente se tiene el alófono prototípico, [m], que se realiza en posición inicial de sílaba, ya que Cuétara (2004) lo reportó como un alófono estable, pero de baja frecuencia, por lo que se consideró como único alófono de /m/.

Cuando el alófono nasal aparece en posición de coda silábica puede alternar su realización con cualquier forma alofónica de /n/. Por ejemplo en *empezar* [empezar(] ~ [eNpezar(]. Quilis (1999:228) menciona dos motivos por lo cuales estos alófonos se neutralizan: “en

primer lugar el fonológico: como no existen diferencias significativas entre las consonantes nasales en posición postnuclear, su lugar articulatorio no es pertinente; en segundo lugar, el puramente fonético, pero, a su vez, en íntima conexión con el anterior: al no ser significativos los distintos lugares de articulación, lo importante, en el plano del habla es que se realice una oclusión bucal y que quede una resonancia nasal". Esta neutralización de los alófono nasales en posición de coda silábica recibe en el nivel T44 la transcripción de [-N], en el alfabeto *Mexbet*. En la Figura 68 se muestra el cuadro con la regla distribucional del fonema bilabial nasal; posteriormente se muestra el proceso de segmentación y transcripción de su alófono.

Fonema	Alófono	Contexto	Grafía
Bilabial nasal /m/	Bilabial nasal [m]	En todos los contextos	*m*

Figura 68. Reglas distribucionales de los alófonos del fonema /m/.

4.1.1. Alófono labial nasal [m]

En el espectrograma, el alófono nasal [m] se nota como un segmento de poca energía, ya que sus formantes son poco visibles. Esta baja energía hace que el segmento del alófono nasal contraste con el oscurecimiento de la imagen de los alófonos vecinos. En la Figura 69 se puede ver la imagen espectrográfica de [m].

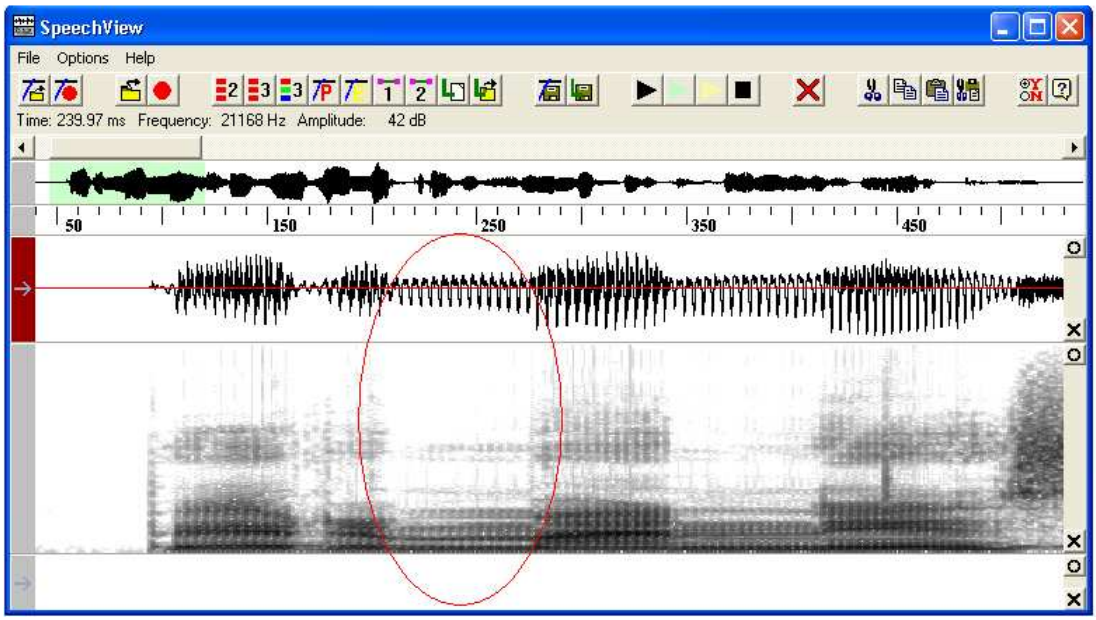

Figura 69. Imagen espectrográfica de [m].

El contraste que se ve entre el segmento nasal y los segmentos que lo anteceden y preceden, ayuda a delimitar con más exactitud al alófono. Lander (1997:50) menciona que "Set the right boundary where the first formant in the spectrogram dies. this should coincide with a point of "radical" change in the waveform". Así, la delimitación de este alófono se hace de donde baja la energía del segmento o la amplitud de la onda sonora, hasta donde vuelven a ascender. En la Figura 70 se pude ver cómo ha sido hecha la segmentación de [m], en los tres niveles de transcripción.

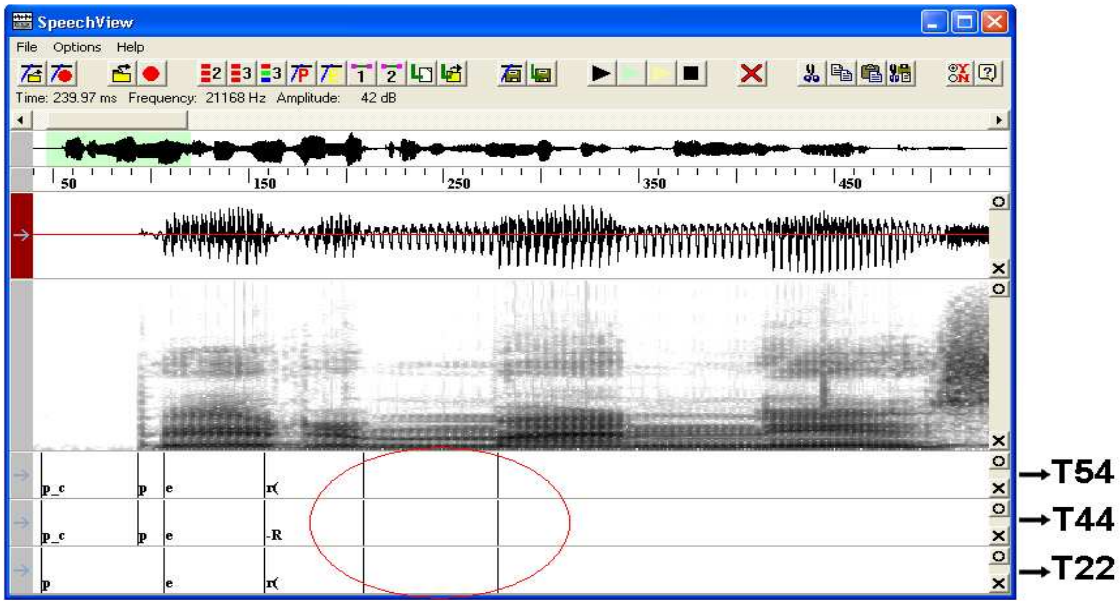

Figura 70. Imagen espectrográfica y segmentación de [m].

La transcripción de este alófono es [m] para las etiquetas de los tres niveles. En la Figura 71 se muestra la imagen espectrográfica de la palabra *permanece*; en ella se ha resaltado la transcripción de [m].

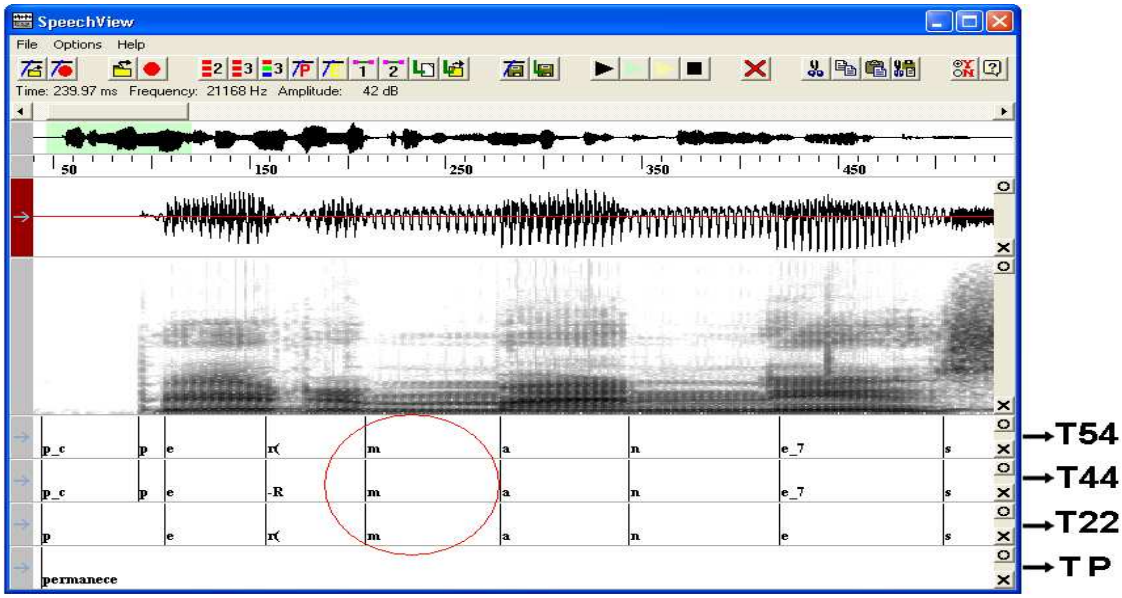

Figura 71. Imagen espectrográfica y transcripción de [m].

Cuando el alófono bilabial nasal se encuentra en posición de coda silábica, se transcribe con la grafía [-N] en el nivel T44; mientras que en los niveles T54 y T22 se hace la transcripción correspondiente al alófono que fue emitido. En la Figura 72 se puede observar la imagen espectrográfica de la palabra *siempre*. En esta imagen se resalta la transcripción de [m] en posición de coda silábica.

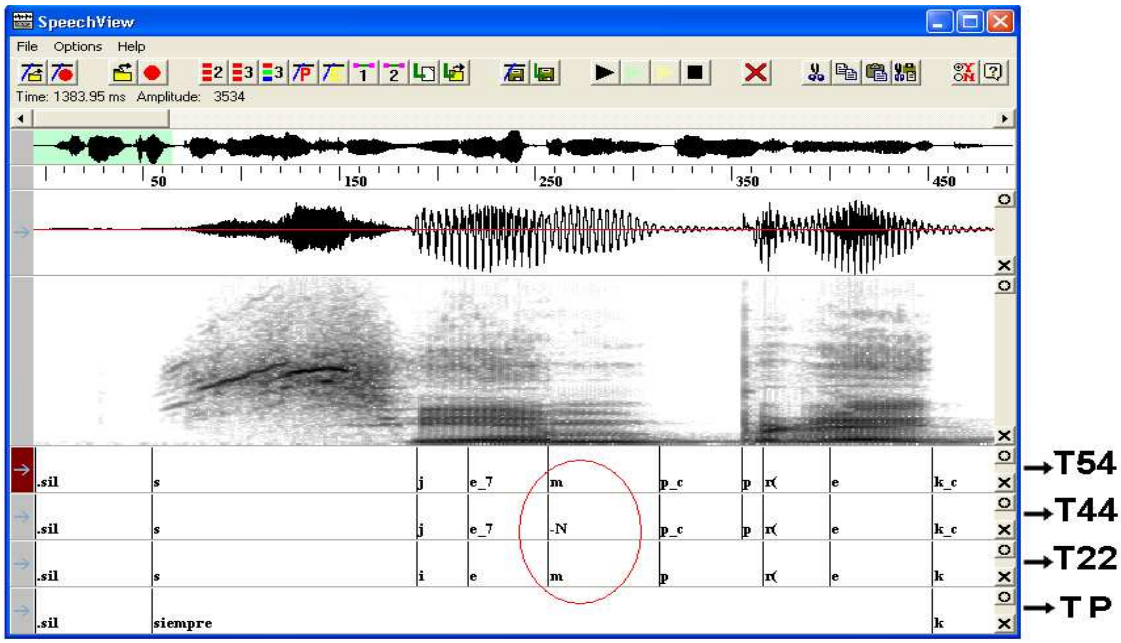

Figura 72. Imagen espectrográfica y transcripción de [-N].

4.2. Fonema alveolar nasal /n/

El fonema alveolar nasal es uno de los fonemas con más frecuencia de aparición en el español. Este fonema presenta una estructura central lo suficientemente moldeable como para asimilarse con la consonante que le precede (Herrera 2002:3); por tal motivo, tiene varias formas alofónicas: una alveolar, otra dental y una más velar.

Cada alófono del fonema alveolar nasal /n/ se realiza en distinto contexto: el alveolar nasal [n] prototípico, que se realiza en cualquier posición de sílaba; el dental nasal [n_[], que ocurre en contexto anterior a [t, d]; y el alófono velar nasal [N], que se realiza en contexto anterior a [k, g].

La imagen espectrográfica de estos alófonos “se caracteriza por la presencia de bandas de resonancia en las cuales hay una concentración de energía que conforman los llamados formantes nasales” (Herrera 2002:3). Aunque normalmente tienen un solo formante bien marcado, en ocasiones se llega a percibir un segundo y tercer formante de energía débil. Con respecto a las transiciones de dichos formantes suelen “presentan la misma forma y dirección de las transiciones del segundo y tercer formantes de las vocales contiguas” (Hidalgo y Quilis 2004:159).

Como se dijo para /m/ (§7), cuando los alófonos de /n/ están en posición de coda alternan su pronunciación con [m], por ejemplo *envase* [enbaze] ~ [embaze]. Esta alternancia sucede porque continuamente /n/ se asimila al punto de articulación de alófono que le precede (Irribarren 2005:290). Matluck (1951) reporta que en el Valle de México la pronunciación de [n] en posición final de sílaba tiende a conservarse aunque no deja de presentar irregularidades, con mayor medida entre las personas de clase inculta.

En la Figura 73 se presentan las reglas distribucionales de los alófonos de /n/, conforme a Cuétara 2004. Después, se muestra el proceso de segmentación y transcripción de cada uno de los alófonos del fonema nasal alveolar.

Fonema	Alófono	Contexto	Grafía
Alveolar nasal /n/	Alveolar nasal [n]	En cualquier posición de sílaba	*n*
	Dental nasal [n_[]	En contexto anterior a [t, d]	*n*
	Velar nasal [N]	En contexto anterior a [k, g]	*n*

Figura 73. Reglas distribucionales de los alófonos del fonema /n/.

4.2.1. Alófono alveolar nasal [n]

La imagen acústica del alófono alveolar nasal [n] se distingue en el espectrograma por ser un segmento de baja frecuencia, en el cual regularmente son casi imperceptibles el segundo y tercer formantes. En la Figura 74 aparece señalada la imagen acústica de [n].

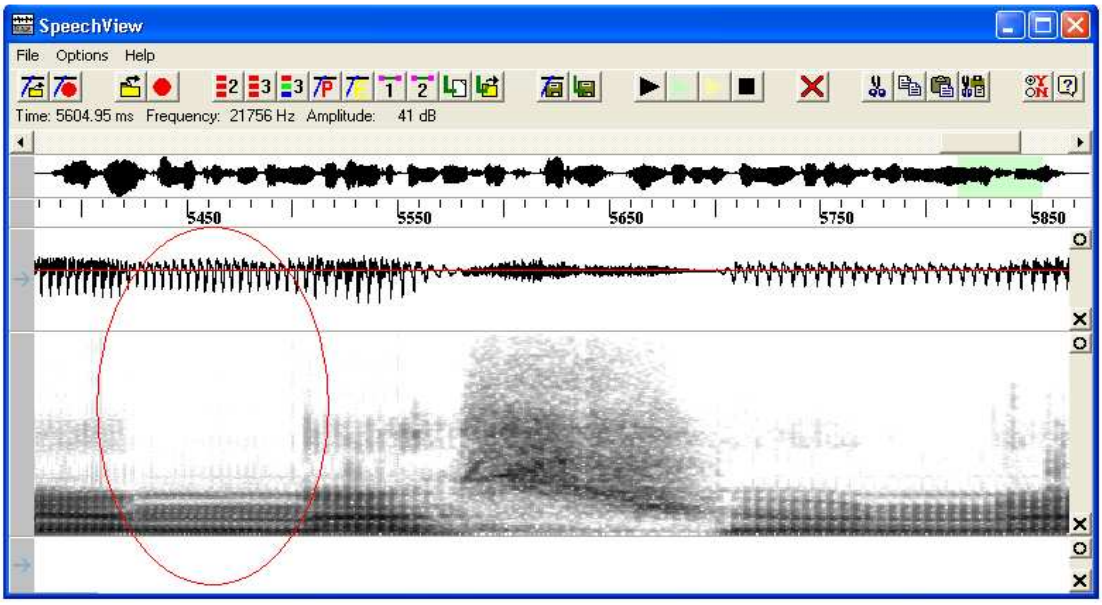

Figura 74. Imagen espectrográfica de [n].

El proceso de delimitación de [n] es el mismo que en el alófono bilabial nasal (remitirse a [m]). La transcripción para este alófono es [n] para las etiquetas de los tres niveles. En la Figura 75 se muestra la imagen espectrográfica de la palabra *nacional*, en la cual se ha señalado la segmentación y transcripción de [n].

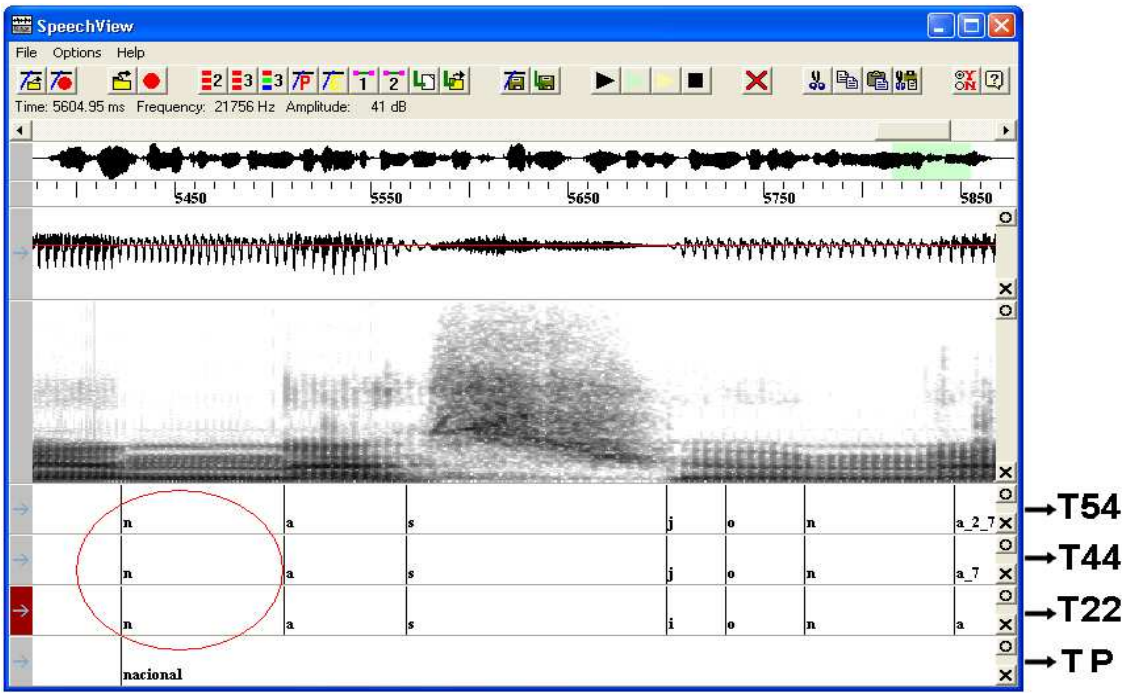

Figura 75. Imagen espectrográfica, segmentación y transcripción de [n].

4.2.2. Alófono dental nasal [n_[]

El fonema dental nasal [n_[] en el espectrograma se distingue como un segmento de baja energía, el cual siempre aparece precedido del cierre o de la barra de explosión correspondiente a los alófonos [t] o [d]. En la Figura 76 se puede observar el segmento espectrográfico correspondiente al alófono dental nasal.

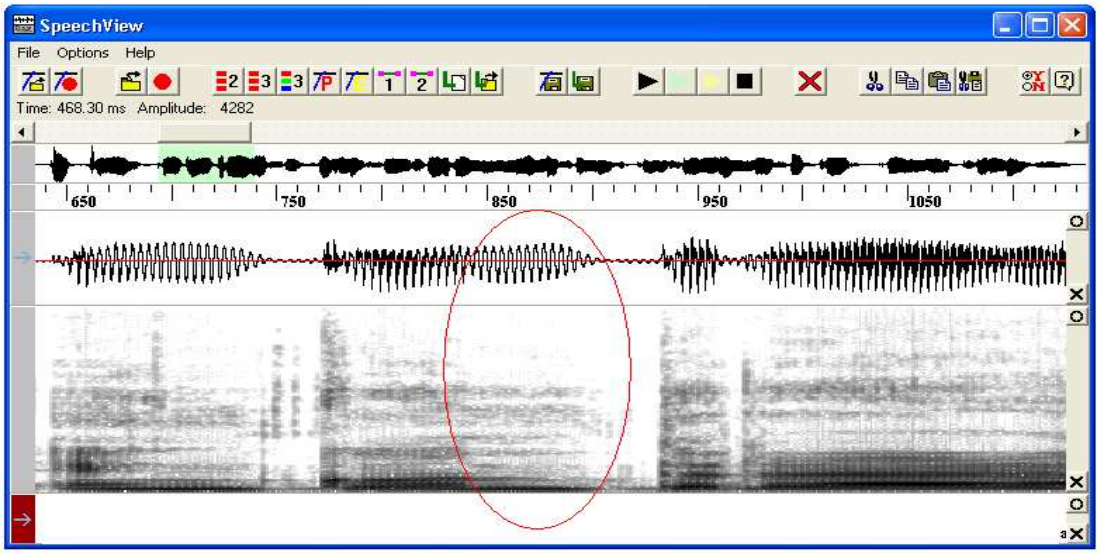

Figura 76. Imagen espectrográfica de [n_[ ].

La segmentación de este alófono para los tres niveles de transcripción, se hace a partir de donde baja la frecuencia fundamental hasta el inicio del cierre o de la barra de explosión de la consonante oclusiva que la precede. Es importante recordar que cuando algún alófono oclusivo se encuentra en posición posterior a una nasal, ésta absorbe su cierre. Machuca *et al*. (1999) mencionan que cuando la nasal está ante una oclusiva sorda, la segunda delimitación se hace al final de la sonoridad de la nasal. En la Figura 77 se pude observar la segmentación de [n_[], en los tres niveles de transcripción (T22, T44, T54).

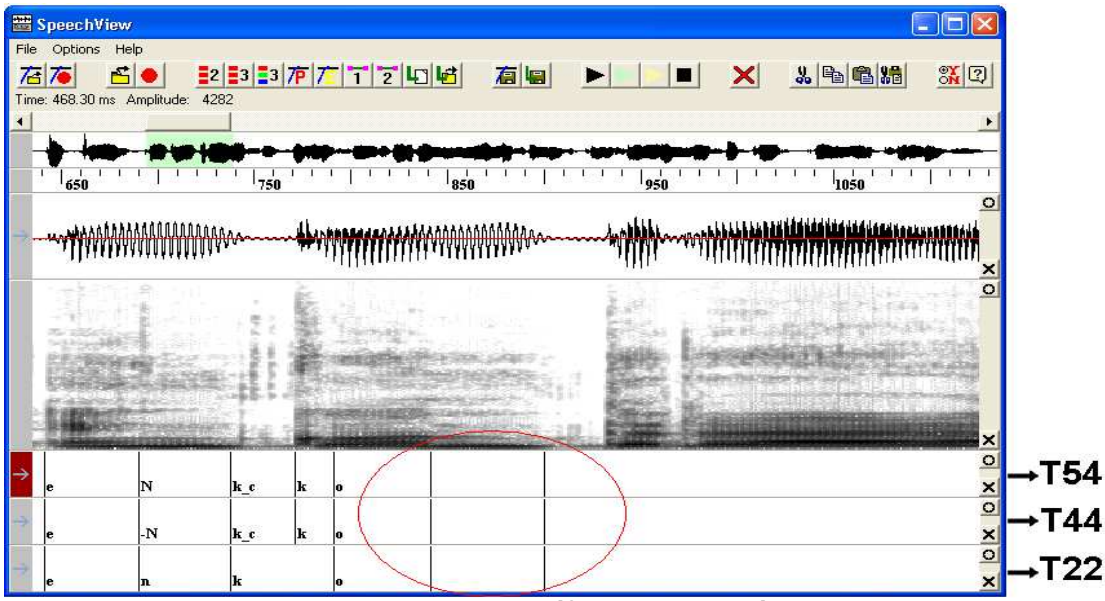

Figura 77. Imagen espectrográfica y segmentación de [n_[].

La transcripción para la etiqueta del nivel T54 es [n_[]. En este contexto, el alófono [n_[] siempre aparecerá en posición de coda silábica, por lo tanto la transcripción para el nivel T44 será [-N]; y para el nivel T22 será [n]. En la Figura 78 podemos observar la imagen espectrográfica de la palabra *encontrar*, en ella se ha señalado la transcripción del alófono dental nasal.

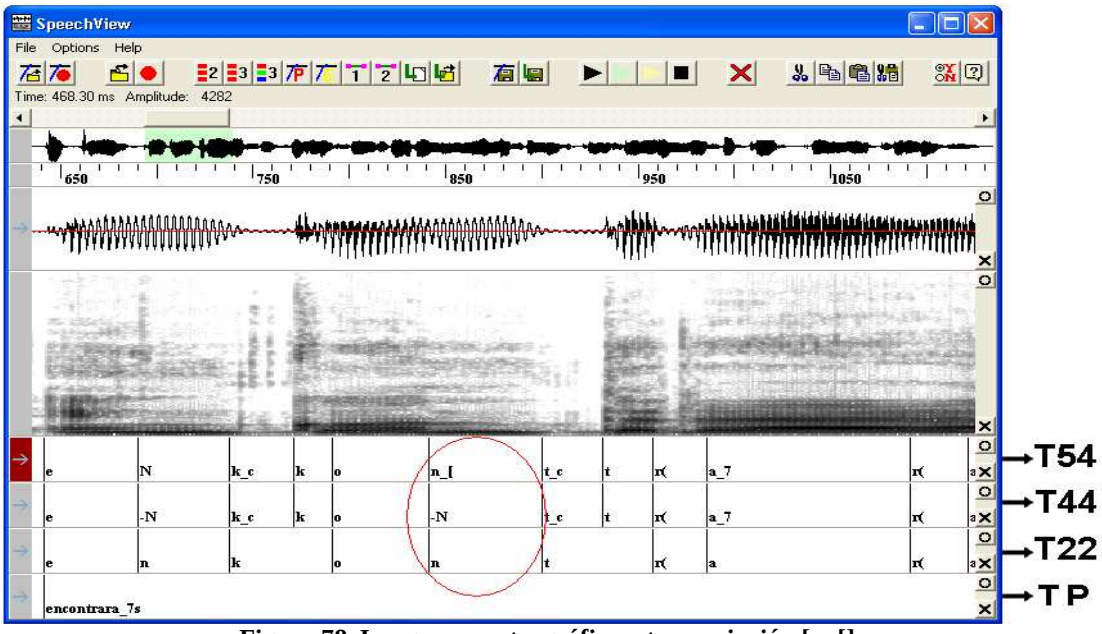

Figura 78. Imagen espectrográfica y transcripción [n_[].

4.2.3. Alófono velar nasal [N]

El alófono velar nasal [N] ocurre en posición anterior a los alófonos velares [k, g], como en las palabras *mango*, *brinco*, *tengo*, etc. La imagen acústica de este alófono se caracteriza por ser un segmento que consta de formantes de baja energía, y al igual que [n_[], siempre está precedido del cierre o de la barra de explosión de [k] o de [g]. En la Figura 79 se puede observar la imagen espectrográfica del alófono dental nasal. En esta imagen también se puede ver que, posterior al segmento nasal, aparece el cierre de la consonante oclusiva.

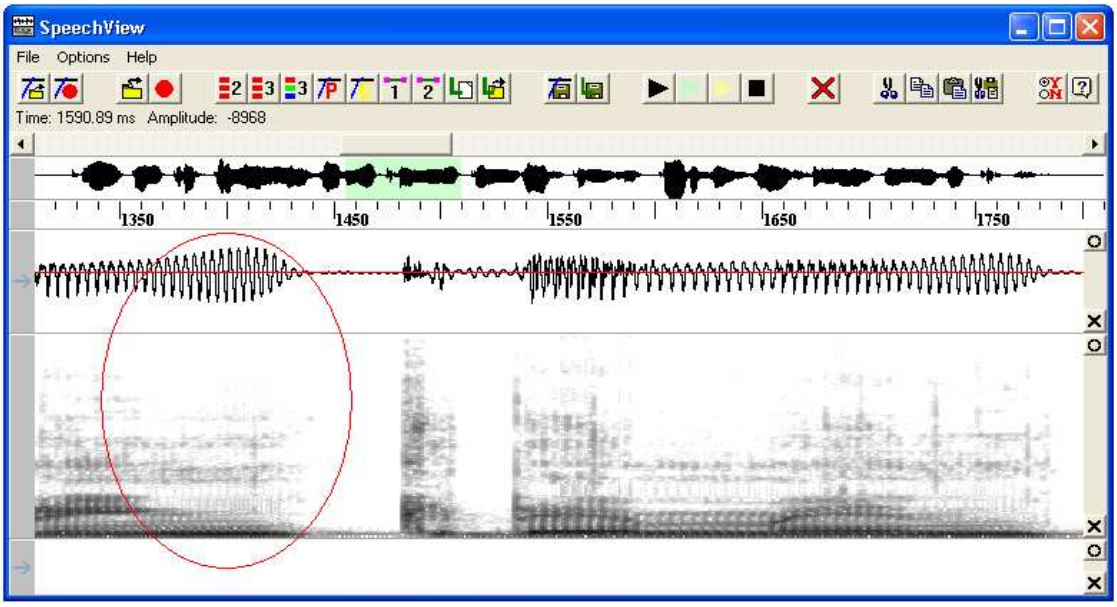

Figura 79. Imagen espectrográfica de [N].

La segmentación del alófono nasal velar se hace de la misma manera que en [n_[] (ver [n_[]). La transcripción de este alófono en el nivel T54 es [N]; para el nivel T44 es [-N], ya que siempre se encontrará en posición de coda silábica; y para el nivel T22 es [n]. En la Figura 80 aparece la imagen acústica de la palabra *incremento*, en ella se ha resaltado la transcripción de [n].

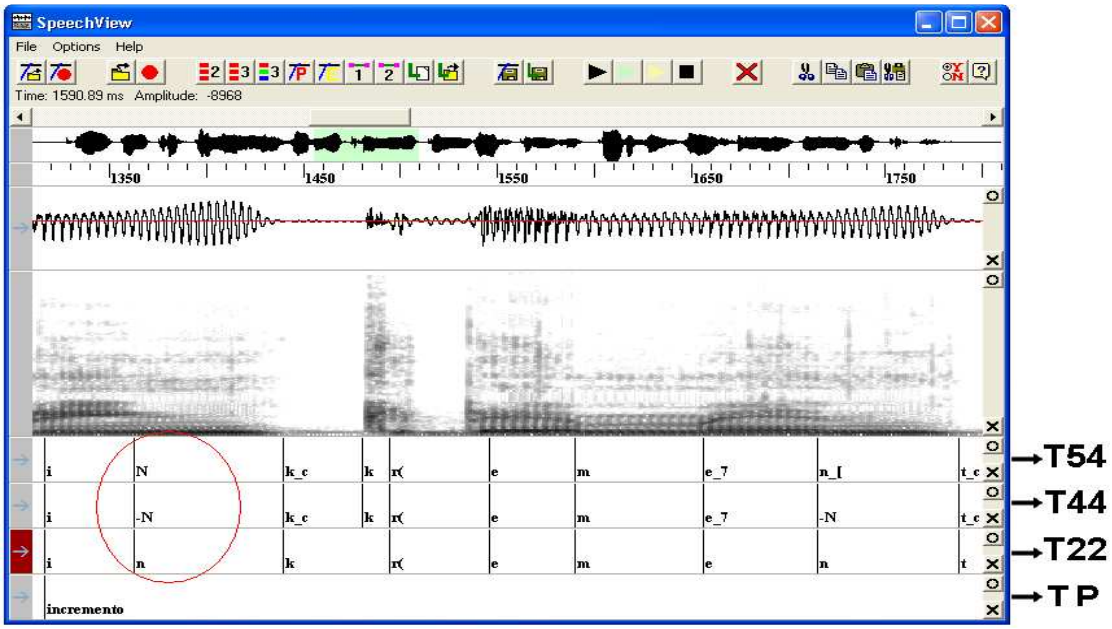

Figura 80. Imagen espectrográfica, segmentación y transcripción de [n].

Cuando algún alófono de /n/ aparecen en posición de coda silábica, se transcribirá en el nivel T44 con la grafía [-N], al igual que en [m] (§7.1). En los otros dos niveles (T54 y T22) se transcribirá el alófono que fue emitido. En la Figura 81 se muestra el ejemplo de la transcripción de [-N] en la palabra *venta*.

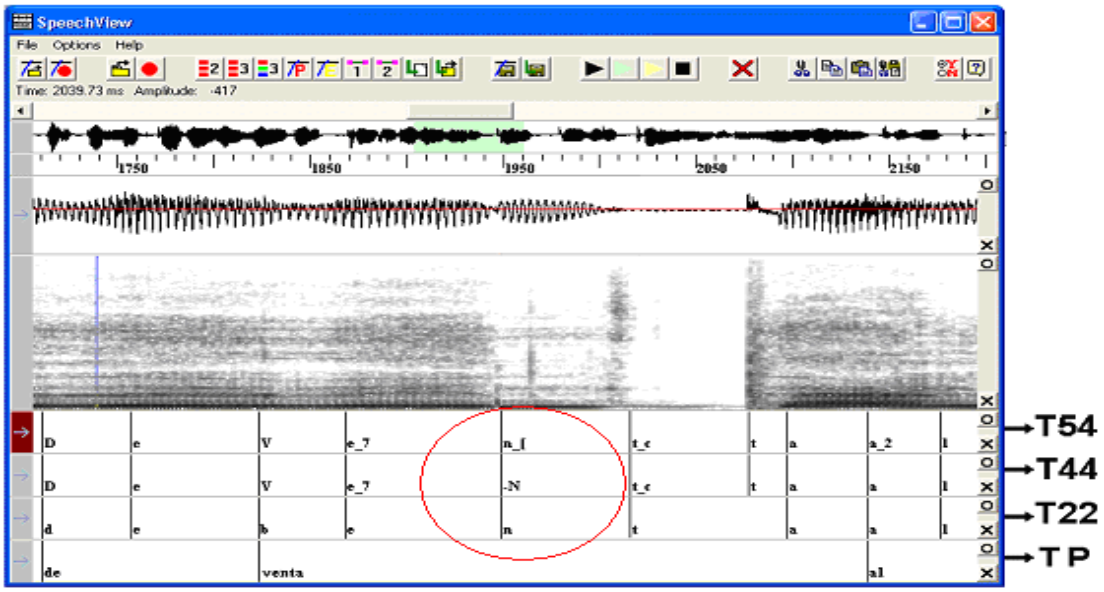

Figura 81. Imagen espectrográfica y transcripción de [-N].

4.3. Fonema palatal nasal /n~/

El fonema palatal nasal, /n~/ tiene un solo alófono en distribución complementaria: el alófono palatal nasal [n~], que ocurre en inicio e interior de sílaba; por ejemplo en *año*, *niño*, *ñandú*, etc. Este alófono se articula en "la región predorsal de la lengua se adhiere a la zona prepalatal, cerrando, de este modo, la salida del aire. El velo del paladar está separado de la pared faríngea; las cuerdas vocales vibran" (Quilis 1999:227).

Cuétara (2004) reportó a este alófono con una baja frecuencia de aparición en el *Corpus DIME*; por tal razón se asignó a [n~] como único alófono para este fonema. En la Figura 82 se muestra un cuadro con las reglas combinatorias del fonema palatal, de acuerdo con Cuétara 2004. Posteriormente, se mostrará el proceso de segmentación y transcripción del alófono de dicho fonema.

Fonema	Alófono	Contexto	Grafía
Palatal nasal /n~ /	Palatal nasal [n~]	En todos los contextos	*ñ*

Figura 82. Reglas distribucionales de los alófonos del fonema /n~/.

4.3.1. Alófono palatal nasal [n~]

En el espectrograma se puede identificar a este alófono por su baja frecuencia y porque las transiciones de sus formantes con respecto a los vocálicos son positivas. En la Figura 83 se señala el segmento espectrográfico correspondiente a [n~].

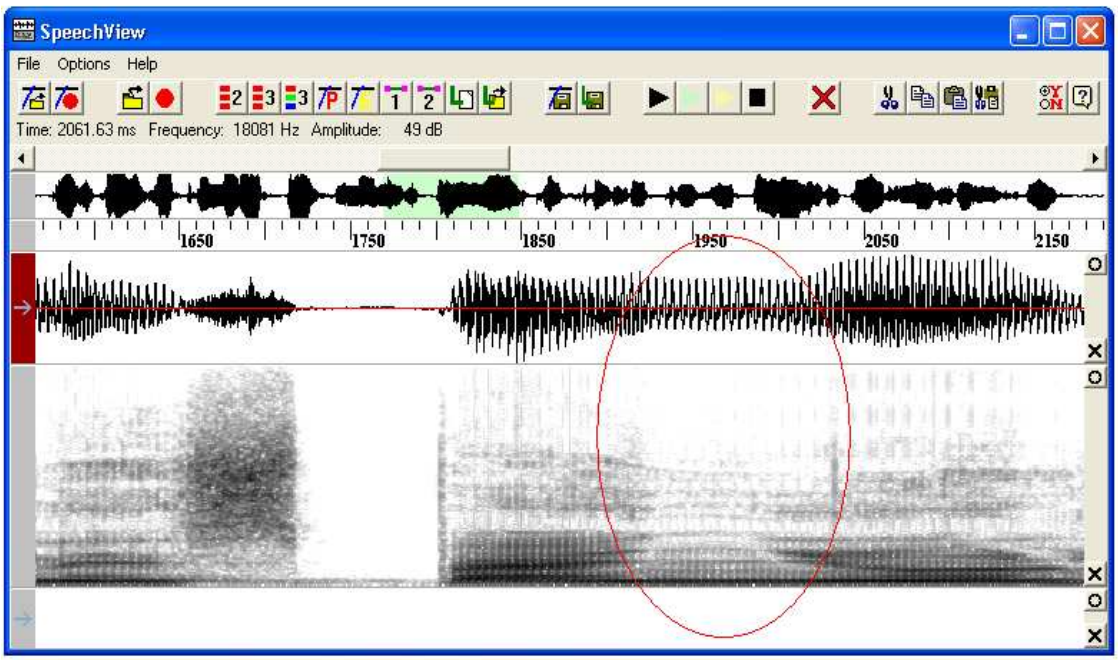

Figura 83. Imagen espectrográfica de [n~].

La delimitación de este alófono se hace en dos segmentaciones: la primera donde la frecuencia fundamental baja, y la segunda donde vuelve a ascender. Esta segmentación es la misma para los tres niveles de transcripción. El símbolo para la transcripción de este alófono es [n~], para los tres niveles de transcripción. En la Figura 84 se puede ver la imagen espectrográfica la palabra *español*. En esta imagen se señala la segmentación y transcripción del alófono [n~], en los tres niveles de transcripción.

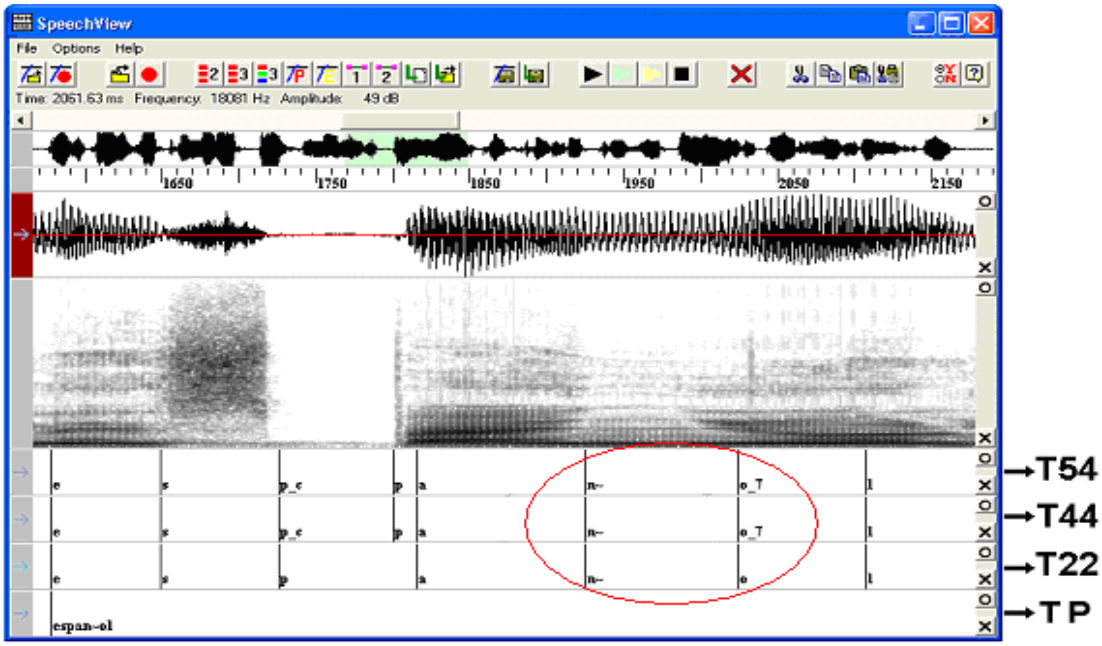

Figura 84. Imagen espectrográfica, segmentación y transcripción de [n~].

5. Fonemas líquidos /l/, /r(/, /r/

Se les llama fonemas líquidos al fonema lateral [l] y a los fonemas vibrantes [r(] y [r]. Quilis (1999:307) explica que la fonética acústica les rehabilitó este término "a causa de la existencia en estas consonantes de ciertas características que les infieren una fisonomía intermedia entre las vocales y las consonantes". Dichos fonemas tienen, al igual que las vocales, un generador armónico y articulatoriamente la abertura de la boca; de las consonantes poseen antirresonancia y obturación del canal longitudinal. Algunas veces la obturación y la abertura alternan como en el caso de las vibrantes (Alarcos 1965).

En el español de la ciudad de México, el fonema lateral alveolar tiene un único alófono: el lateral alveolar [l]. Algunos autores documentan una realización dental de /l/ (Perissinotto1975, Quilis 1999) y otra palatal (Perissinotto 1975). Sin embargo, en *Mexbet* el fonema alveolar lateral /l/ tiene un solo alófono en distribución complementaria: alveolar lateral [l], que ocurre en posición inicial y final de sílaba; ya que en el *Corpus DIME*, el alófono dentalizado presentó poca frecuencia de aparición (Cuétara 2004).

Los fonemas vibrantes articulatoriamente se caracterizan porque "el ápice de la lengua o la úvula, realiza varias oclusiones apoyándose en el órgano pasivo, es decir, los alvéolos o el dorso posterior de la lengua, respectivamente (aunque también existen vibrantes bilabiales, en las que un labio golpea contra otro, son poco frecuentes)" (Gil 1990:100). Acústicamente, se caracterizan por ser interruptos, porque durante su emisión existen intervalos de silencio; sin embargo estos fonemas se reconocen más por sus vibraciones que por sus oclusiones. Espectrográficamente, se diferencian por el número de vibraciones y oclusiones: "la vibrante simple sólo posee una interrupción, mientras que la múltiple tiene dos o más" (Martínez Celdrán 1998:94).

En *Mexbet* aparecen dos fonemas vibrantes: el simple /r(/ y el múltiple /r/. El fonema vibrante simple tiene un solo alófono: el alveolar vibrante simple [r(], que ocurre en inicio y final de sílaba, también puede aparecer en contexto intervocálico y en los grupos /pr/, /tr/,

/kr/, /br/, /gr/. El fonema alveolar vibrante múltiple, /r/, presenta un alófono: el alveolar vibrante múltiple, [r], este alófono ocurre en inicio de sílaba y en contexto intervocálico.

Estos alófonos, al igual que los oclusivos y las nasales, presentan alternancia en posición de coda silábica. Cuando al alófono vibrante simple [r(] está en posición de coda silábica, en ocasiones, se refuerza y es realizado como el alófono vibrante múltiple [r]. Matluck (1951:85) dice que este es un fenómeno de baja frecuencia y social: “Las gentes cultas y semicultas la pronuncian ordinariamente como vibrante sencilla sonora: *verde*, *carne*, *alarma*, *virgen*, *corcho*, *puerto*, etc. Algunas personas la refuerzan como *rr* vibrante múltiple, pero es raro”.

A continuación, en la Fgura 85, aparecen las reglas de distribución de fonema lateral, posteriormente se muestra el proceso de etiquetado de su alófono. Después, se mostrarán en las Figuras 85 y 88 las reglas distribucionales de los fonemas vibrantes, conforme a Cuétara 2004; y, finalmente, la segmentación y transcripción de sus alófonos.

5.1. Fonema alveolar lateral /l/

Fonema	Alófono	Contexto	Grafía
Alveolar lateral /l/	Alveolar lateral [l]	En inicio, interior o final de sílaba	*l*

Figura 85. Reglas distribucionales de los alófonos del fonema /l/.

El fonema alveolar lateral /l/ tiene sólo un alófono, el alveolar lateral [l], que ocurre en posición inicial y final de sílaba. Este alófono se detecta tanto en el oscilograma como en el espectrograma por una ligera disminución de energía, Croot y Taylor (1995:§3.5) señalan que los alófono líquidos “are vowel-like in appearance although usually of lower intensity than vowels. They display long formant transitions and gradual changes in intensity, and consequently have no clear boundaries adjacent to other liquids or glides, or vowels”. Lander (1997:50) da otra característica de la imagen del los fonemas líquidos: “The onset

of a liquid is marked by the disappearance of f3 (after a vowel) or the appearance of f1 and/or f2 (after a nasal or obstruent)". En la Figura 86 se puede observar la imagen espectrográfica correspondiente a [l]. En la imagen se puede constatar que [l] carece del tercer formante y posee baja energía con respecto a sus alófonos vecinos.

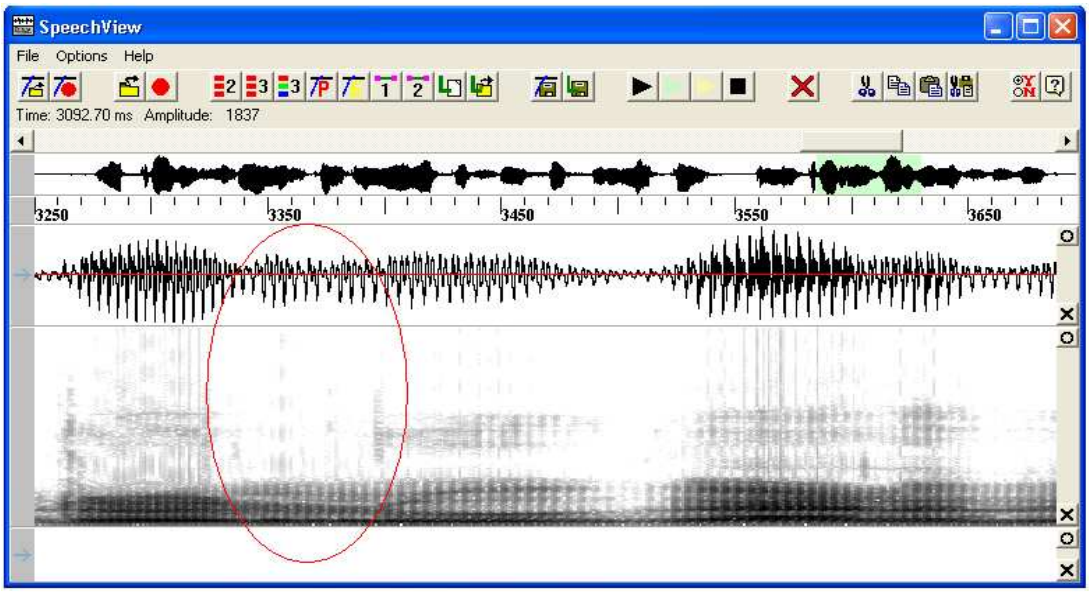

Figura 86. Imagen espectrográfica de [l].

La segmentación de este alófono se hace donde baja la frecuencia del primer formante, hasta donde vuelve a ascender. Machuca *et al*. (1999) sugieren que la segmentación se haga donde se ve el cambio de amplitud de la onda. La transcripción para las etiquetas es [l] en los tres niveles. En la Figura 87 se puede ver la imagen espectrográfica de la palabra *lugar*, en ella se señala la segmentación y la transcripción de este alófono.

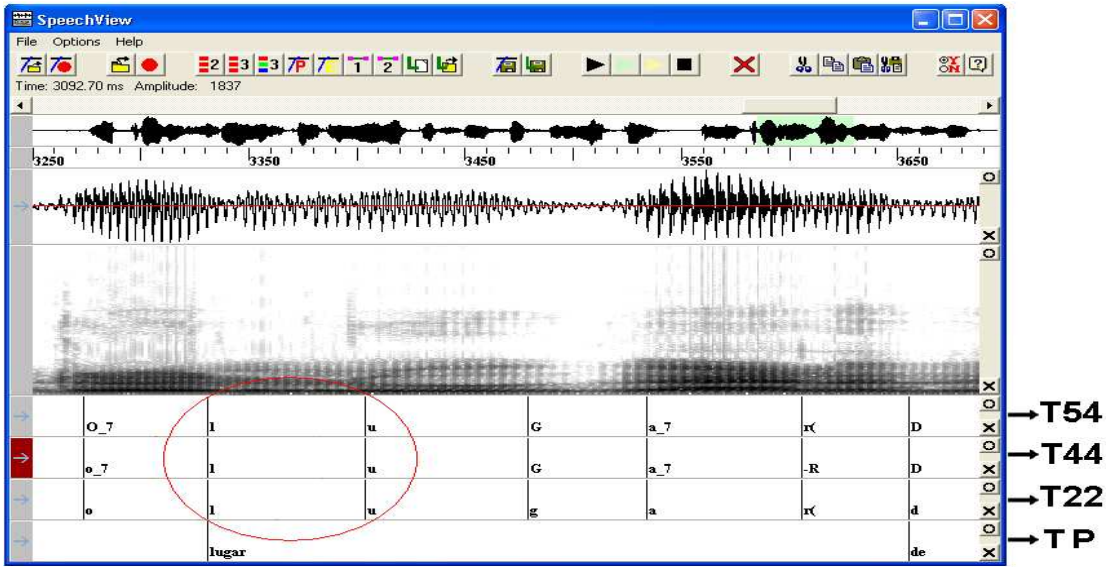

Figura 87. Imagen espectrográfica, segmentación y transcripción de [l].

5.2. Fonema alveolar vibrante simple /r(/

Fonema	Alófono	Contexto	Grafía
Alveolar vibrante simple /r(/	Alveolar vibrante simple [r(]	En inicio y final de sílaba	*r*

Figura 88. Reglas distribucionales de los alófonos del fonema /r(/.

El fonema alveolar vibrante simple /r(/ tiene un solo alófono, el alveolar vibrante simple [r(], que ocurre en posición inicial y final de sílaba, y en los grupos [pr], [tr], [kr], [br], [gr] y [fr].

En el oscilograma, este alófono se localiza por mostrar breves intervalos de frecuencia alta y baja. En el espectrograma se distingue por estar conformado de tres segmentos: en el primero se ve una pequeña vibración, el segundo se presenta como un segmento parecido al cierre de las oclusivas y en el tercero vuelve a haber otra pequeña vibración. En la Figura 89 se señala el segmento espectrográfico de [r(]. En el se pueden ver claramente los tres segmentos descritos anteriormente.

Figura 89. Imagen espectrográfica de [r(].

La segmentación de esté alófono se hace de donde comienza la primera vibración hasta el final de la segunda. En la Figura 90 se puede observar la segmentación de la imagen espectrográfica del alófono vibrante simple.

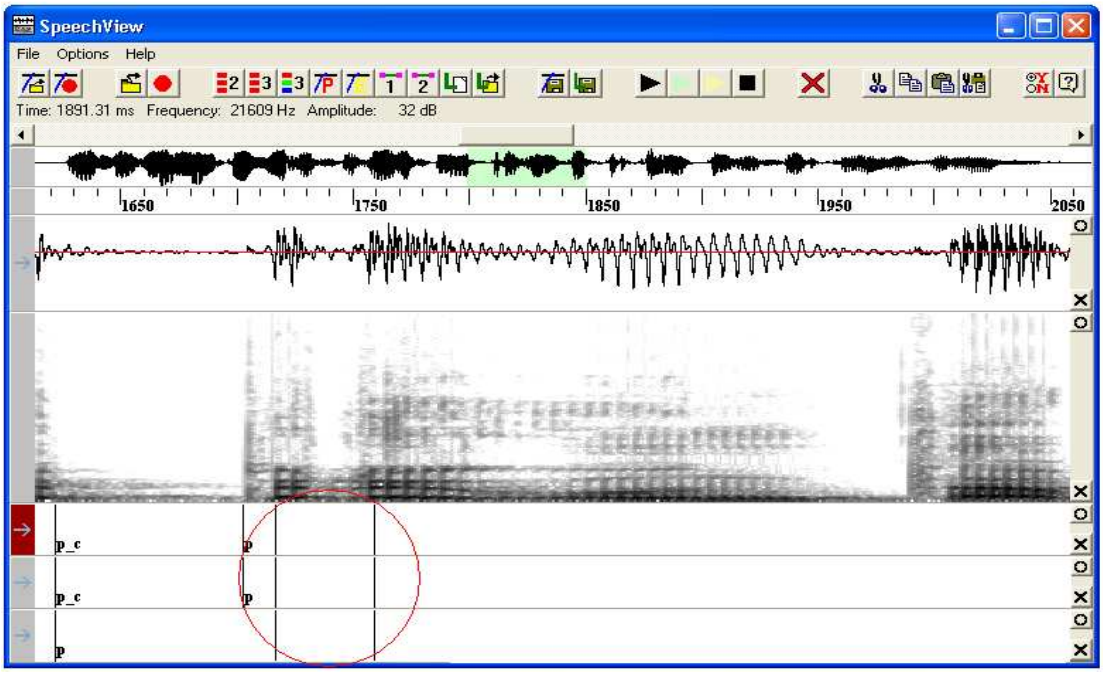

Figura 90. Imagen espectrografía y segmentación de [r(].

La transcripción de este alófono es [r(] para los tres niveles de transcripción. En la Figura 91 se puede observar la imagen espectrográfica de la palabra *pregunta*, en ella se resalta la transcripción de [r(] en los tres niveles.

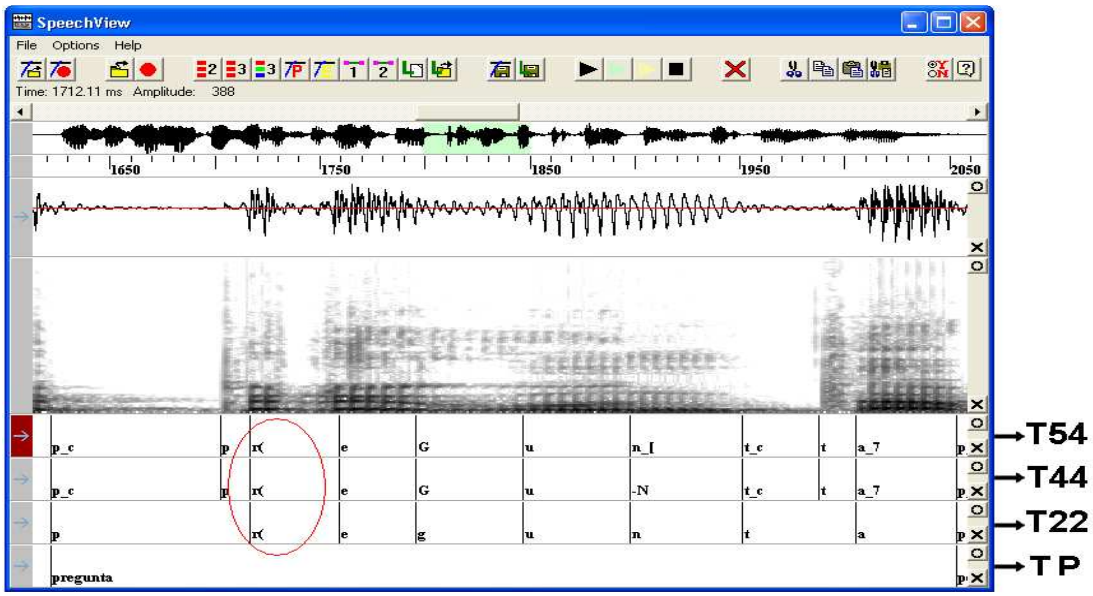

Figura 91. Imagen espectrográfica y transcripción de [r(].

Cuando [r(] está en posición de coda silábica, se transcribe en el nivel T44 con el símbolo [-R]; en los otros niveles se transcribe con la misma grafía [r(]. En la Figura 92 se muestra el ejemplo de transcripción para la vibrante simple en posición de coda, en la palabra *hermosa*.

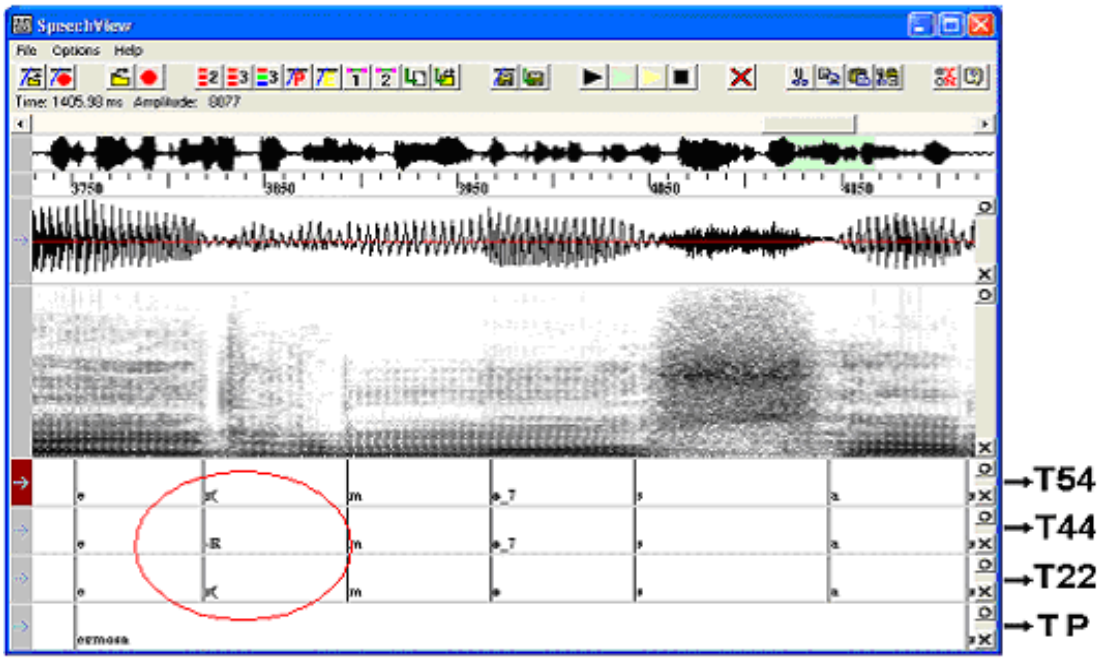

Figura 92. Imagen espectrográfica y transcripción de [-R].

5.3. Fonema alveolar vibrante múltiple /r/

El fonema alveolar vibrante múltiple tiene un solo alófono: el alveolar vibrante múltiple, [r], que se realiza en posición inicial de sílaba y en contexto posterior a [n, l, s]. En la Figura 93 aparece su regla distribucional, conforme a Cuétara 2004.

Fonema	Alófono	Contexto	Grafía
Alveolar vibrante múltiple /r/	Alveolar vibrante múltiple [r]	En inicio de sílaba y posterior a [n, l, s]	*r*, *rr*

Figura 93. Reglas distribucionales de los alófonos del fonema /r/.

En el oscilograma, este alófono se localiza porque muestra pequeños intervalos de baja y alta frecuencia. Su imagen espectrográfica se caracteriza por tener de tres a cuatro vibraciones durante su emisión. En la Figura 94 se puede observar la imagen espectrográfica de [r] señalada con un círculo.

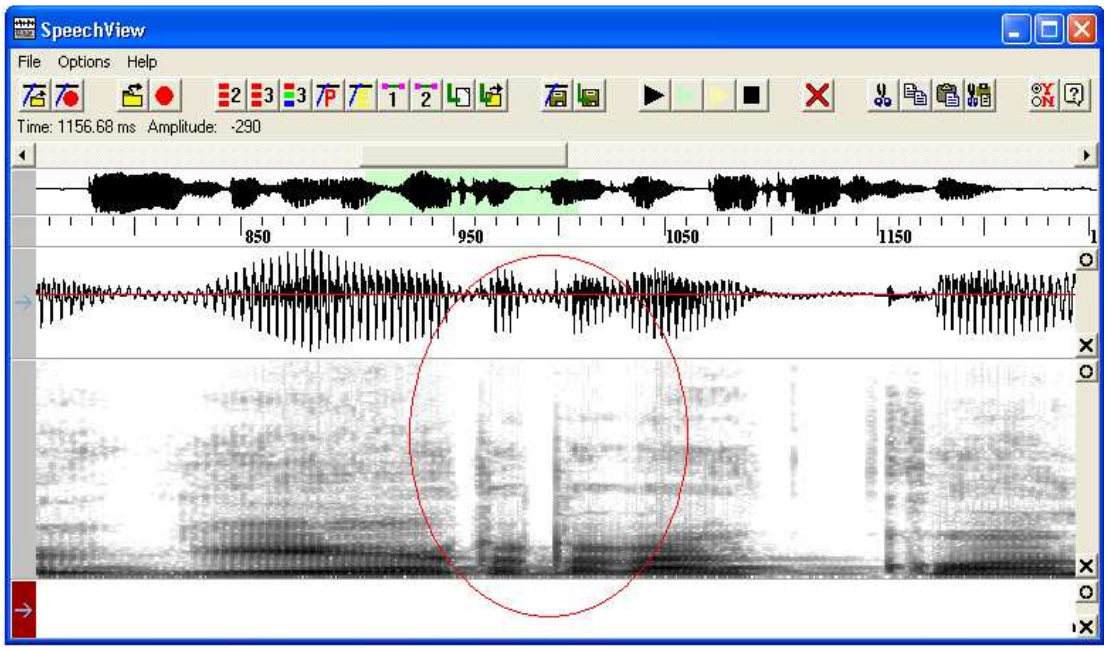

Figura 94. Imagen espectrográfica de [r].

La delimitación de este alófono se hace de la primera vibración hasta el final de la última que se detecten en el espectrograma. Esta segmentación se hace en los tres niveles de transcripción. En la Figura 95 se puede observar la segmentación de la imagen espectrográfica de este alófono en los tres niveles de transcripción.

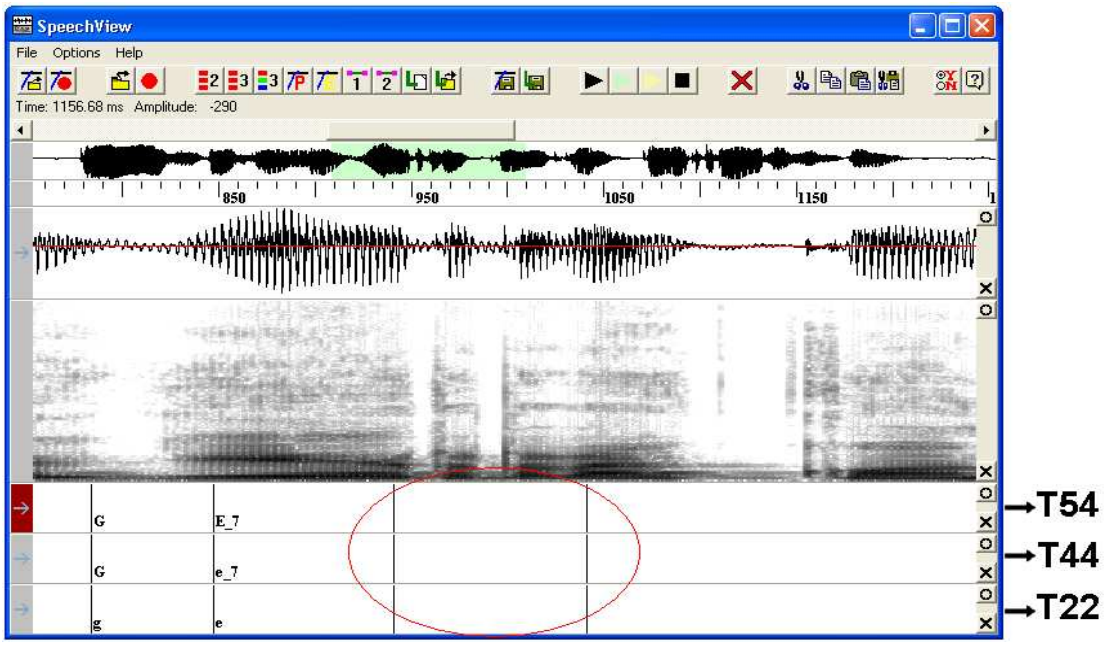

Figura 95. Imagen espectrográfica y segmentación de [r].

La transcripción de este alófono es [r] para las etiquetas de los tres niveles de transcripción. En la Figura 96 se puede observar la imagen espectrográfica de la palabra guerra, en la cual se resalta la transcripción del alófono vibrante múltiple.

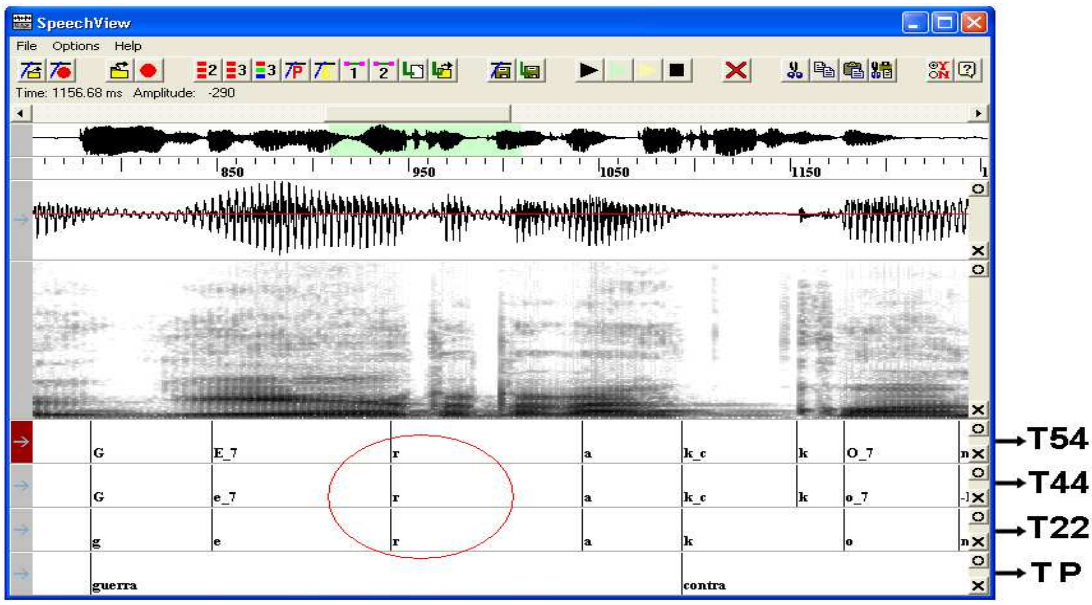

Figura 96. Imagen espectrográfica y transcripción de [r].

II. FONEMAS VOCÁLICOS

1. Fonemas vocálicos /a/, /e/, /i/, /o/, /u/

El español de México tiene cinco fonemas vocálicos, los cuales se dividen por su abertura en abiertos, medios y cerrados. La abertura de cada uno depende de la proximidad de la lengua al velo del paladar (Quilis 1999). De esta manera [a] es abierta, [e] y [o] medias, [i] y [u] cerradas. Por la zona de articulación se dividen en anteriores, centrales y posteriores; estos términos hacen referencia a la región que ocupa la lengua en el paladar al emitir la vocal; así, [i] y [e] son anteriores porque se articulan en el paladar duro; [a] es central ya que la posición de la lengua no va hacia el velo del paladar ni hacia al paladar duro y toma una posición semiplana (Gili 1975); [u] y [o] son posteriores, porque se articulan en el paladar blando. Durante la realización de los fonemas vocálicos el aire pasa libremente por la cavidad bucal, ya que ningún órgano impide su salida (Garrido *et al.* 1998)

Cada fonema vocálico tiene sus propios alófonos en distribución complementaria. De manera que el fonema vocálico central abierto /a/ tiene tres alófonos: uno prototípico [a], otro palatal [a_j] y uno más velar [a_2]. El fonema anterior medio /e/ tiene dos alófonos: uno prototípico [e], y otro abierto [E]. El anterior cerrado /i/ tiene un alófono prototípico y una paravocal [j]. El posterior medio tiene un alófono prototípico [o] y otro abierto [O]. Finalmente, el fonema posterior cerrado /u/ tiene un alófono prototípico [u] y una paravocal [w] (Cuétara 2004). La realización de los alófonos está condicionada por reglas combinatorias (las cuales se exponen en el apartado de cada fonema vocálico); aunque cabe mencionar que, en ocasiones, puede aparecer una realización distinta a la esperada en cierto contexto o regla.

Espectrográficamente, las vocales se pueden diferenciar de las consonantes porque son segmentos de mayor energía, la cual se ve reflejada por el oscurecimiento de su segmento, y también por tener una estructura formántica muy estable. Cada fonema vocálico tiene sus propios formantes, de los cuales los dos primeros son los más importantes para el reconocimiento de cada vocal (Quilis 1999), ya que reflejan la abertura y la zona de articulación de la vocal.

El primer formante (F1) se relaciona con la abertura de la cavidad bucal. Entre más abierta sea la vocal mayor será la frecuencia del F1. El segundo formante (F2) se relaciona con el punto de articulación. Cuando la vocal sea anterior, el F2 aparecerá a una alta frecuencia, y viceversa, cuando la vocal sea posterior aparecerá a una baja frecuencia (Gil 1990). De esta manera, el F1 de [a] es el de más alta frecuencia por ser la vocal con mayor abertura, y su F2 aparece a una altura media por su articulación central. El F1 de [e] tiene una altura media y el F2 alta. El F1 de [i] tiene una frecuencia baja, por lo que aparece en la parte inferior del espectrograma, y su F2 tiene una frecuencia alta. En [o] el F1 tiene una frecuencia media y el F2 baja. Finalmente, en [u] ambos formantes aparecen con una muy baja frecuencia. En la Figura 97 se puede observar la frecuencia de los formantes vocálicos, conforme a los parámetros anteriores.

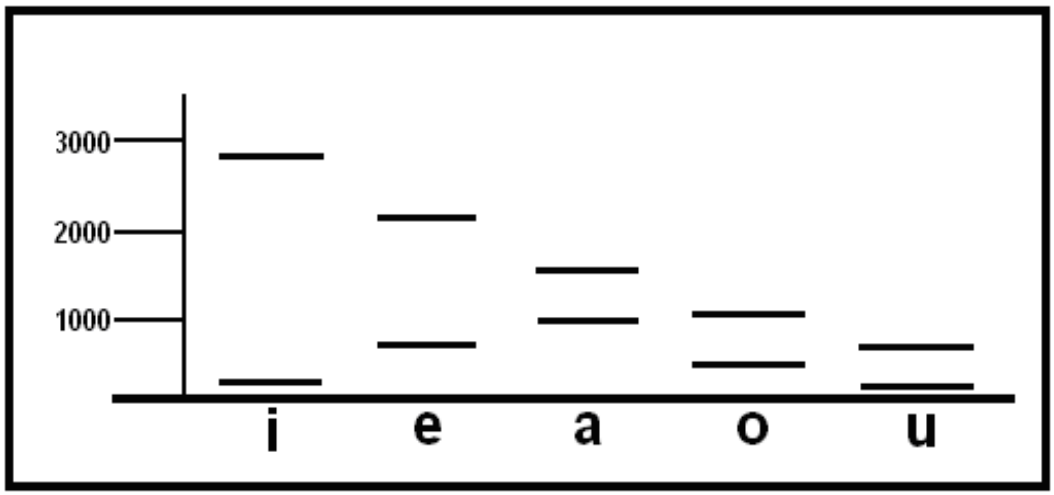

Figura 97. Frecuencias de los primeros formantes vocálicos (Quilis 1999).

En cuanto a las transiciones de dichos formantes, el segundo es el que más varía y se moverá dependiendo la articulación de cada alófono. En el caso de [a], cuando palataliza [a_j], la transición del segundo formante sube, es decir, es positiva. Por el contrario, cuando velariza [a_2] la transición es negativa.

En los alófonos abiertos de las vocales medias [e] y [o], el segundo formante regularmente se mueve hacia la frecuencia del F2 de [a] (Alarcos 1965:149). En las paravocales [j] y [w], de las vocales altas /i/ y /u/, el primer formante siempre aparecerá a una baja frecuencia.

Croot y Taylor (1995) mencionan que en que la paravocal [w] es común que no aparezcan el segundo y tercer formante. En ambas paravocales, la transición siempre será positiva porqué se unen a una vocal con más alto grado de abertura.

En las vocales tónicas se pueden observar las mismas características formánticas, que en las átonas; sin embargo, en las primeras se observará un grado mayor de energía y de duración en el segmento, lo que hará que se diferencien de las vocales átonas (*Albalá et al.* 2008). En *Mexbet* se tienen un inventario de nueve alófonos tónicos que son los mismos que en las vocales átonas, pero se excluyen las paravocales [j], [w]. Se les representa por medio del diacrítico [_7]. En la Figura 98 se puede observar la representación fonética de los alófonos tónicos en *Mexbet*.

Vocal tónica	Representación en *Mexbet*
a	a_j_7
	a_7
	a_2_7
e	e_7
	E_7
i	i_7
o	o_7
	O_7
u	u_7

Figura 98. Representación fonética de los alófonos vocálicos tónicos en *Mexbet.*

1.1. Paravocales [j], [w]

Son llamados *paravocales* aquellos alófonos que ocupan el lugar de margen silábico en un diptongo[2] (Cuétara 2004). En *Mexbet* se tienen dos formas paravocálicas: [j] y [w], las cuales aparecen como alófonos de /i/ y /u/ (Cuétara 2004). En la cadena hablada podemos encontrar dos tipos de diptongos llamados *creciente* y *decreciente*. En los primeros, la cavidad oral va de cerrada a abierta. A este grupo pertenecen las combinaciones [j], [w] + [a], [e], [o]; como en las palabras *rueda* [rweda] y *viejo* [bjexo]. En el segundo tipo de diptongos, la cavidad oral va de abierta a cerrada, ya que primero se pronuncia la vocal nuclear y posteriormente la paravocal. En esta posición pueden aparecer las combinaciones [a], [e], [o] + [j], [w], como en las palabras *auto* [awto], *hoy* [oj], *deuda* [deuda] (Quilis 1999). Algunos autores mencionan que si la paravocal aparece en el margen anterior al núcleo es una semiconsonante, pero si aparece en posición posterior al núcleo es una semivocal (Perissinotto 1975, Gili G. 1975; Gil 1990, Quilis 1999; Hidalgo y Quilis 2004).

Cada uno de los diptongos tiene su imagen espectrográfica característica, en la cual la transición del primer formante es el parámetro para saber cuándo el diptongo es creciente o decreciente. En el diptongo creciente, el F1 irá en ascenso durante su emisión, pues como se dijo en el párrafo anterior, en la primera posición está la paravocal y en segunda la vocal. En los diptongos decrecientes sucede lo contrario, el primer formante va en descenso, ya que en primera posición se encuentra la vocal y en segunda la paravocal; por lo tanto los órganos se desplazan de abiertos a cerrados (ver Figura 129).

1.2. Fonema vocálico central abierto /a/

El fonema vocálico central abierto tiene tres alófonos: uno abierto [a], que se realiza en cualquier posición de sílaba; otro palatalizado [a_j], que ocurre cuando está en contexto anterior a [tS, n~, Z, j]; y otro velarizado [a_2], que se realiza cuando está en contexto anterior a [u, x] o en silaba trabada por [l]. En la Figura 99 se muestra un cuadro con las

[2] En el *Proyecto DIME* se hizo una tesis de licenciatura sobre los diptongos del español de México. *El estudio de los diptongos del español de México para su aplicación en un reconocedor de habla* (López 2004).

reglas combinatorias del fonema /a/, conforme a Cuétara 2004. Perissinotto (1975) señala que de las tres realizaciones del fonema /a/, la más frecuente es la prototípica [a], le sigue la velar [a_2] y, finalmente, la palatal [a_j] que se presenta esporádicamente.

Fonema	Alófono	Contexto	Grafía
Vocálico central abierto /a/	Central abierto	En cualquier posición de sílaba	*a*
	Abierto palatal [a_j]	En contexto anterior a [tS, n~, Z, j]	*a*
	Abierto velar [a_2]	En contexto anterior a [u, x] y en sílaba trabada por [l]	*a*

Figura 99. Reglas distribucionales de los alófonos del fonema /a/.

1.2.1. Alófono central abierto [a]

El alófono central abierto [a] ocurre en cualquier posición de sílaba, siempre y cuando no se encuentre en alguno de los contextos donde se realizan los alófonos velar y palatal. En el espectrograma, el alófono central [a] se caracteriza por ser un segmento de alta energía; por lo que tiene un oscurecimiento mayor al de los segmentos que lo preceden y anteceden. También se caracteriza porque su primer y segundo formantes están bien definidos. En la Figura 100 se puede observar la imagen espectrográfica de [a].

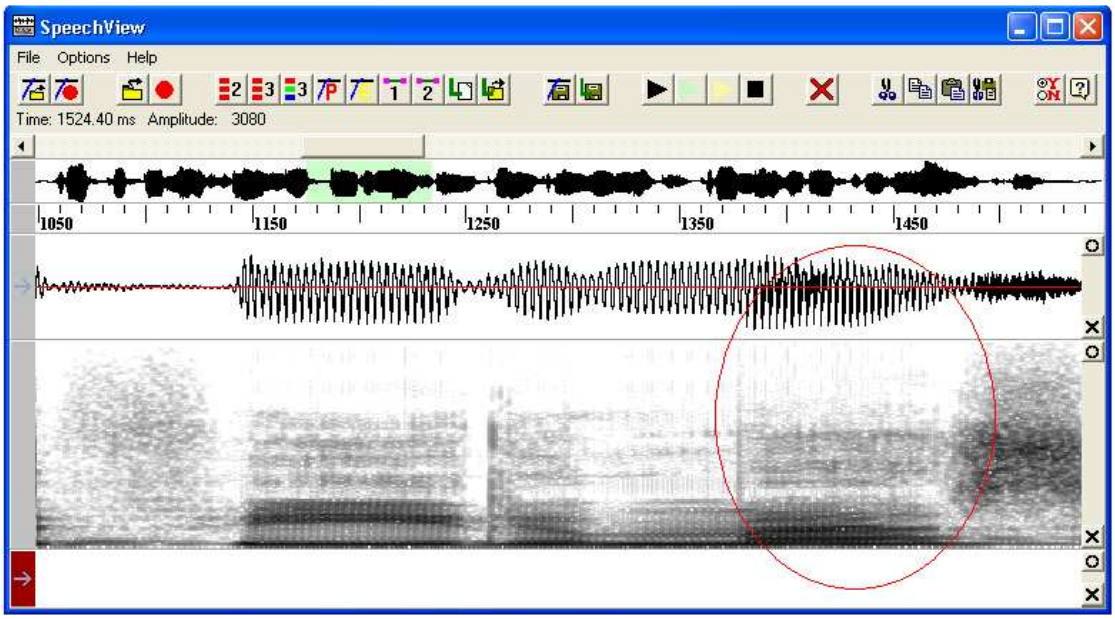

Figura 100. Imagen espectrográfica de [a].

La segmentación de este alófono se hace, si se ve el oscilograma, en los cambio de amplitud de la onda; es decir, de dónde sube la frecuencia hasta donde vuelve a descender. Si se ve el espectrograma, la segmentación va de donde comienza la estructura formántica de [a] hasta donde los formantes se desvanecen, o bien, donde comienza la frecuencia del siguiente alófono. En la Figura 101 se puede observar la segmentación de la imagen acústica de [a] en los tres niveles de transcripción.

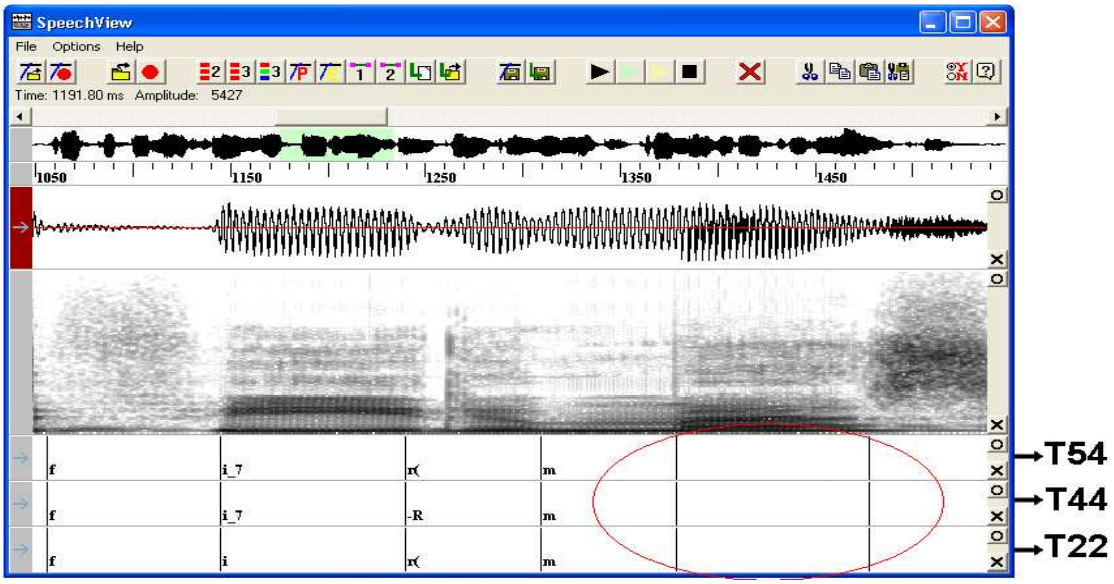

Figura 101. Imagen espectrográfica y segmentación de [a].

La transcripción para este alófono es [a] en los tres niveles de transcripción (T54, T44 y T22). En la Figura 102 se muestra la imagen espectrográfica de la palabra *firmas*, transcrita en los tres niveles, en ella se ha resaltado la transcripción de [a].

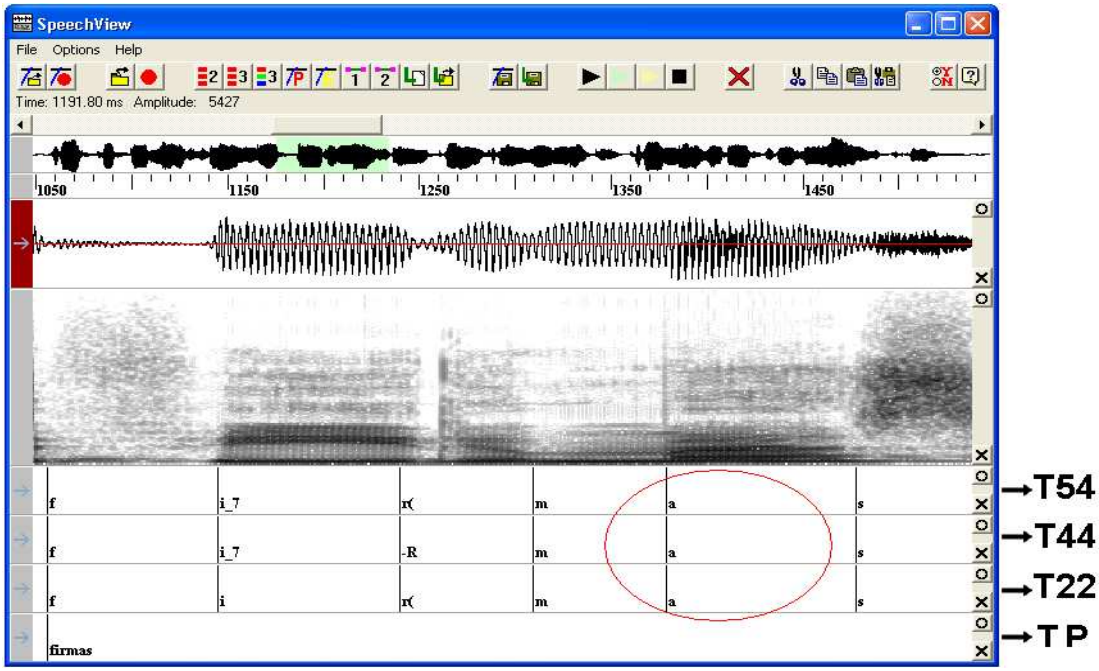

Figura 102. Imagen espectrográfica y transcripción de [a].

1.2.2. Alófono abierto palatal [a_j]

El alófono palatal [a_j] ocurre cuando se encuentra en posición anterior a los alófonos palatales [tS, n~, Z, j], como en las palabras *muchacho*, *año*, *ayer*, etc. En el espectrograma se distingue por ser un segmento de alta energía. La palatalización de este alófono se puede notar por la elevación en la transición del segundo formante con respecto al alófono que lo precede. En la Figura 103 se puede observar la imagen del alófono palatalizado, señalada con un círculo.

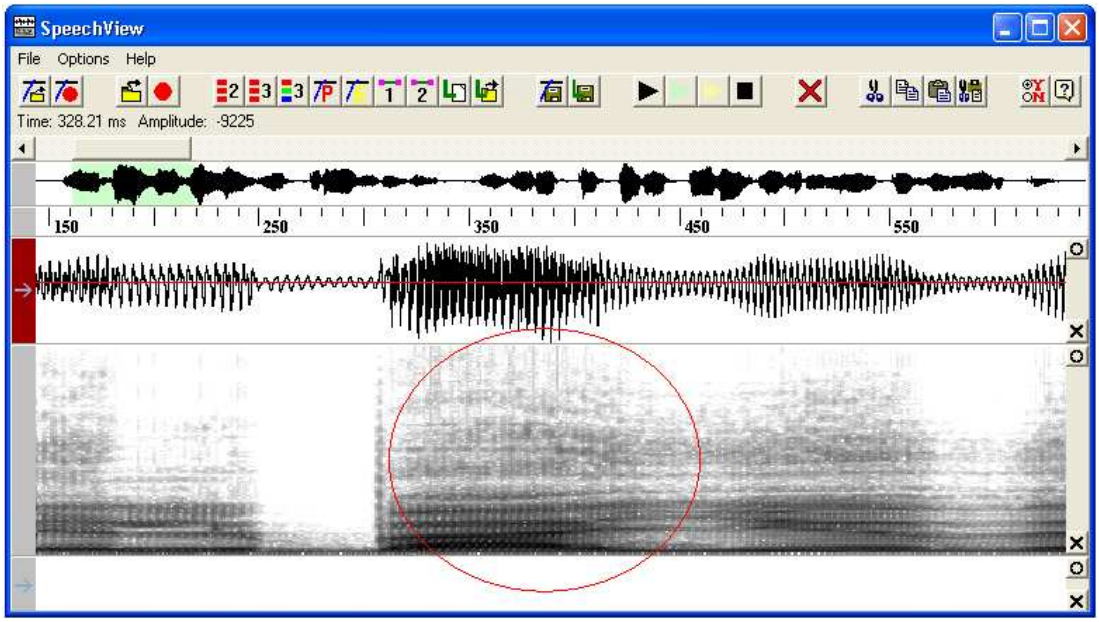

Figura 103. Imagen espectrográfica de [a_j].

La segmentación de este alófono, si se ve en el oscilograma, se hace de donde sube la amplitud de la onda hasta donde vuelve a descender. En el espectrograma, la segmentación se hace donde comienza la región formántica de [a_j] hasta donde el formante sube para unirse al del alófono palatal. También se puede notar por el cambio de energía en los formantes, ya que estos parecieran difuminarse. Esta misma segmentación se hace para los tres niveles de transcripción. En la Figura 104 se puede observar la segmentación de la imagen del alófono palatal [a_j].

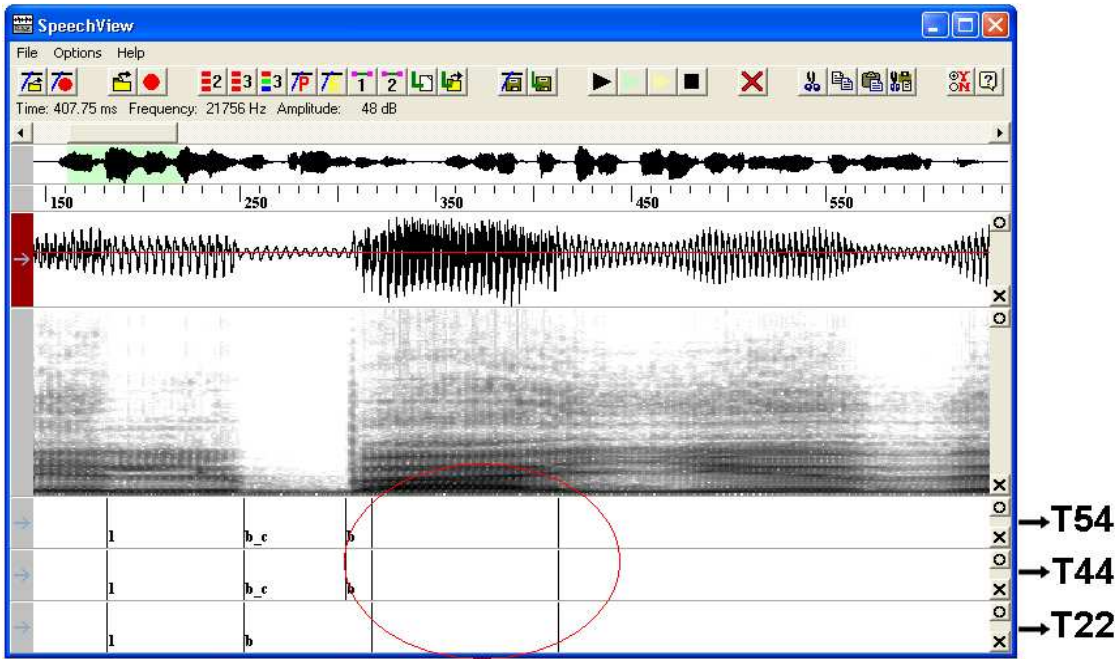

Figura 104. Imagen espectrográfica y segmentación de [a_j].

La transcripción de este alófono es [a_j] para la etiqueta del nivel T54, y [a] para los niveles T44 y T22. En la Figura 105 se puede observar la imagen espectrográfica de la palabra *valle*, en ella se ha resaltado la transcripción del alófono palatal en los tres niveles.

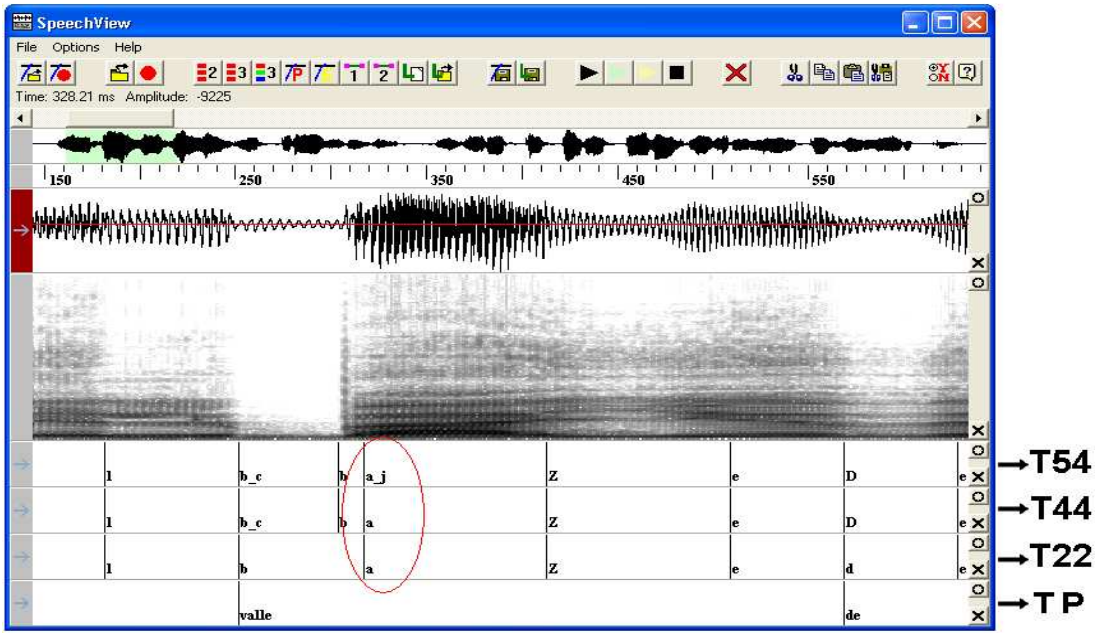

Figura 105. Imagen espectrográfica y transcripción de [a_j].

1.2.3. Alófono abierto velar [a_2]

El alófono abierto velar [a_2] ocurre cuando está en contexto anterior a [u, x] y en sílaba trabada por [l], como en las palabras *ajo*, *aullar*, *sal*. En el espectrograma se puede identificar por que la transición de sus formantes es decreciente con respecto al alófono que lo precede. En la Figura 106 se puede ver el segmento espectrográfico del alófono velar [a_2]. En éste se puede observar el primer formante de alta energía, cuya transición desciende con respecto al alófono siguiente.

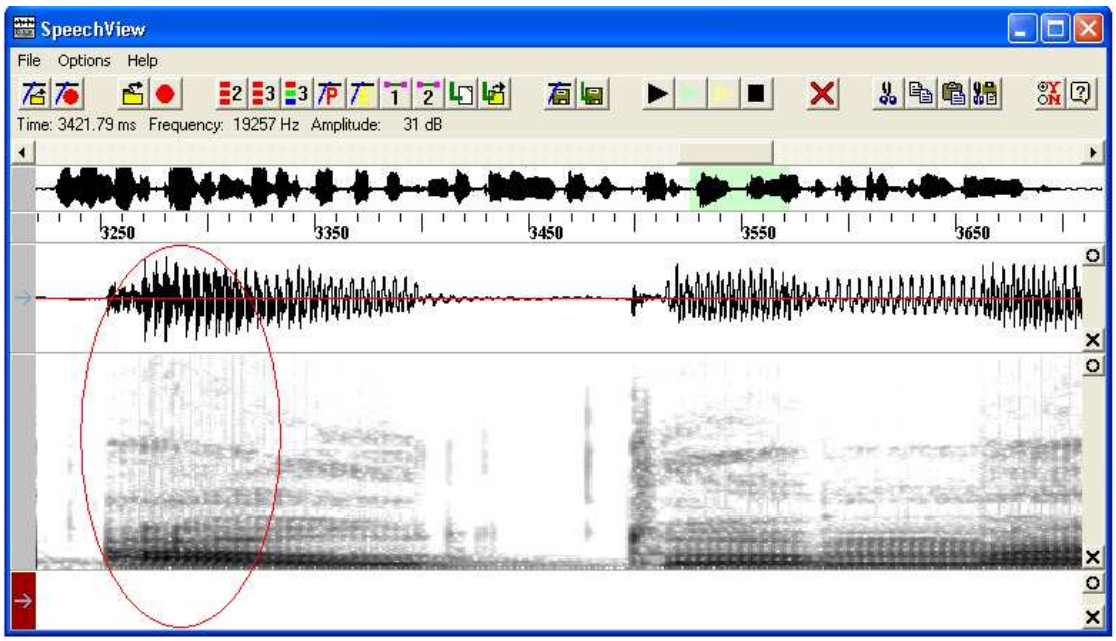

Figura 106. Imagen espectrográfica de [a_2].

La delimitación de este alófono se hace en dos segmentaciones: la primera donde comienza la estructura formántica de [a_2], hasta la transición de los formantes. Si esta transición no se distingue, la segmentación se hace a la mitad de ambos segmentos (Croot y Taylor 1999). Esta misma segmentación se hace para los tres niveles de transcripción. En la Figura 107 se muestra una imagen espectrográfica, en la que se puede ver cómo se ha hecho la segmentación de la imagen del alófono velar, en los tres niveles de transcripción.

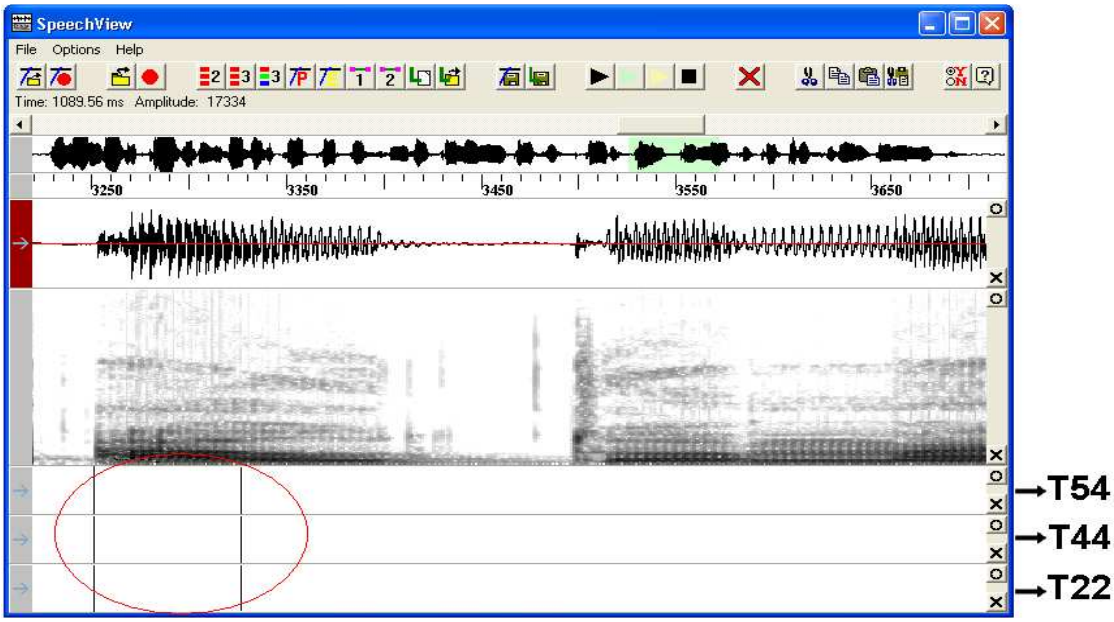

Figura 107. Imagen espectrográfica y segmentación de [a_2].

La transcripción de este alófono es [a_2] para la etiqueta del nivel T54; para las etiquetas de los niveles T44 y T22 es [a]. En la Figura 108 se muestra la imagen espectrográfica segmentada y transcrita de la palabra *automático*, en la cuál se señala la transcripción del alófono velar en cada uno de los niveles de transcripción.

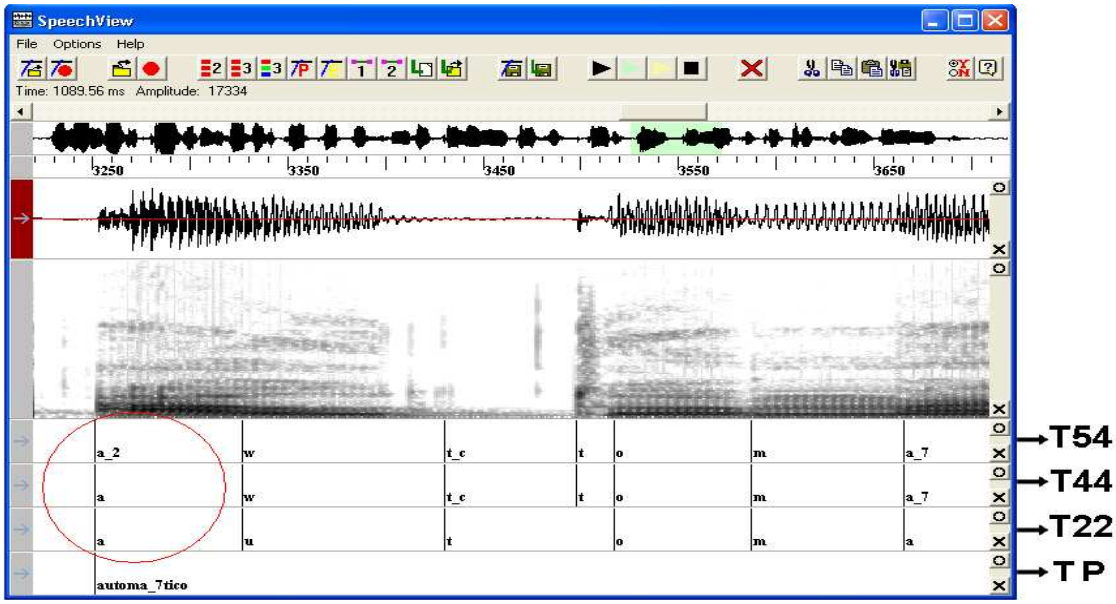

Figura 108. Imagen espectrográfica y transcripción de [a_2].

1.3. Fonema medio palatal /e/

El fonema medio palatal tiene dos alófonos: uno medio palatal [e], que se realiza en cualquier posición de sílaba; y el alófono medio palatal abierto [E], que ocurre cuando está en posición anterior y posterior al alófono alveolar vibrante múltiple. En la Figura 109 se muestran dichas reglas distribucionales, conforme a Cuétara 2004.

Fonema	Alófono	Contexto	Grafía
Vocálico medio palatal /e/	Medio palatal [e]	En cualquier posición de sílaba	*e*
	Medio palatal abierto [E]	En contexto anterior y posterior a [r]	*e*

Figura 109. Reglas distribucionales de los alófonos del fonema /e/.

1.3.1. Alófono medio palatal [e]

El alófono medio palatal, prototípico [e], en el espectrograma, se distingue por ser un segmento de corta duración, cuyos formantes aparecen a una frecuencia media. En la Figura 110 se puede observar el segmento espectrográfico de [e].

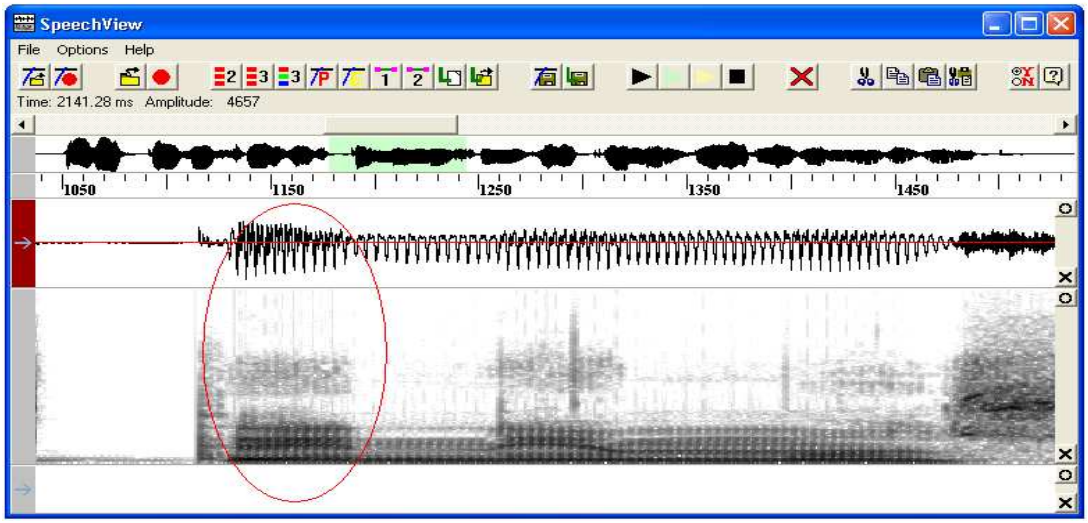

Figura 110. Imagen espectrográfica de [e].

La segmentación de este alófono se hace de la misma manera que el alófono central abierto (remitirse a [a]). La transcripción para las etiquetas de los tres niveles es [e]. En la Figura 111 se puede observar la imagen espectrográfica de la palabra *tenemos*, en la cual se ha resaltado la segmentación y transcripción de [e] en los tres niveles.

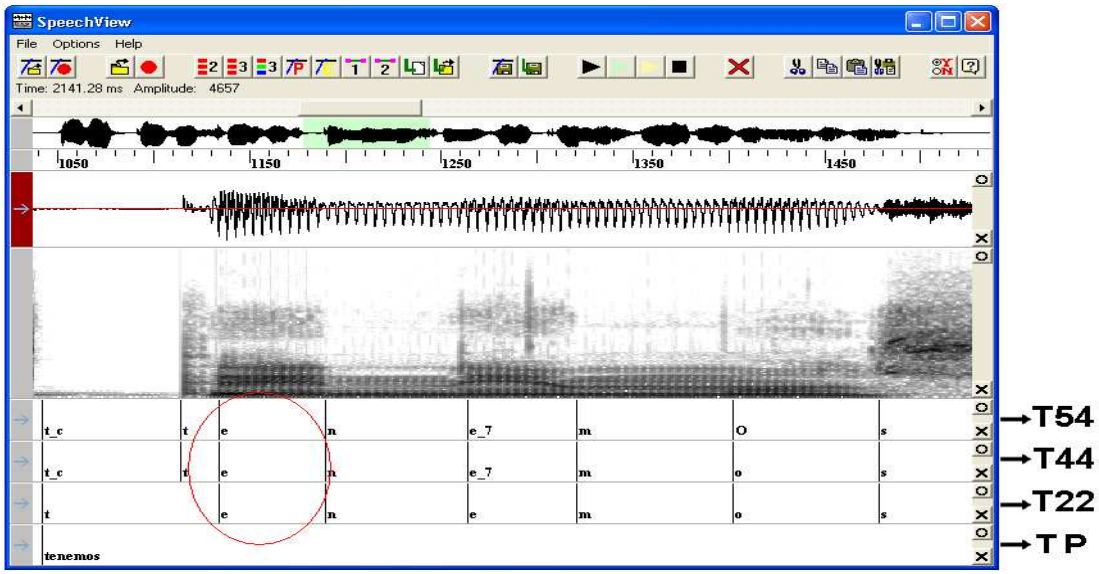

Figura 111. Imagen espectrográfica, segmentación y transcripción de [e].

1.3.2. Alófono medio palatal abierto [E]

El fonema medio palatal se abre cuando está en contacto con el alófono vibrante múltiple [r], como en las palabras *cerrar*, *correr*, *recio*, etc., dando como resultado, la realización del alófono medio palatal abierto [E]. Este alófono se puede ver en el espectrograma como un segmento de amplia duración, y siempre estará condicionado a parecer junto a las vibraciones de [r]. En la Figura 112 se puede observar la imagen del alófono medio palatal abierto [E]. En esta imagen se puede ver cómo aparece junto a las vibraciones del alófono vibrante múltiple. Cabe mencionar que a diferencia del alófono prototípico su imagen espectrográfica es de mayor duración.

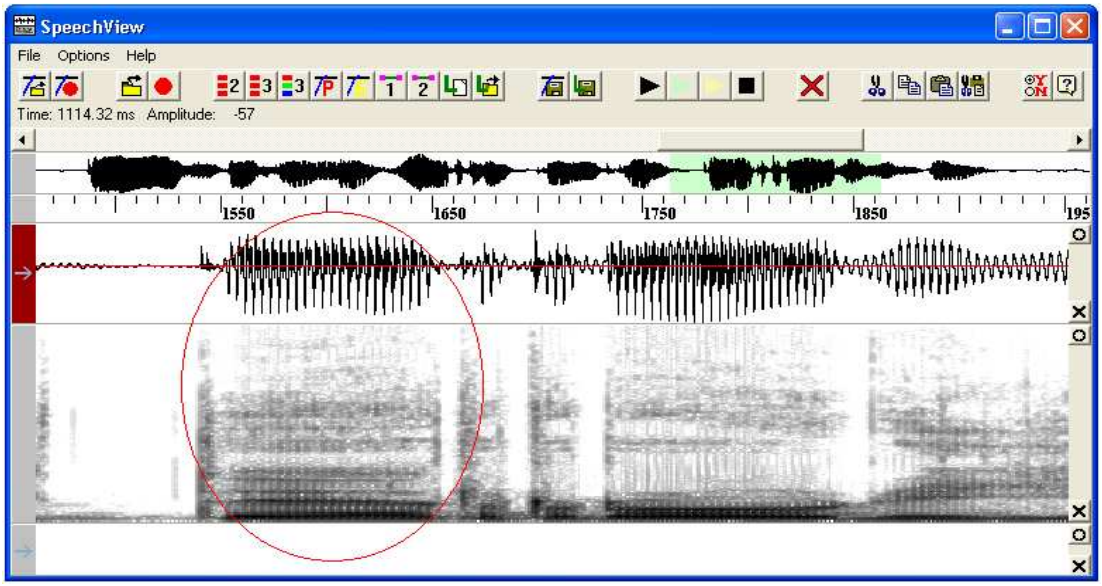

Figura 112. Imagen espectrográfica de [E].

La segmentación de este alófono se hace de donde termina [r] hasta la transición formántica de [E] con el alófono siguiente. Cuando [E] está antes de la vibrante múltiple, la segmentación se hará desde del comienzo de la estructura formántica de [E] hasta el inicio de [r], el cuál se pude notar por el debilitamiento de los formantes de la vocal. En la Figura 113 se puede observar la segmentación del alófono vocálico, el cual aparece precedido por la vibrante múltiple [r].

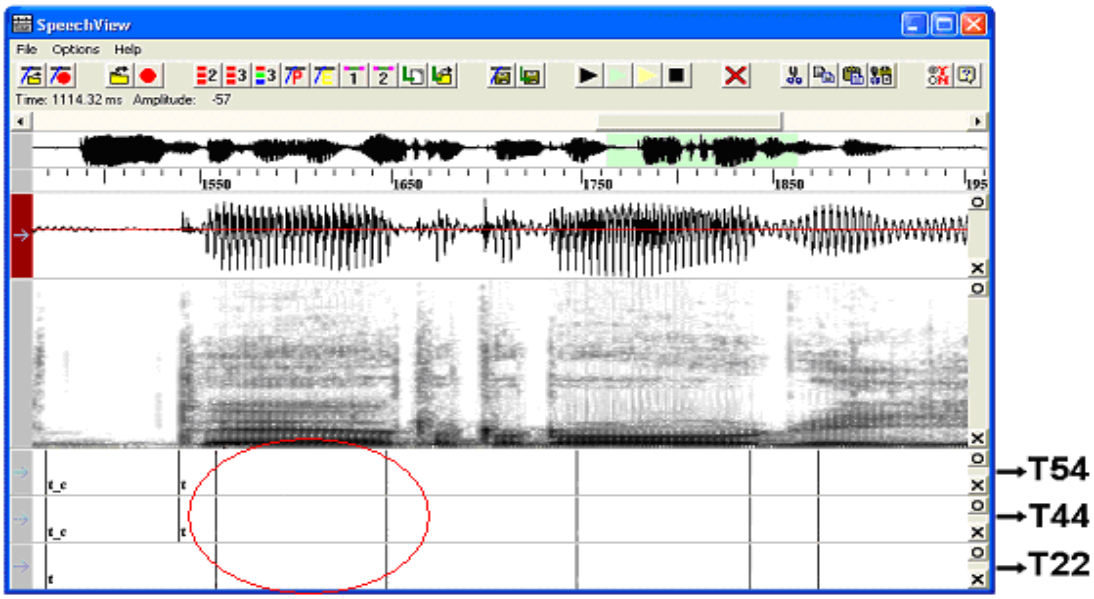

Figura 113. Imagen espectrográfica y segmentación de [E].

La transcripción de este alófono para el nivel T54 es [E]; y [e] para los niveles T44 y T22. En la Figura 114 se puede observar la imagen espectrográfica de la palabra *terrorismo*, en la cual se resalta la transcripción del alófono [E].

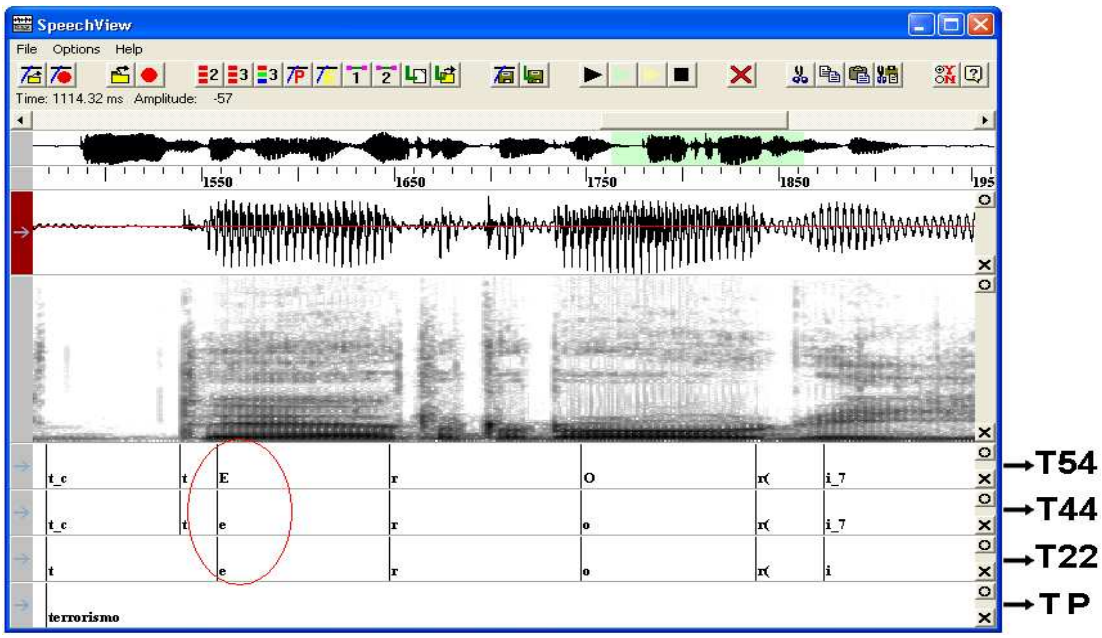

Figura 114. Imagen espectrográfica y transcripción de [E].

1.4. Fonema vocálico cerrado palatal /i/

El fonema vocálico cerrado palatal tiene dos alófonos: uno cerrado palatal, prototípico, [i], que ocurre en cualquier posición de sílaba; y otro paravocal palatal [j], que se realiza en contexto anterior o posterior a las vocales [a, e, o, u]. En la Figura 115 se muestran las reglas distribucionales de este fonema, de acuerdo con Cuétara 2004.

Fonema	Alófono	Contexto	Grafía
Vocálico cerrado palatal /i/	Cerrado palatal [i]	En cualquier posición de sílaba.	*i, y*
	Paravocal palatal [j]	En contexto anterior y posterior a [a, e, o, u]	*i, y*

Figura 115. Reglas distribucionales de los alófonos del fonema /i/.

1.4.1. Alófono cerrado palatal [i]

El alófono cerrado palatal [i] se distingue en el espectrograma por ser un segmento breve de mayor energía que las consonantes, pero menor a las de los alófonos vocálicos [a] y [e]. Sus formantes se caracterizan por aparecer separados: F1 con baja frecuencia y F2 a una alta frecuencia. En la Figura 116 se puede observar la imagen acústica de [i].

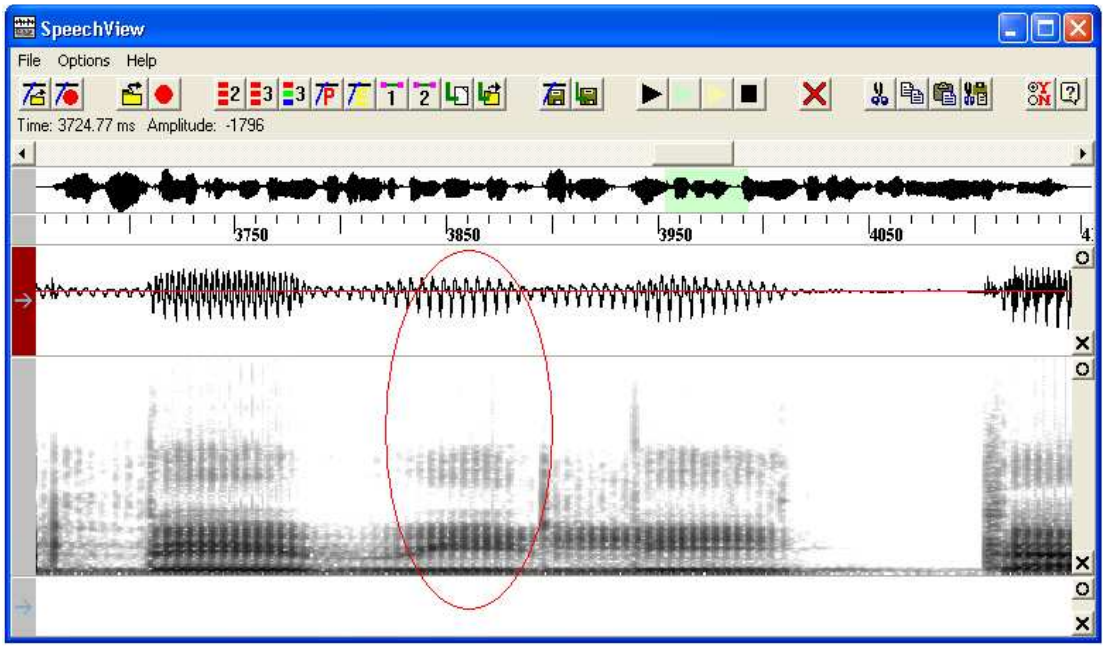

Figura 116. Imagen espectrográfica de [i].

Para la segmentación de este alófono seguir el mismo proceso del alófono central abierto [a]. La transcripción es [i] para las etiquetas de los tres niveles. En la Figura 117 se puede ver la imagen espectrógrafica de la palabra *debilitamiento*, en la cual muestra la segmentación y la transcripción de [i].

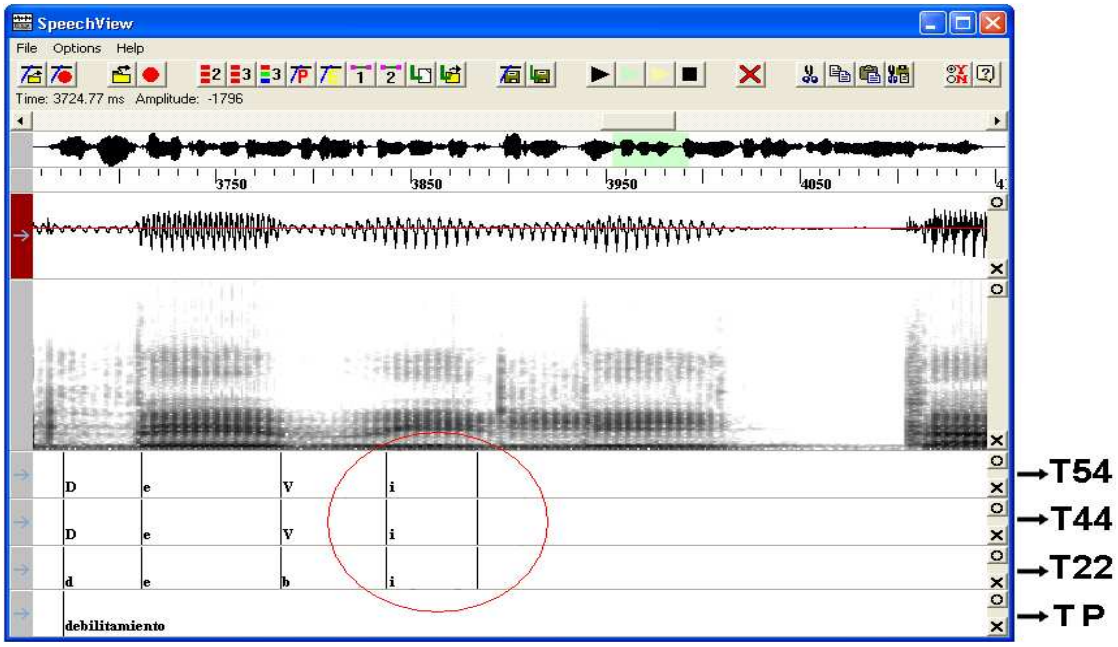

Figura 117. Imagen espectrográfica, segmentación y transcripción de [i].

1.4.2. Alófono paravocal palatal [j]

El alófono paravocal palatal [j] se distingue en el espectrograma por la transición de su primer formante, la cual será descendente o ascendente según su posición. Cuando [j] está en posición posterior a [a, e, o, u], la transición del formante desciende; por el contrario, cuando está anterior a las vocales, la transición es ascendente. En la Figura 118 se puede observar la imagen espectrográfica de [j]. En la imagen, el alófono anterior paravocal [j] está en posición anterior a alguna vocal, esto lo podemos notar en la transición del primer formante, ya que ésta es ascendente con respecto al alófono siguiente.

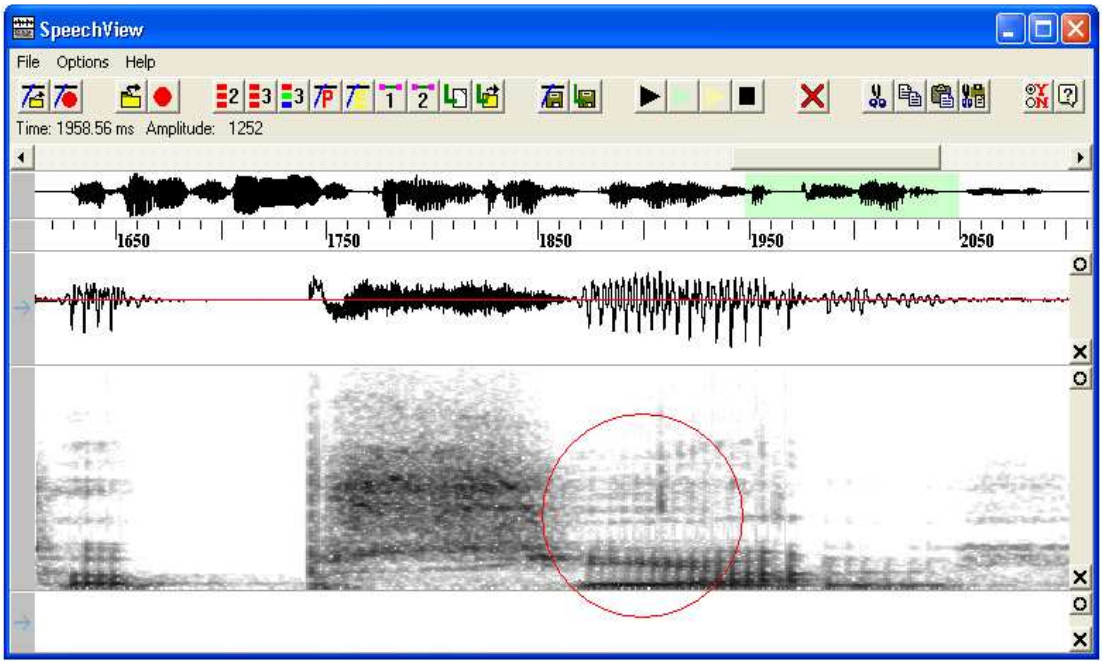

Figura 118. Imagen espectrográfica de [j].

La segmentación de este alófono se hace a partir de donde comienzan los formantes de [j], hasta donde empieza a descender o ascender el primer formante, según sea el caso. Si la transición no es muy notoria, lo recomendable es hacer la segunda segmentación a la mitad de los dos formantes. Lander señala que en este caso "After a semivowel or vowel, it can be practically impossible to determine the exact onset of a vowel. To be consistent, we have chosen to place the boundary in the middle of the transition period. If the formants never level o on either the semivowel or the vowel, divide the segment in half" (1997:58). Esta segmentación es la misma para los tres niveles de transcripción. En la Figura 119 se puede observar la segmentación de la imagen espectrográfica del alófono paravocal.

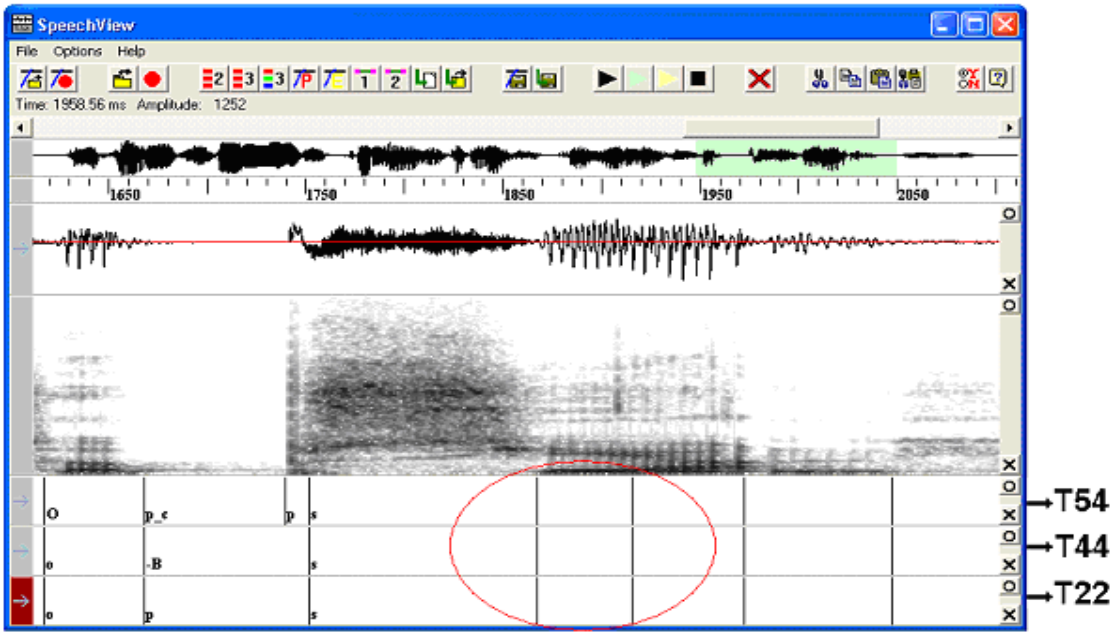

Figura 119. Imagen espectrográfica de segmentación de [j].

La transcripción de este alófono es [j] para los niveles T54 y T44; mientras que para el nivel T22 es [i]. En la Figura 120 se puede observar la imagen espectrográfica de la palabra *opciones*, en la cual se resalta la transcripción de este alófono en los tres niveles de transcripción.

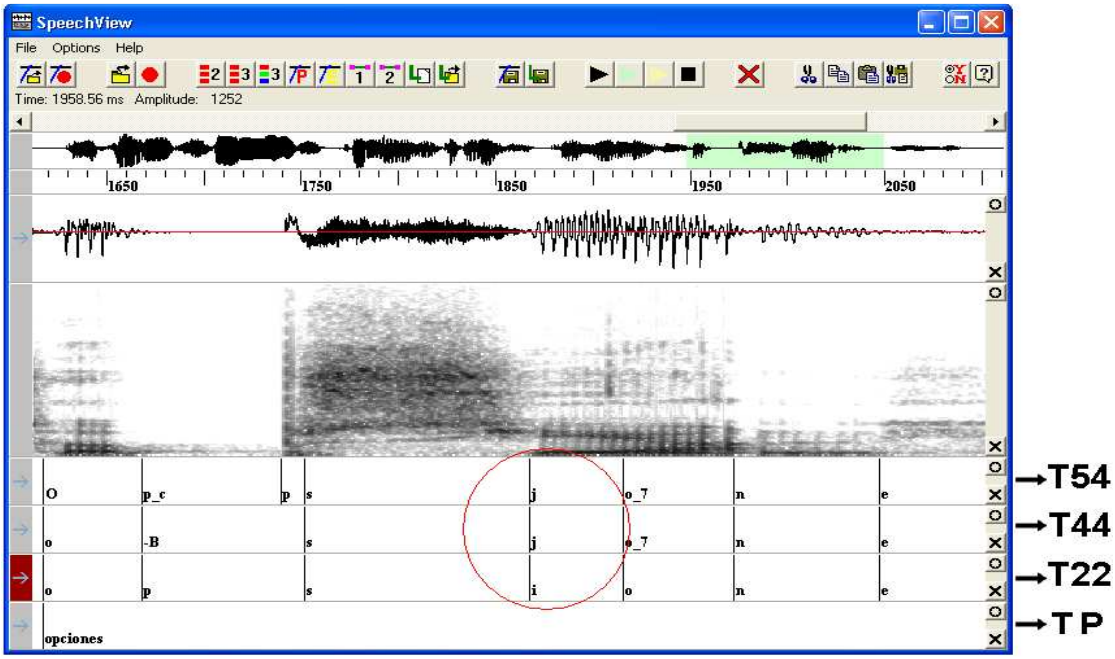

Figura 120. Imagen espectrográfica y transcripción de [j].

1.5. Fonema medio velar /o/

El fonema vocálico medio velar tiene dos alófonos: uno prototípico medio velar [o], que se realiza en cualquier posición de sílaba, y otro, medio velar abierto [O], que ocurre en contacto con [r] y en sílaba trabada por cualquier consonante. En la Figura 121 se muestran las reglas distribucionales de los alófonos de /o/, de acuerdo con Cuétara 2004.

Fonema	Alófono	Contexto	Grafía
Vocálico medio velar /o/	Medio velar [o]	En cualquier posición de sílaba	*o*
	Medio velar abierto [O]	En contexto anterior y posterior a [r] y en sílaba trabada por consonante	*o*

Figura 121. Reglas distribucionales de los alófonos del fonema /o/.

1.5.1. Alófono medio velar [o]

El alófono medio velar, [o], se identifica en el espectrograma por ser un segmento corto de alta energía, pero menor que la de [a] y la de [e]. Sus formantes tienen una frecuencia media, por lo que la distancia entre su F1 y F2 no es grande. En la Figura 122 se puede ver la imagen espectrográfica de [o] delimitada en un círculo.

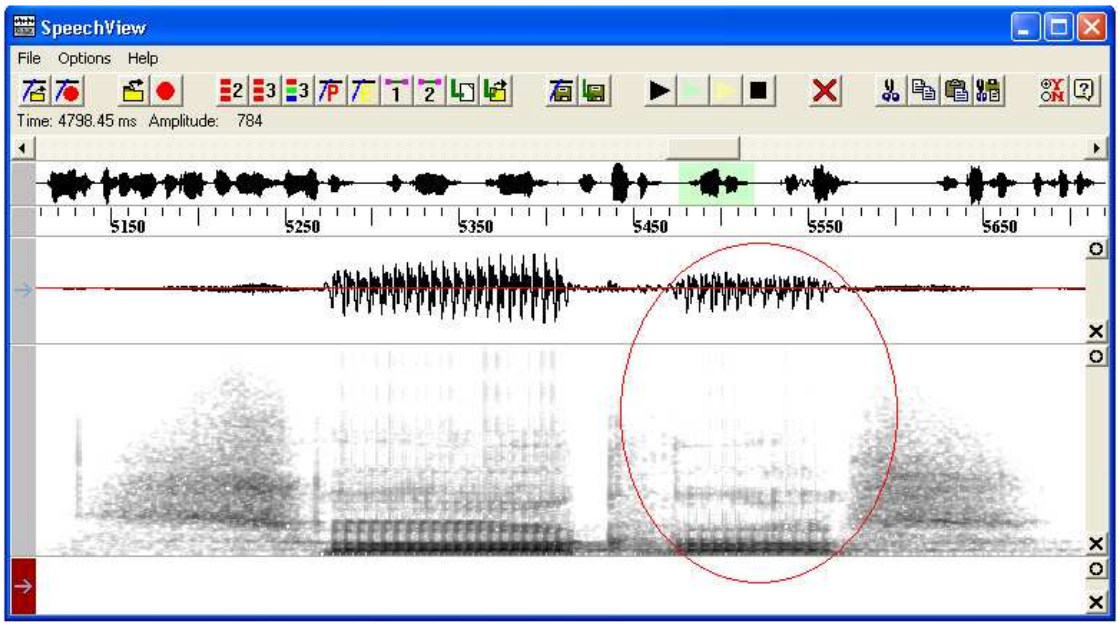

Figura 122. Imagen espectrográfica de [o].

La segmentación del alófono velar [o] se hace de la misma manera que en [a] (ver a [a]). La transcripción es [o], para las etiquetas de los tres niveles de transcripción. En la Figura 123 se puede ver la imagen espectrográfica de la palabra *zorros*, en la cual se resalta la segmentación y transcripción de [o].

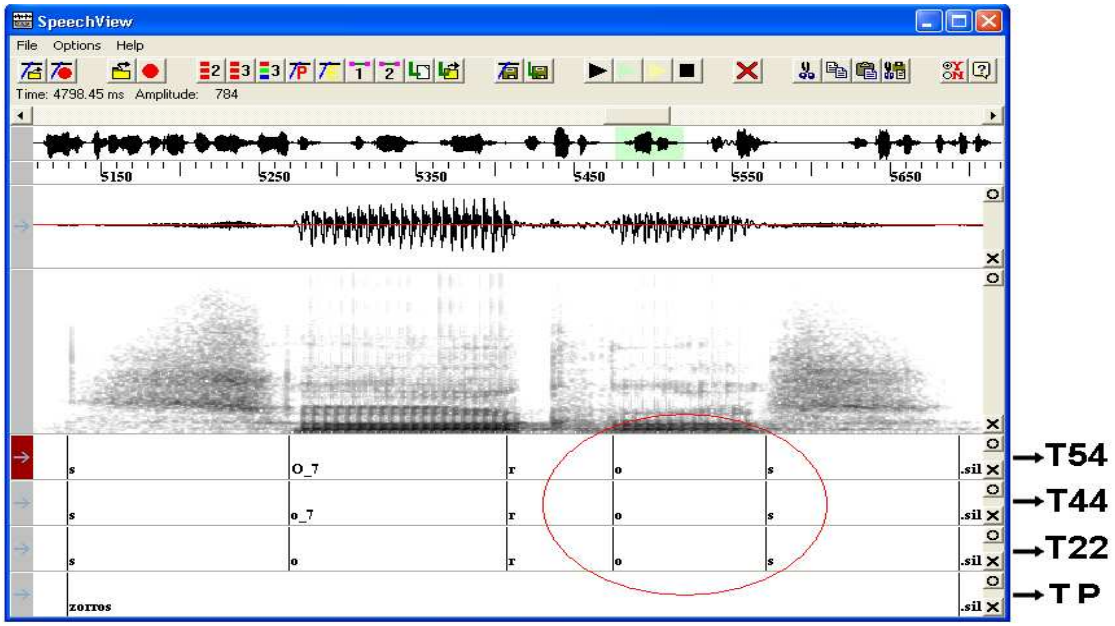

Figura 123. Imagen espectrográfica, segmentación y transcripción [o].

1.5.2. Alófono medio velar abierto [O]

El alófono medio velar abierto [O], se ve en el espectrograma como un segmento amplio de alta energía. Éste, al igual que [E], siempre se encontrará antecedido o precedido de las vibraciones del alófono alveolar vibrante múltiple, ya que es el contexto en el que se realiza. En la Figura 124 se puede ver la imagen espectrográfica de [O]. En ella se puede observar que la antecede el alófono vibrante.

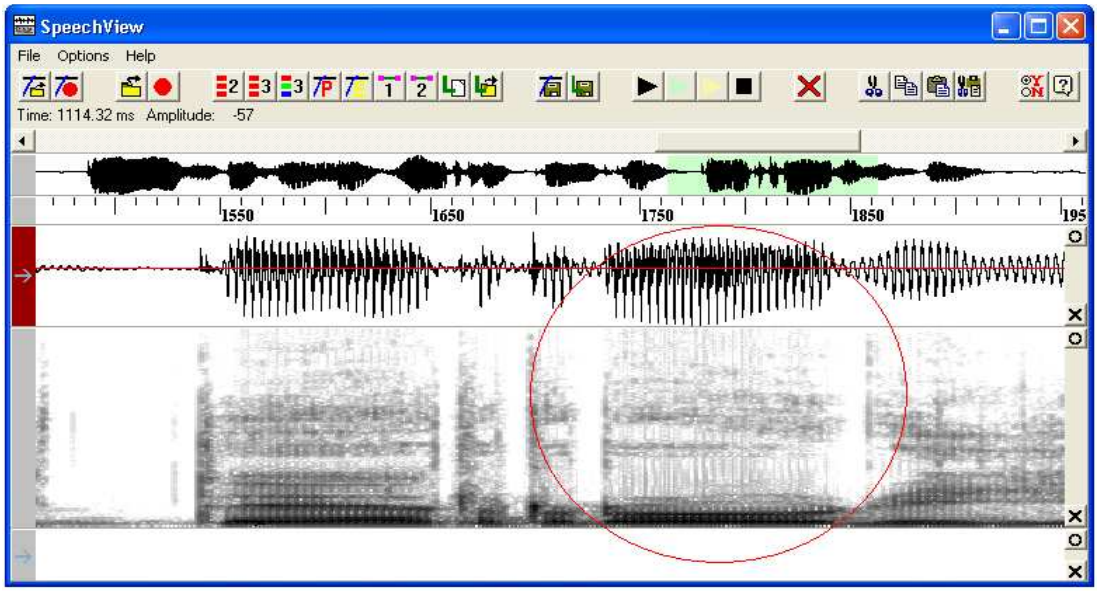

Figura 124. Imagen espectrográfica de [O].

Para la segmentación de [O] remitirse a [E]. La transcripción para este alófono es [O] para el nivel T54, y [o] para los niveles T44 y T22. En la Figura 125 se puede observar la imagen espectrográfica de la palabra *terrorismo*, en ella se ha señalado la segmentación y trascripción de [O], en los tres niveles.

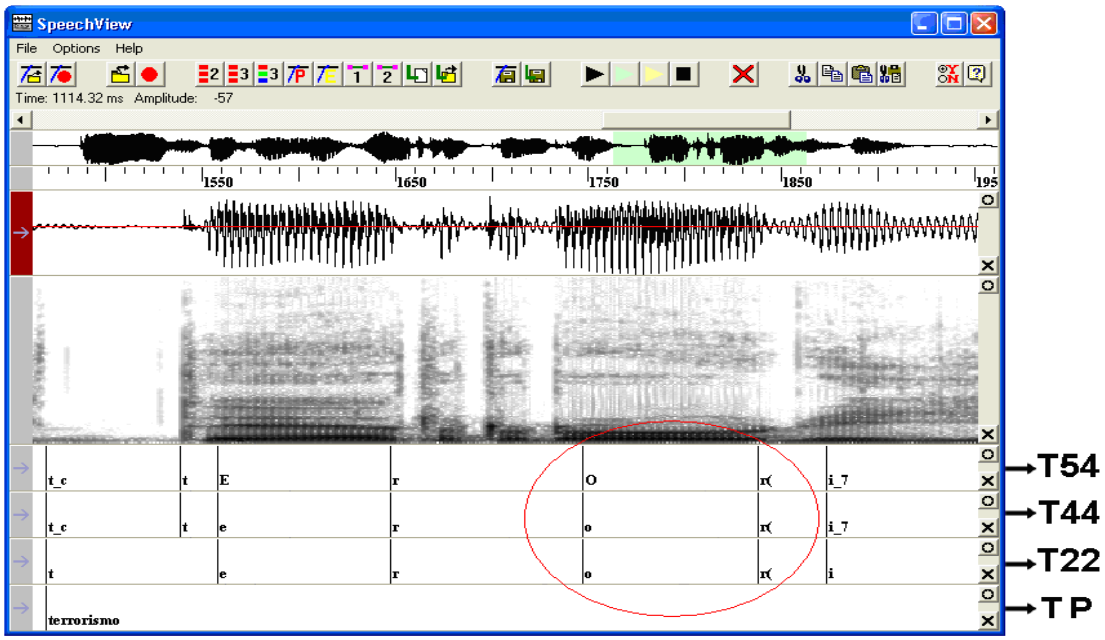

Figura 125. Imagen espectrográfica, segmentación y transcripción de [O].

1.6. Fonema cerrado velar /u/

El fonema vocálico posterior cerrado /u/ tiene dos alófonos: el prototípico, cerrado velar [u], que se realiza en cualquier posición de sílaba; y el paravocal velar [w], que ocurre cuando se encuentra en posición anterior o posterior a las vocales [a, e, i, o]. En la Figura 126 se muestran las reglas distribucionales de este fonema, de acuerdo con Cuétara 2004.

Fonema	Alófono	Contexto	Grafía
Vocálico cerrado velar /u/	Cerrado velar [u]	En cualquier posición de sílaba	*u*
	Paravocal velar [w]	En contexto anterior y posterior a [a, e, i, o]	*u*

Figura 126. Reglas distribucionales de los alófonos del fonema /u/.

1.6.1. Alófono cerrado velar [u]

El alófono cerrado velar [u] se caracteriza en el espectrograma por ser un segmento corto pero de alta energía, cuyos primero y segundo formantes aparecen en la parte inferior del espectrograma. En la Figura 127 se señala la imagen acústica que corresponde a este alófono.

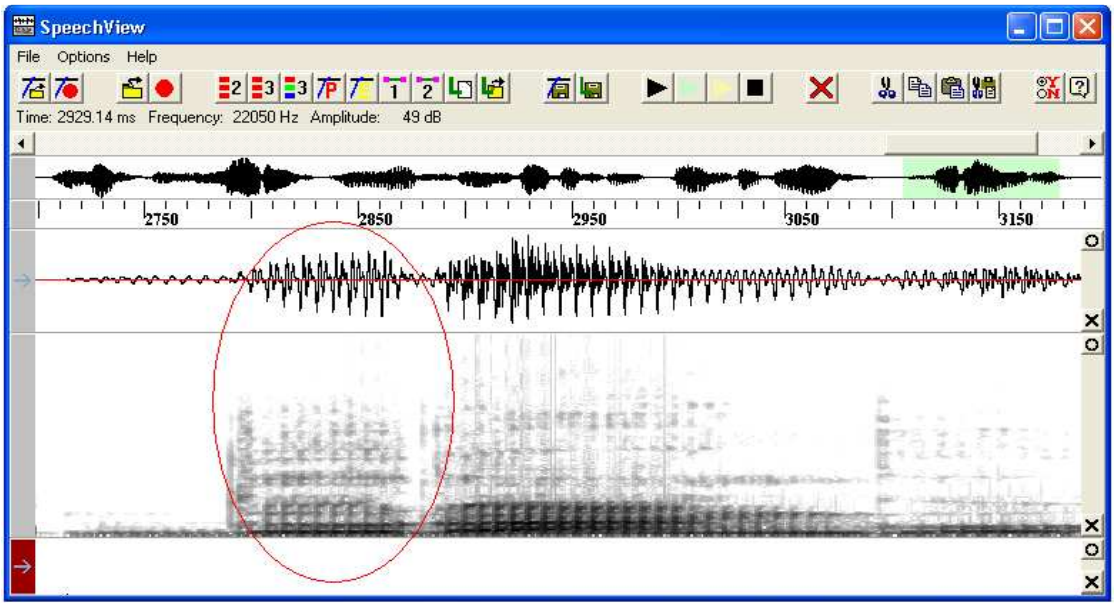

Figura 127. Imagen espectrográfica de [u].

Para la segmentación seguir el mismo proceso que en el alófono [a]. La transcripción es [u] para las etiquetas de los tres niveles de transcripción. En la Figura 128 se puede ver la imagen espectrográfica de la palabra *Durango*, en ella ha señalado la segmentación y la transcripción de [u].

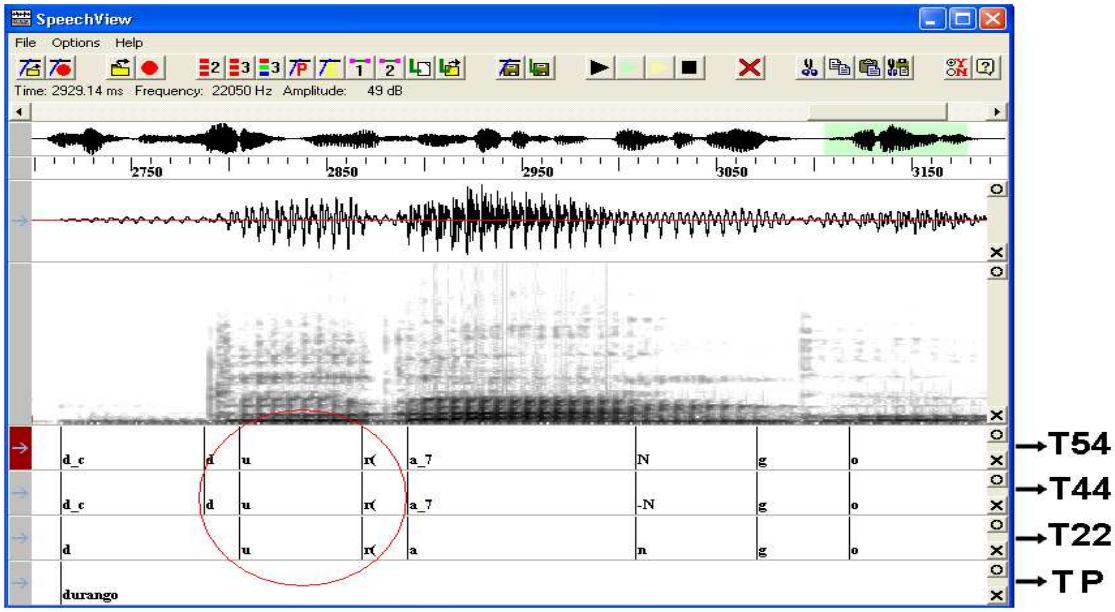

Figura 128. Imagen espectrográfica, segmentación y transcripción de [u].

1.6.2. Alófono paravocal velar [w]

El alófono paravocal velar [w] se caracteriza por ser un segmento en el cual el segundo y tercer formante tienden de disminuir de energía o desaparecer (Croot y Taylor 1995). También se distingue por presentar un primer formante, cuya transición asciende o desciende, según su posición: cuando está en anterior a [a, e, i, o] el formante asciende, cuando está posterior el formante desciende. En la Figura 129 se puede ver observar la imagen espectrográfica del alófono [w]. En esta imagen se puede observar la transición descendiente del primer formante; por lo tanto, se pude deducir que la paravocal está después de la vocal, pues ya que los formantes de este alófono siempre aparecerán a una baja frecuencia.

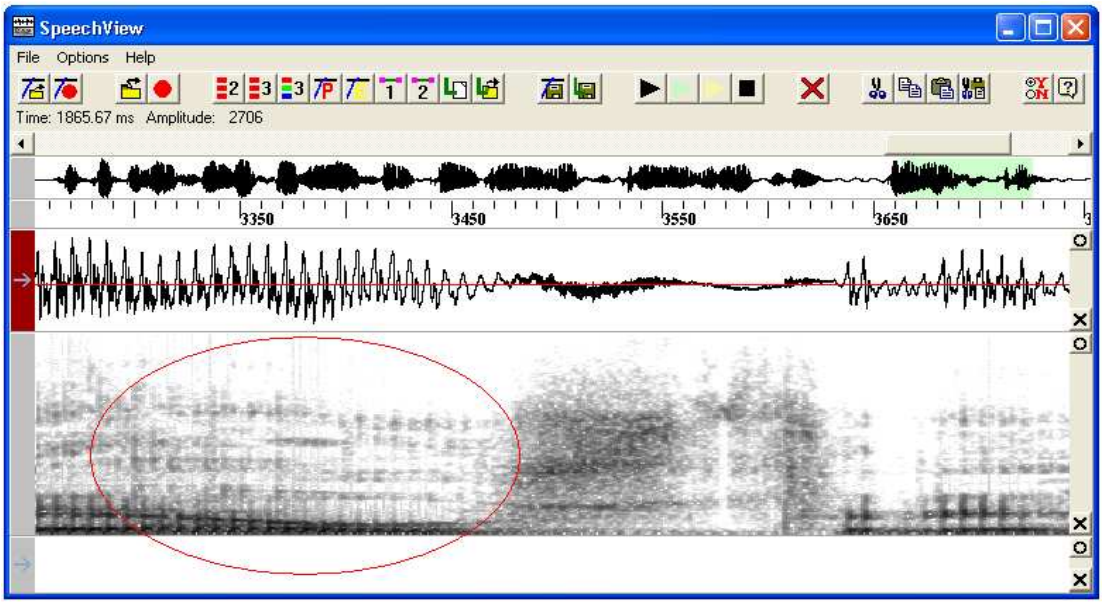

Figura 129. Imagen espectrográfica de [w].

La segmentación de este alófono se hace aplicando los mismos criterios de segmentación que en [j]. La transcripción de este alófono es [w] para la etiqueta del nivel T54, y [u] para las etiquetas de los niveles T44 y T22. En la Figura 130 se puede ver en la imagen acústica de la palabra *Austria*, en la cuál se ha señalado la transcripción de [w] en los tres niveles de transcripción.

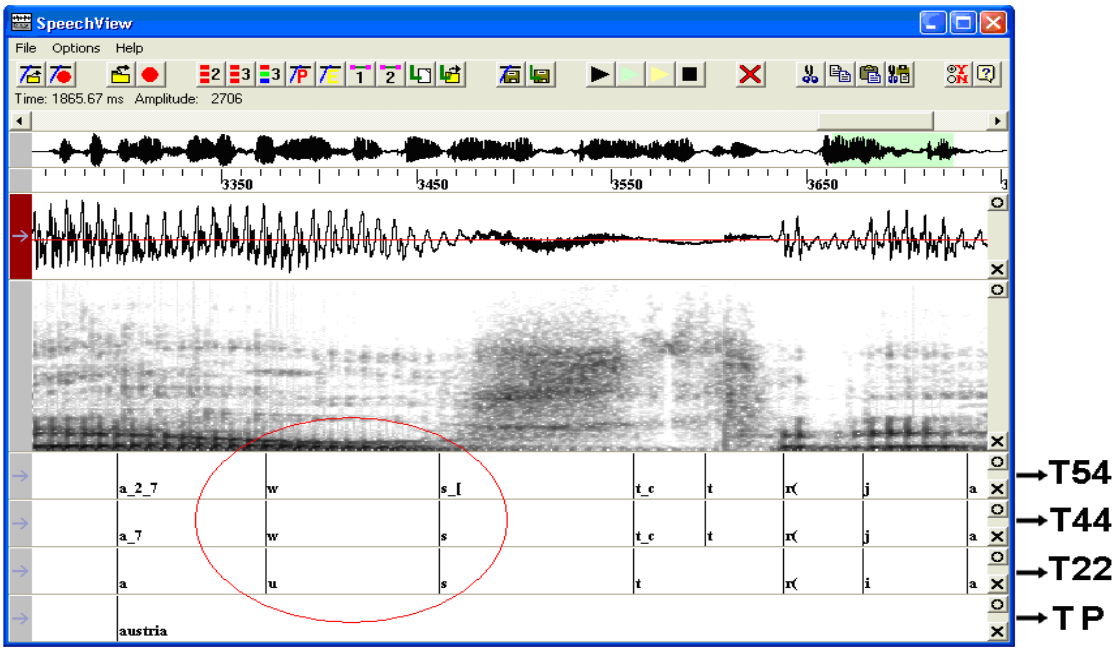

Figura 130. Imagen espectrográfica, segmentación y transcripción de [w].

III. FENÓMENOS FONÉTICOS

1. Fenómenos fonéticos

Durante el proceso del habla los sonidos no se emiten aisladamente, sino que se van hilando para dar como resultado una cadena hablada, la cual tiene como objetivo entablar una comunicación que pretende ser eficiente a un bajo costo; por tal razón el hablante durante el acto de comunicación tiende a variar o modificar la pronunciación en las palabras, dichas modificaciones son conocidas como *fenómenos fonéticos*.

Dentro de la clasificación de fenómenos fonéticos existen varios tipos como la asimilación, la relajación de sonidos, etc. En este *Manual* únicamente se documentan dos, que son: la homologación de sonidos idénticos y la elisión,[3] dentro del segundo se reúnen tres tipos: aféresis, síncopa y apócope. A continuación se explica en lo que consiste cada uno de los fenómenos ejemplificándolos con imágenes espectrográficas.

1.1. Homologación de sonidos idénticos

La homologación de sonidos idénticos se refiere a la unión dos sonidos idénticos, vocálicos o consonánticos, que se caracterizan por fundirse en una misma emisión. Para que eso suceda, los sonidos deben ser el último de una palabra y el inicial de la otra; así, cuando el hablante pronuncia ambas palabras, las une por medio de los sonidos en común, haciendo una reducción de ellos. La reducción de los sonidos que se unen en palabra suele ocurrir por la velocidad con que son emitidos. Por ejemplo, en la oración *Quiero de estos mismos*, el hablante pronunciaría *Quiero destos mismos*. En la Figura 131 podemos observar la homologación del fonema /e/ en la frase *que es una*. Como se puede ver en la imagen espectrográfica, únicamente se pronunció un fonema /e/ para ambas palabras. Esto se puede

[3] En el *Proyecto DIME* se han realizado dos tesis de licenciatura sobre estos fenómenos: *Fenómeno de pérdida en Corpus DIME para su inclusión en un reconocedor de habla* (Ceballos 2007) y *Sistematización del fenómeno de silabicación en el Corpus DIME* (Espinoza 2007).

comprobar porque el segmento del alófono es pequeño, a diferencia de cuando hay una pronunciación de los dos alófonos, el segmento es largo.

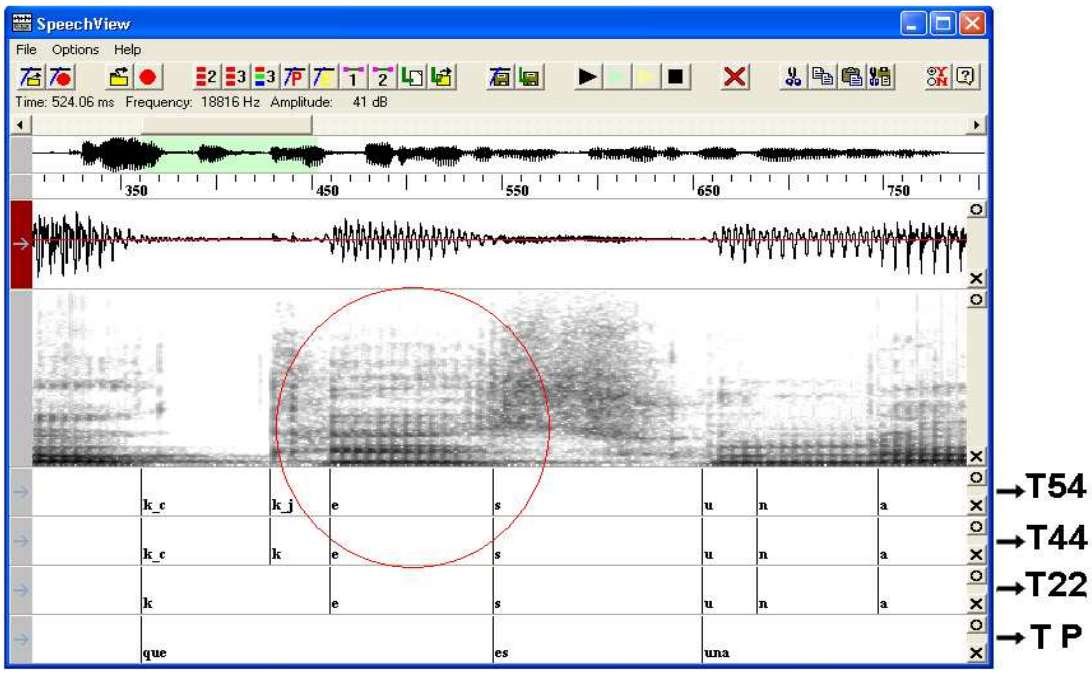

Figura 131. Imagen espectrográfica de l fenómeno de homologación.

1.2. Elisión

La elisión consiste en la supresión de algún sonido en una palabra, o bien, dentro de la cadena hablada. La elisión suele ocurrir por relajamiento en la pronunciación de algunos. Existen tres tipos de elisión: aféresis, síncopa y apócope. A continuación se explica cada uno de ellos y su respectivo ejemplo en el espectrograma.

- **Aféresis.** En la aféresis se suprime el sonido inicial de una palabra. En la Figura 132 aparece la imagen espectrográfica, segmentada y transcrita, de la frase *esto es todo*, en la cual se puede observar la pérdida del fonema /e/, en la palabra *es*.

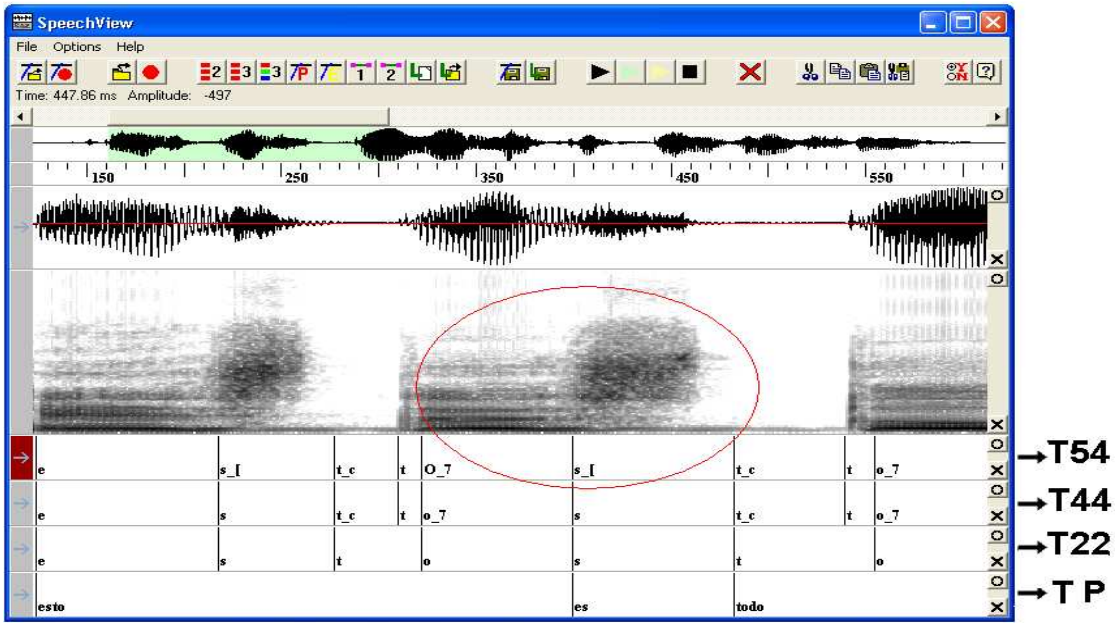

Figura 132. Imagen espectrográfica de l fenómeno de aféresis.

- **Síncopa.** La síncopa consiste en la elisión de algún fonema que se encuentra al interior de la palabra. Lass menciona que este término suele utilizarse con más frecuencia en la pérdida de vocales, pero también se ha usado para sonidos consonánticos (1984:187). En la Figura 133 se puede observar la imagen espectrográfica de la palabra *diputados*, en la cual se ha elidido el segundo alófono /d/.

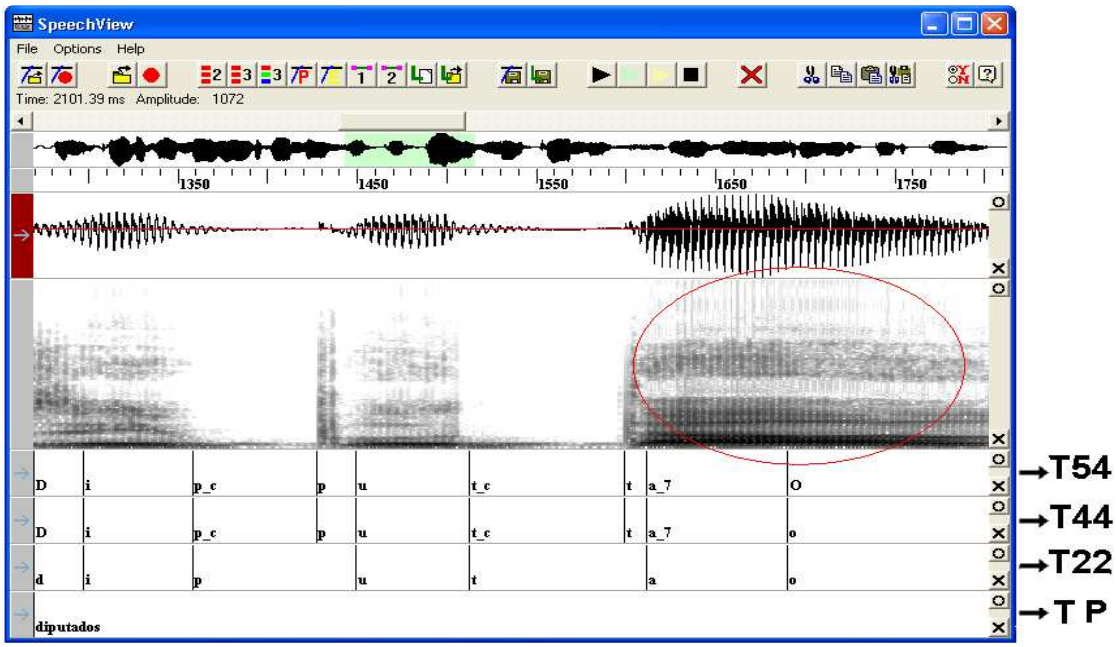

Figura 133. Imagen espectrográfica del fenómeno de síncopa.

- **Apócope.** En la apócope se pierde el último sonido de una la palabra o de la cadena hablada. En la Figura 134 se puede observar la imagen espectrográfica de la palabra *universidad*, en la cual se ha perdido el último alófono /d/.

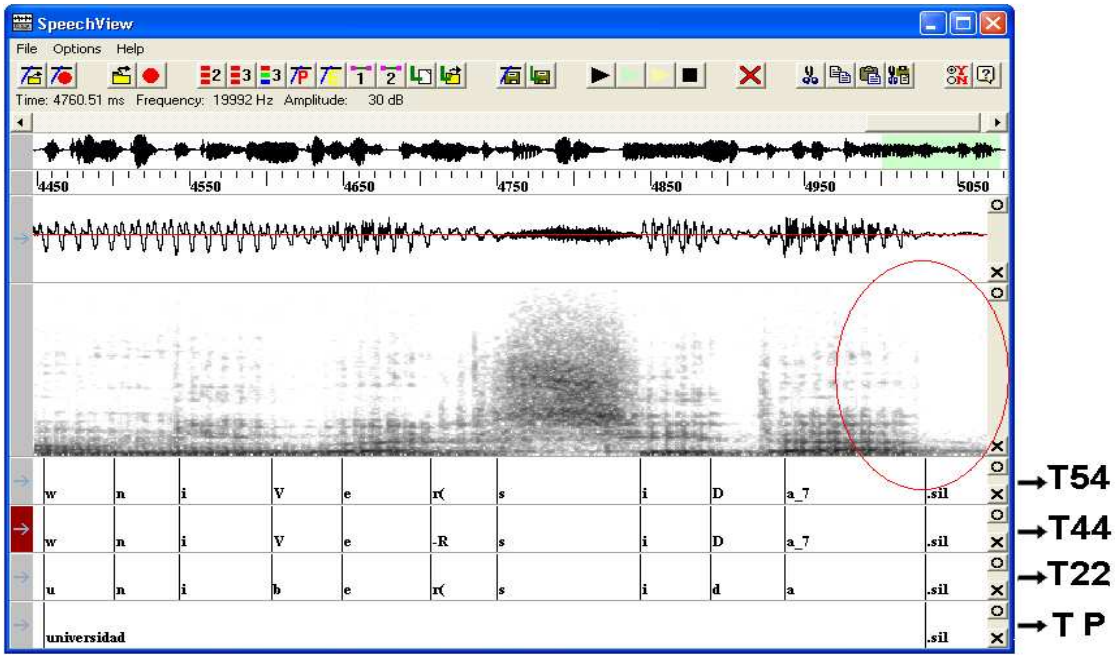

Figura 134. Imagen espectrográfica del fenómeno de apócope.

IV. Apéndice. TABLA DE EQUIVALENCIAS ENTRE LOS ALFABETOS *AFI*, *RFE* Y *MEXBET* (Cuétara 2004:144-145)

Alófonos del español de la ciudad de México			
Alófono	**AFI**	**RFE**	**Mexbet**
Labial oclusivo sordo	p	p	p
Dental oclusivo sordo	t	t	t
Velar oclusivo sordo	k	k	k
Bilabial oclusivo sonoro	b	b	b
Dental oclusivo sonoro	d	d	d
Velar oclusivo sonoro	g	g	g
Palatal africado sordo	t͡ʃ	ĉ	tS
Palatal africado sonoro	d͡ʒ	ŷ	dZ
Labiodental fricativo sordo	f	f	f
Dental fricativos sordo	s̪	ș	s_[
Alveolar fricativo sordo	s	s	s
Velar fricativo sordo	x	x	x
Bilabial fricativo sonoro	β̞	ⱱ̟	V
Dental fricativo sonoro	ð̞	ɾ̪	D
Alveolar fricativo sonoro	s̬	z	z
Palatal fricativo sonoro	ʒ	y	Z
Velar fricativo sonoro	ɣ̞	ɾ̞	G
Bilabial nasal	m	m	m
Dental nasal	n̪	ȵ	n_[

Alveolar nasal	n	n	n
Palatal nasal	ɲ	n̪	n~
Velar nasal	nˠ	ŋ	N
Alveolar vibrante simple	ɾ	r	r(
Alveolar vibrante múltiple	r	r̄	r
Alveolar lateral	l	l	l
Vocal central abierta	a	a	a
Vocal abierta palatal	a+	a̪	a_j
Vocal abierta velar	a̲	a.	a_2
Vocal media palatal	e	e	e
Vocal media palatal abierta	e̬	e̞	E
Vocal cerrada palatal	i	i	i
Paravocal palatal	j	j / i̯	j
Vocal media velar	o	o	o
Vocal media velar abierta	o̬	o̞	O
Vocal cerrada velar	u	u	u
Paravocal velar	u	w / u̯	w


"Pies para qué los quiero si tengo alas pa' volar" F.H.