UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

PROGRAMA DE MAESTRÍA Y DOCTORADO EN INGENIERÍA

FACULTAD DE INGENIERÍA

RECONOCEDOR DE COMANDOS DE VOZ DEL NÁHUATL DE UNA COMUNIDAD DE LA SIERRA NORTE DEL ESTADO DE PUEBLA

T E S I S

QUE PARA OPTAR POR EL GRADO DE:

MAESTRO EN INGENIERÍA

ELÉCTRICA – PROCESAMIENTO DIGITAL DE SEÑALES

P R E S E N T A :

ADOLFO HERNÁNDEZ HUERTA

TUTOR:
DR. JOSÉ ABEL HERRERA CAMACHO

2007

JURADO ASIGNADO:

Presidente:	Dr. Gerardo Eugenio Sierra Martínez.
Secretario:	Dr. Felipe Orduña Bustamante.
Vocal:	Dr. José Abel Herrera Camacho.
1er. Suplente:	Dr. Alfonso Medina Urrea.
2do. Suplente:	M. en Ling. Leopoldo José Manuel Valiñas Coalla.

Lugar donde se realizó la tesis:

Laboratorio de Procesamiento de Voz, Facultad de Ingeniería, UNAM.

TUTOR DE TESIS:

Dr. José Abel Herrera Camacho

FIRMA

A mis padres,

Adolfo Hernández Reyes y
María del Carmen Huerta Espinosa,

por manifestarme una vez más su amor
al apoyarme incondicionalmente
en toda mi formación profesional.

Agradecimientos

A la Facultad de Ingeniería y a la UNAM, por la formación profesional que me han brindado a través de esta maestría. También agradezco a mis compañeros, y ahora amigos, de la maestría en ingeniería por las horas de estudio conjunto y su apoyo en mi estancia de estudio en la ciudad de México.

A mi tutor, el Dr. Abel Herrera, no sólo por su asesoría, también por su constante apoyo y valiosos consejos durante su acompañamiento a lo largo de la maestría para ser un mejor profesional en la investigación.

A mis profesores, por la transmisión de sus conocimientos y experiencia, en especial a la Dra. Lucía Medina por su apoyo en mi formación profesional.

Al M.L. Polo Valiñas, del Instituto de Investigaciones Antropológicas (IIA) de la UNAM, por sus invaluables recomendaciones y hacerme descubrir la maravilla de la ciencia del lenguaje.

Al Lic. Samuel Herrera y al Dr. Mario Castillo, del IIA, por sus muy oportunas orientaciones.

Al grupo de personas que me permitieron grabarlas:

Esteban Arrieta	Francisco Ortigoza	Miguel Osorio
Yolanda Argueta	Linda Ortigoza	Vicente Flores
Ocotlán Osorio	Clara Osorio	Neri Hernández
Micaela Osorio	Eduardo Osorio	Alfredo Álvarez
Ernesto Vázquez	José Heradio Vázquez	

Gracias por su apoyo y paciencia en las sesiones, especialmente a Esteban Arrieta y a Miguel Osorio.

Agradezco a mi familia, amigos y seres queridos por su apoyo manifestado de tan diversas maneras y en todo momento.

Mi profundo agradecimiento a la comunidad jesuita de Pedregal Santo Domingo por recibirme con los brazos abiertos.

Al Consejo Nacional de Ciencia y Tecnología por el apoyo brindado durante los estudios de maestría a través de una beca de 24 meses.

ÍNDICE

INTRODUCCIÓN

El procesamiento digital de señales es una de las tecnologías más poderosas en la actualidad, transforma nuestro mundo día tras día y constantemente muestra su vigor y evolución continua. Uno de los campos de investigación más interesantes y prometedores, aunque estudiados desde hace décadas, es el procesamiento digital de voz o procesamiento digital del habla. En este trabajo emplearemos indistintamente los términos *voz* y *habla*, sin embargo resulta oportuno comentar que el término *voz* hace referencia al fenómeno físico y proceso fisiológico, mientras que *habla* tiene una connotación lingüística.

La tecnología de voz es un tema multidisciplinario e interdisciplinario, requiere el conocimiento de campos tan diversos tales como: procesamiento de señales, electrónica, ciencias computacionales, lingüística y fisiología. Hay tres áreas principales en la tecnología de la voz: síntesis, reconocimiento y codificación de la voz.

El procesamiento digital de voz permite salvar las diferencias del lenguaje y permite potencialmente a cada individuo participar en la revolución de la información. Desafortunadamente, y como ocurre en tantas otras tecnologías, la construcción de sistemas de procesamiento de voz requiere de recursos significativos. Con aproximadamente 4600 lenguas en el mundo, el procesamiento de voz es prohibitivo para muchas de ellas excepto para las lenguas económicamente viables.

Los problemas de comunicación entre personas de diferentes lenguas siempre ha sido un reto. Tan sólo en la República Mexicana existen aproximadamente seis millones de hablantes en lengua indígena; de ellas, la lengua dominante es el náhuatl, más de un millón y medio de personas en veintiún estados del país lo hablan en cualquiera de sus diferentes variantes de acuerdo a cada región geográfica. En su XII censo nacional de población realizado en el año 2000, el INEGI advirtió un alto porcentaje de monolingüismo de lengua indígena (16.9%) en la población hablante de lengua indígena. Para este sector de la población el acceso efectivo de servicios públicos y privados -tales como educación, salud, jurídicos, agrarios- aún encuentra dificultades. Las consecuencias del monolingüismo indígena incluso llegan hasta los Estados Unidos de Norteamérica, donde se han registrado detenciones de indígenas mexicanos monolingües indocumentados sometidos a juicios injustos por no contar con traductores que los apoyaran empezando porque las autoridades desconocían de qué lengua se trataba.

La creación de sistemas que procesen la producción oral de diversas lenguas de las culturas indígenas de México aportará grandes beneficios a sus comunidades y a quienes nos interesamos en ellas. En este sentido muchas ideas se pueden desarrollar; ejemplo de ello es la existencia de un sistema identificador de lenguas indígenas de México [Cdi]; sin embargo es importante estimular el desarrollo de otras tecnologías de voz para estas lenguas, como por ejemplo sistemas de reconocimiento. Por otro lado, el estudio acústico de los sonidos lingüísticos, o estudios fonético-acústicos, de las lenguas indígenas brindan la oportunidad de conocer a

mayor profundidad las propiedades intrínsecas de dichos sonidos y que los caracterizan como tales, así como la interacción entre ellos. Estos estudios tienen gran importancia no sólo para la ciencia de la lingüística, sino también para las tecnologías de voz, donde la identificación y el conocimiento de las características acústicas de los sonidos de una lengua determinada permiten el desarrollo de sistemas de procesamiento de voz más eficientes, por ejemplo, en reconocimiento y síntesis de voz.

El presente trabajo es una sincera y entusiasta exploración en este amplio horizonte del procesamiento de señales de voz y la fonética acústica aplicados a una de las lenguas indígenas más importantes y estudiadas de México y que no había sido abordada de esta manera.

El trabajo se divide en dos partes: la primera tiene por objetivo la descripción fonética de la lengua náhuatl por medio del análisis acústico de uno de sus dialectos. Se consideró el náhuatl de San Miguel Tzinacapan, poblado perteneciente al municipio de Cuetzalan del Progreso, ubicado en la Sierra Norte del Estado de Puebla. A través del estudio fonético-acústico se analizan sus sonidos lingüísticos, comenzando desde las características fundamentales de cada uno de ellos hasta la interacción entre sí. Los efectos coarticulatorios y la posición silábica de los sonidos repercuten de manera seria en sus rasgos. Se realizan mediciones de las propiedades temporales y espectrales de estos sonidos y se explica el desafío de esta tarea. Hasta donde he averiguado es el primer estudio de su tipo del náhuatl.

La segunda parte consiste en el desarrollo de un sistema de reconocimiento de sesenta y un comandos de voz de dicho dialecto. El objetivo es desarrollar un reconocedor por cuantización vectorial (VQ) para analizar la eficiencia de esta técnica de voz, que es bien conocida y una de las más empleadas, ante un idioma que no se había procesado con ella. Se utilizaron coeficientes LPC ya que es otra de las técnicas más poderosas en el análisis de la señal de voz pues brinda una representación precisa y económica de parámetros relevantes de esta señal y reduce los cálculos en el reconocimiento de voz. Esta tarea ha sido exitosa por el alto porcentaje de reconocimiento logrado, los principales problemas existentes en el reconocimiento son identificados.

Este trabajo de investigación tiene un claro sentido didáctico y a la vez exploratorio para sentar bases no sólo a mi formación, sino también inicia una prospectiva de desarrollo de sistemas sofisticados de procesamiento digitales de voz aplicado a lenguas indígenas en el Laboratorio de Procesamiento de Voz de la Facultad de Ingeniería de la UNAM.

Por último, la estructura del presente trabajo es la siguiente:

El capítulo 1 describe los aparatos fonador y auditivo, los cuales son esenciales para la producción y percepción de la voz.

El capítulo 2 presenta conceptos básicos de fonética y fonología, los cuales se emplean en este trabajo.

El capítulo 3 brinda fundamentos del procesamiento digital de voz así como presenta sus métodos más conocidos; también se presenta el papel de algunos de estos métodos en los estudios fonéticos acústicos, proporcionando muchos ejemplos de sus usos.

El capítulo 4 aborda el tema del reconocimiento de voz; introduce un panorama de esta tecnología, sus características más importantes, las principales líneas de investigación. Tras dicho panorama se desarrollan las técnicas del reconocimiento de voz por cuantización vectorial.

El capítulo 5 presenta la lengua náhuatl, su diversidad dialectal, su distribución en el país; de manera más importante, se describe la metodología seguida para el estudio fonético-acústico y se presenta el sistema fonológico del náhuatl hablado en San Miguel Tzinacapan. Por último se describe la silabación del náhuatl y se presentan una serie de consideraciones finales para este estudio.

El capítulo 6 inicia el estudio fonético-acústico, se abordan las propiedades estáticas de los sonidos del náhuatl cuando están ubicados a principio y final de sílaba.

Los capítulos 7 a 10 conforman el estudio fonético-acústico desde un enfoque dinámico, es decir, analizando las interacciones entre los sonidos, primordialmente a través de las transiciones de un sonido a otro.

El capítulo 11 cierra el estudio fonético-acústico presentando variaciones acústicas de los hablantes. Mientras que en los capítulos anteriores se presentan segmentos de sonidos bien formados y que fueran buenos representantes de las realizaciones, o pronunciaciones, de los hablantes grabados, este capítulo muestra cómo los hablantes no siempre pronunciaban con los mismos rasgos acústicos de los demás.

El capítulo 12 muestra el sistema de reconocimiento desarrollado en este trabajo, describiendo las diferentes etapas que lo conforman. Al final de este capítulo se presentan los resultados.

Adicionalmente, se presentan las secciones correspondientes a las conclusiones, apéndices y referencias. Cabe mencionar que en los apéndices se presentan el cuestionario de las palabras que se grabaron, con su traducción al español, así como una relación de las figuras del estudio fonético-acústico con los archivos de audio de las pronunciaciones grabadas, esta relación permite identificar al hablante con su figura correspondiente.

1. PRODUCCIÓN Y PERCEPCIÓN DE LA VOZ

En este capítulo se presentarán los órganos que intervienen en la producción del habla, así como del sistema auditivo. Las secciones 1.1 (aparato fonador) y 1.4 (aparato auditivo) se extraen de Antonio Quilis en [Qui99] para resumirlas y presentarlas en este trabajo, varias figuras no pertenecen a dicho autor y son respectivamente referenciadas.

1.1 Aparato fonador

En la producción del sonido interviene un conjunto de órganos que se conoce con el nombre de aparato fonador. Estos órganos tienen, además, otras funciones estrictamente fisiológicas: la respiración, la deglución, etc. Lo importante es que la corriente de aire, que proviene de los pulmones, va a sufrir una serie de transformaciones a su paso por el aparato fonador y se va a convertir en sonidos propios para la comunicación humana. En la figura 1.1 se muestra un esquema del aparato fonador.

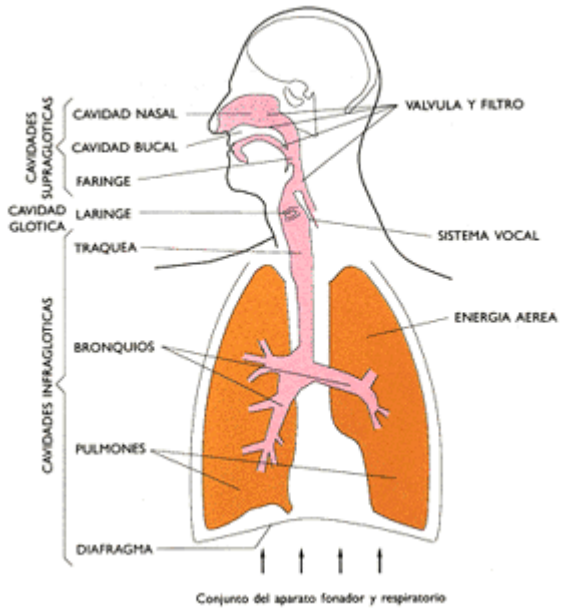


Figura 1.1: Aparato fonador [Deu]

El conjunto de órganos que intervienen en la fonación se puede clasificar en tres grupos: 1. Órgano respiratorio ó cavidades infraglóticas; 2. Órgano fonador ó cavidad laríngea; 3. Cavidades supraglóticas.

Cavidades Infraglóticas

Están formadas por los órganos propios de la respiración: pulmones, bronquios, tráquea. Los pulmones son los que desempeñan el papel más relevante. Su misión es doble: por un lado, fisiológica, en cuanto que son instrumento de la respiración, con toda la serie de transformaciones bioquímicas que en ellos se originan; por otro, el de servir de proveedores de la cantidad de aire suficiente para que el acto de la fonación se realice.

Los pulmones realizan constantemente dos movimientos: el de inspiración, absorbiendo aire, y el de espiración, expulsándolo. Durante este segundo movimiento se puede producir el sonido articulado.

El aire contenido en los pulmones va a parar a los bronquios, y de aquí a la tráquea, órgano formado por anillos cartilaginosos superpuestos, que desemboca en la laringe.

Cavidad laríngea u órgano fonador

La laringe, como puede verse en las figuras 1.1 y 1.2, está situada inmediatamente por encima de la tráquea. Está compuesta por cuatro cartílagos: el cricoides, que tiene forma de anillo, constituye la base; el tiroides (llamado también nuez o bocado de Adán), en forma de escudo, va unido al cricoides por medio de los cuernecillos; está abierto por su parte alta y posterior; los dos aritenoides, especie de pirámides pequeñas situadas sobre el engaste del cricoides; se mueven sobre él merced a un sistema de músculos. Desde los aritenoides parten los músculos que abren y cierran la glotis.

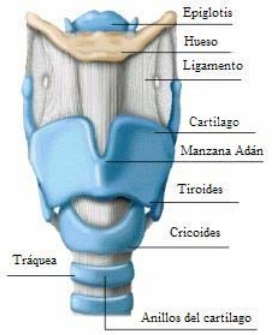


Figura 1.2: Laringe [Deu]

En el interior de la laringe están situadas las cuerdas vocales, que son como dos tendones o dos pliegues, vea la figura 1.3. Están situadas horizontalmente en dirección anteroposterior. Por su parte anterior están unidas al interior del cartílago tiroides, y por la posterior a los aritenoides. El paso que queda entre las cuerdas vocales cuando están abiertas recibe el nombre de glotis.

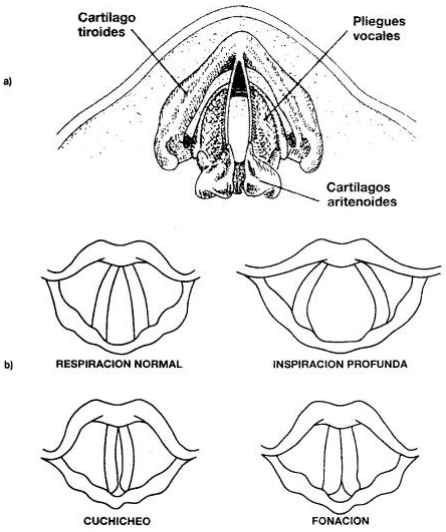


Figura 1.3: Cuerdas vocales en actitud de reposo [Mag01]

El movimiento de los aritenoides, realizado, como ya hemos dicho, por un sistema de músculos, puede variar la forma de la glotis haciendo que adopte diversas posiciones.

La acción de cuerdas vocales produce la primera gran clasificación de los sonidos articulados: si las cuerdas vocales vibran, los sonidos son sonoros, como las vocales y algunas consonantes: [b, d, g, m, n, l, r], etc. Si no vibran, los sonidos son sordos, como [s, f, x, Θ], etc.

¿Cómo se produce la vibración de las cuerdas vocales? cuando se va a iniciar la fonación, la glotis se cierra. Se produce entonces una presión del aire infraglótico contra los lados de la tráquea y contra la glotis, cuyos bordes se separan dejando salir una cantidad determinada de aire, que pasa entre las cuerdas, las cuales, por su elasticidad interior, se aproximan nuevamente, pero por su parte inferior, llegando a cerrar la glotis. Esta oclusión se desplaza hacia lo alto; el mismo movimiento se repite una y otra vez: pequeñas masas de aire, una detrás de otra, pasan a través de la glotis, desplazando su punto de colusión de abajo hacia arriba a medida que la presión del aire infraglótico tiende a separar las cuerdas vocales, que se cierran nuevamente después del paso de cada pequeña masa de aire. Estas interrupciones en la salida de la corriente de aire, debidas al cierre y abertura repetidos de la glotis y a la tensión de las cuerdas vocales, originan vibraciones de aire de la misma frecuencia fundamental que los cierres y aberturas de la glotis; por lo consiguiente, la frecuencia vibratoria de las cuerdas vocales y la frecuencia del fundamental de la onda sonora que se origina son iguales [Qui99].

Las cuerdas poseen un sistema de finísimas fibras musculares que pueden modificar su grosor y su grado de tensión: unas cuerdas vocales gruesas originan una frecuencia fundamental baja, y viceversa. Esta diferencia de grosor puede ser intencionada o constitutiva del individuo: edad y sexo; uno de los promedios es el siguiente: la frecuencia media del fundamental en los niños es de 264 cps (ciclos por segundo); en las mujeres, es de 223 cps y en los hombres de 123 cps. La voz de la mujer y del niño tiene un fundamental más alto que el del hombre, debido a que

sus cuerdas vocales son más delgadas y cortas. Si la tensión es débil, la frecuencia del fundamental será débil, y viceversa [Qui99].

La presión infraglótica también es importante: su aumento incide en la elevación de la frecuencia del fundamental. Así mismo, actúa sobre la amplitud de las vibraciones: cuanto mayor es la presión infraglótica, mayor es la amplitud de las vibraciones y el sonido es más fuerte.

Cavidades supraglóticas

Al pasar la corriente de aire (vibrando o no, según haya sido la situación de las cuerdas vocales) por la zona laríngea, entra en la cavidad de la faringe laríngea (o laringofaringe) y luego en la faringe oral, donde se va a producir otra gran división de los sonidos, según la acción del velo del paladar:

Si el velo del paladar está adherido a la pared faríngea, se producen los sonidos articulados orales, como [p, b, s, k], etc.

Si el velo del paladar desciende de la pared faríngea se articulan los sonidos consonánticos nasales, como [m, n].

Si están abiertas simultáneamente la cavidad bucal y la cavidad nasal, se originan los sonidos vocálicos nasales, o mejor los sonidos oronasales, como [ã], [ẽ], etc.

Cuando el sonido es oral, la única gran cavidad existente es la bucal. Al poder cambiar fácilmente su volumen y su forma gracias a la movilidad de la lengua, de los labios y del maxilar inferior, se pueden originar diferentes cavidades de resonancia que son las que producen, al actuar como filtros, los distintos sonidos articulados.

La parte superior de la cavidad bucal está constituida por el paladar, dividido en dos zonas: la anterior, ósea, conocida con el nombre de paladar duro, y la posterior, con el nombre de paladar blando o el velo del paladar. En la parte inferior de la boca está la lengua, órgano activo por excelencia (vea la figura 1.4).

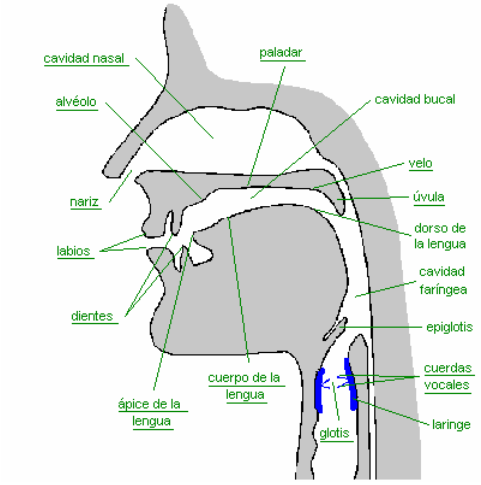


Figura 1.4: Corte vertical de los órganos fonadores y sus cavidades [Deu]

Con el objeto de fijar el lugar de articulación de los sonidos, es necesario señalar unas zonas en cada uno de los órganos articulatorios anteriormente mencionados.

El paladar duro comienza inmediatamente por detrás de los alvéolos, y queda dividido en prepaladar, mediopaladar y postpaladar.

El paladar blando o velo del paladar está dividido en dos zonas: zona prevelar y postvelar.

La lengua, el órgano más móvil, se divide en su cara superior, o dorso, en predorso, mediodorso y postdorso. Su punta o extremo anterior, se llama ápice, opuesto a su raíz, en el extremo posterior.

Cerrando la cavidad bucal por su parte anterior encontramos, en primer lugar, los dientes superiores, incisivos superiores, y los inferiores, incisivos inferiores. Entre los incisivos superiores y el comienzo del paladar duro, existe una zona de transición conocida con el nombre de alvéolos. Y como últimos órganos que cierran la cavidad bucal, y por su extraordinaria movilidad para cambiar fácilmente su volumen, modificando, por consiguiente, el timbre del sonido, se encuentran los labios (superior e inferior).

1.2 Clasificación articulatoria de los sonidos del habla

1.2.1 Consonantes

La manera en que los fonemas son agrupados depende de cuáles son las propiedades en común que están siendo consideradas.

Las consonantes son identificadas tradicionalmente de la siguiente manera:

- Si las cuerdas vocales vibran o no.
- El *modo* en que el sonido se produce.
- La locación de la obstrucción principal en el tracto vocal.

Cabe mencionar que los ejemplos que se muestran a continuación son con fines ilustrativos y no se presenta un listado exhaustivo. Tampoco hacen referencia particular a los sonidos del náhuatl, éstos serán revisados en la sección del estudio fonético.

Sonoridad

La posición de las cuerdas vocales cuando el aire pasa a través de ellas es una de las mayores distinciones clasificatorias de los fonemas. Cuando las cuerdas vocales se juntan y el aire pasa a través de ellas, se pueden producir vibraciones si la corriente del aire es lo suficientemente grande. Los sonidos producidos con la vibración de las cuerdas vocales son llamados sonidos **sonoros**. Por otra parte, también pueden producirse sonidos lingüísticos con las cuerdas vocales relajadas y separadas. Cuando el aire pasa a través de las cuerdas abiertas, éstas no vibran. Los sonidos hechos con las cuerdas vocales abiertas son llamados sonidos **sordos**.

Lugar de articulación

Los sonidos del habla pueden ser identificados por la locación en el tracto vocal donde las articulaciones forman una constricción. Este criterio clasificatorio es llamado **lugar de articulación**. Debido a que las consonantes involucran cierto tipo de constricción del tracto vocal, los puntos de esa constricción pueden ser identificados. Los sonidos **bilabiales** son hechos con la cierre de los labios, algunos ejemplos son /p/, /b/, /m/, /w/. Los sonidos **labiodentales** son producidos con el labio inferior cerca de los dientes superiores, por ejemplo /f/, /v/. Los sonidos **dentales** son producidos con la punta de la lengua ya sea entre los dientes frontales, a través del borde filoso de los dientes o atrás de los dientes superiores, tenemos como ejemplo /d/.

Cuando la lengua se acerca al arco alveolar (detrás de los dientes superiores delanteros), los sonidos producidos son llamados **alveolares**, ejemplo de ellos son /t/, /s/, /z/, /n/, /l/. Presionando la lengua contra el paladar duro, localizado en el centro del techo de la boca, es también una manera de producir sonidos del habla. Cuando la lengua está alzada hacia el paladar duro, se producen sonidos **palatales**, como /y/. Los sonidos hechos con la lengua entre las regiones alveolares y palatales se llaman consonantes **alveo-palatales**. Los lugares de articulaciones dentales, alveolares y alveo-palatales son a veces llamados **coronales**. Hacia la

parte trasera del paladar está el velo, los sonidos del habla producidos en esa región se llaman fonemas **velares**, algunos ejemplos son /k/, /g/.

Finalmente, el espacio entre las cuerdas vocales es llamado la glotis y allí se producen los sonidos **glotales,** tal es el caso de /h/.

Modo de articulación

Además de las clasificaciones anteriores, hay otra muy valiosa, se trata del **modo de articulación** del fonema. Esta clasificación describe el grado y modo de cierre en el tracto vocal, o la forma de los articuladores.

Oclusivas

Las **oclusivas** se producen obstruyendo totalmente el paso de la corriente de aire. La obstrucción de la corriente de aire resulta en ausencia o poco sonido, éste es usualmente seguido de una súbita explosión de aire cuando la constricción es liberada. Existen oclusivas sonoras (como /b/, /d/ y /g/) y sordas (como /p/, /t/ y /k/). De esta manera el símbolo /p/, por ejemplo, representa un fonema que puede ser descrito como **oclusiva sorda.**

Fricativas

Las **fricativas** se producen cuando el tracto vocal es altamente constrictivo pero no está completamente cerrado. Debido a la fricción, el aire que pasa a través de la constricción es turbulento y produce un siseo o un ruido apagado. Estas constricciones ocurren en bastantes lugares del tracto vocal. Existen fricativas sonoras (como /v/, /z/) y sordas (como /f/, /s/).

Africadas

Este sonido es una combinación de una oclusiva con una fricativa con el mismo lugar de articulación y la misma sonoridad, siempre y cuando la combinación funcione como una sola unidad. Ejemplos de dos africadas son: /č/ y /¢/; /č/ es una combinación de [t] y [ʃ];/¢/ es una combinación de [t] y [s]. Aunque hay muchas combinaciones posibles oclusiva-fricativa, sólo son sonidos africados aquellos que funcionan como una sola unidad.

Nasales

Resulta útil también en la clasificación de los fonemas la identificación del paso de la corriente de aire después de que ésta deja la laringe; el aire puede abandonar el tracto vocal a través de la boca o de la nariz, inclusive ambos. Cuando el puerto del velo está cerrado, la cavidad nasal está bloqueada y el aire sale a través de la boca. Los sonidos producidos con el velo del puerto cerrado son llamados sonidos **orales**. Cuando el puerto del velo está abierto, el aire puede escapar a través de la cavidad nasal. Durante la producción de las consonantes nasales el aire es forzado a través de la nariz porque la cavidad oral está bloqueada o considerablemente reducida. Los sonidos producidos de esta manera son llamados **nasales**. Ejemplos de nasales son: /m/, /n/. Ya

que hay una completa obstrucción en la cavidad oral, las nasales son similares a las oclusivas; así, /m/ es similar a /b/, pero la primera es una nasal, mientras que la segunda es una oclusiva oral. La principal diferencia entre los dos fonemas con respecto al modo de articulación es si el puerto del velo está abierto o cerrado.

Líquidas

Las **líquidas** se producen con una ligera constricción del tracto vocal, causando poca o ninguna fricción. Ejemplos de dos fonemas líquidos son /l/ y /r/. La líquida /l/ se produce con la lengua hacia el arco alveolar con el aire escapando a los lados de la lengua, /l/ es llamada líquida **lateral**. Para articular la líquida /r/ la lengua se retrae un poco y toma leve una posición arriba, hacia la parte trasera del arco alveolar, dando a /r/ el nombre de **retroflejo**.

Aproximantes

Las **aproximantes**, también llamadas **semivocales**, son articuladas casi sin constricción; se forman definiendo la boca en configuraciones particulares. /w/ se produce redondeando los labios y levantando la parte trasera de la lengua hacia el velo; /y/ se articula sin redondeo de labios y con una ligera constricción en la región palatal.

Clasificación de consonantes

Los fonemas consonánticos pueden ser agrupados por categorías articulatorias usando la información de sonoridad, lugar de constricción en el tracto vocal y el modo de constricción. Una vez que se clasifican los fonemas de acuerdo a sus rasgos, es posible agruparlos en base a características comunes. Estos agrupamientos son llamados **clases naturales**. Las clases naturales de los fonemas comparten algunos rasgos articulatorios (ya sea lugar o modo).

1.2.2 Vocales

Las vocales y las consonantes se producen de una manera algo distinta, por eso se emplean diferentes criterios para describirlas. Las consonantes pueden ser adecuadamente identificadas en términos de cómo y donde la corriente e aire es restringida en el tracto vocal. Sin embargo, las vocales no son fácil o exactamente descritas con estos parámetros debido a que las vocales se producen con muy poca constricción. El lugar de articulación de las vocales se describe en términos de un subconjunto de los lugares de articulación consonánticos, entre palatal y velar. En la clasificación vocálica, el lugar palatal de la articulación es llamado **anterior** y el lugar velar de la articulación es llamado **posterior**. La manera de articulación de las vocales se describe en términos del grado de cierre, así como de **articulaciones secundarias** tales como el redondeo de labios. Debido a que las vocales son articulaciones formadas con una corriente de aire con mínima fricción, la abertura más estrecha que la lengua puede hacer sin causar fricción es llamada **cerrada**. La abertura más amplia de la cavidad oral en la producción de vocales es llamada **abierta**. Pueden distinguirse varios rangos intermedios de abertura entre estas aberturas. Las vocales ubicadas aproximadamente a mitad de cierre y abierta se llaman vocales **medias**. Por

lo anterior, para diferenciar a las vocales se les clasifica en términos de la configuración de la lengua, su distancia desde el paladar, su anterioridad o posterioridad en la región palatal/velar y la forma de la abertura entre los labios.

El tracto vocal puede asumir diferentes formas o configuraciones por:

- Levantar o bajar la lengua.
- Adelantar o retraer el cuerpo de la lengua.
- Levantar o bajar la quijada.
- El redondeo de los labios.

En el habla normal todas las vocales son usualmente sonoras. Sin embargo, hay casos en el habla, especialmente en los susurros, donde las vocales y las consonantes sonoras pueden ser producidas sin la vibración de las cuerdas vocales.

1.3 Modelo del tracto vocal

Rabiner y Schafer [Rab78] propusieron un modelo del tracto vocal como un sistema lineal variante en el tiempo que adopta las propiedades de la voz en su salida, vea la figura 1.5. La función de excitación es esencialmente un tren de impulsos cuasi-periódicos (para sonidos de voz sonoros) y una fuente de ruido aleatorio (para sonidos sordos). La fuente de excitación es elegida mediante un switch y es escalada a través de la ganancia G para así entrar a un filtro digital que es controlado por los parámetros del tracto vocal y que cambian en el tiempo.

En sonidos continuos como las vocales los parámetros cambian muy lentamente y el modelo funciona muy bien. Sin embargo, en sonidos transitorios, como las oclusivas, el modelo no es tan bueno. Se asume que los parámetros son constantes en intervalos de tiempo de 10-20 ms; la intención es definir la estructura de un modelo cuyos parámetros varían lentamente en el tiempo. El modelo cuenta con limitaciones para modelar sonidos nasales o fricativos sonoros; ante esto, se han desarrollado modelos más sofisticados a partir de este modelo general.

El modelo de la figura 2.5 es conveniente para el análisis LPC que será discutido en capítulos posteriores.

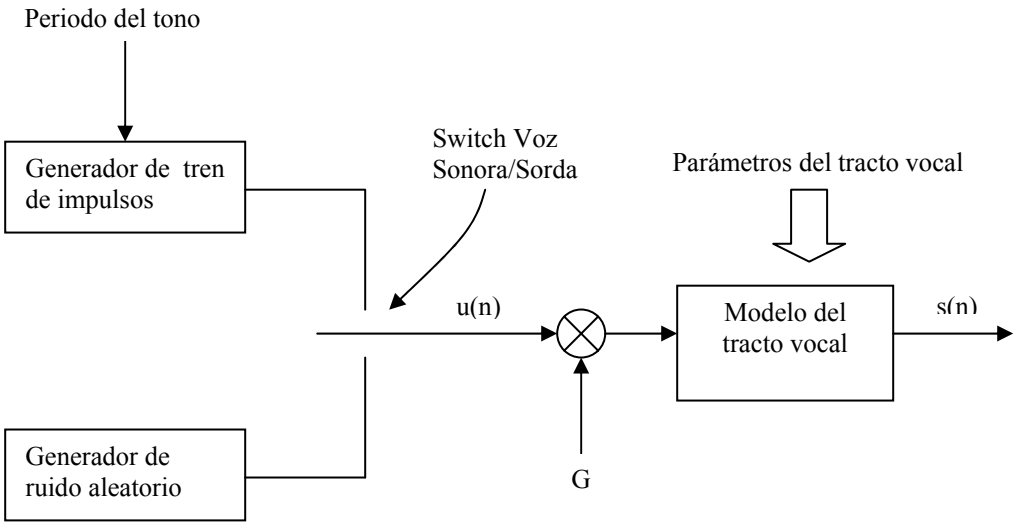


Figura 2.5: Modelo de producción de la voz

1.4 Aparato auditivo

Una fase de la comunicación es la de la recepción de la onda sonora a través del órgano receptor que conocemos con el nombre de oído. El aparato auditivo es un transductor extraordinariamente complejo, que, para su estudio, se divide en tres partes: el oído externo, el medio y el interno.

El oído externo está constituido por el pabellón auditivo u oreja (fig. 1.6), cuya misión es recoger la onda acústica y canalizarla hacia el oído medio. El pabellón auditivo desemboca en el conducto o canal acústico externo (fig. 1.6), que es una especia de resonador de unos 25 mm de largo y 8 mm de diámetro. En su parte interior termina en el tímpano.

Este conducto acústico externo actúa como un resonador, que refuerza las ondas sonoras que coinciden con sus frecuencias de resonancia: aproximadamente entre los 2.500 y 4.000 cps. En estas frecuencias, la presión del sonido que llega al tímpano es de dos a cuatro veces mayor que la presión con la que entró en el conducto acústico externo; la sensibilidad del oído mejora así notablemente en esta gama de frecuencias. De este modo el oído puede captar sonidos que por su debilidad no percibiría si el tímpano estuviese situado en el mismo pabellón auditivo, donde comienza el conducto acústico externo.

El oído medio es una cavidad llena de aire; en ella, la energía acústica de la presión de las ondas de aire se convertirá en vibraciones mecánicas. Esta conversión se realiza en la membrana del tímpano, que vibra como respuesta a los cambios en la presión del aire que llega por el conducto acústico externo.

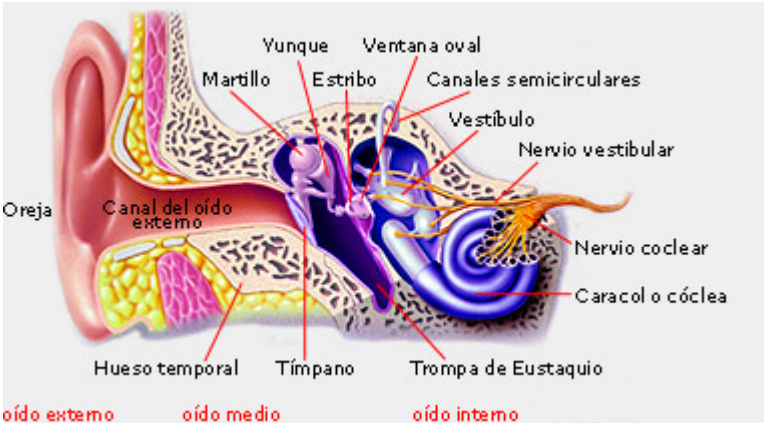


Figura 1.6 Esquema del oído [Dis]

El tímpano (fig. 1.6) es el comienzo del oído medio y su órgano esencial; es una delgada membrana elástica, relativamente rígida, con forma de cono dirigido hacia su interior. Tiene aproximadamente 1 cm. de diámetro y una superficie de unos 0.8 cm^2. El tímpano no vibra siempre de la misma forma: con las bajas frecuencias, vibra todo, pero con las altas, diferentes partes de la membrana responden a diferentes frecuencias. Sus vibraciones se encaminan hacia la cóclea a través de la cadena de huesecillos.

El oído medio termina en el oído interno hacia el cual se abre por medio de la ventana oval (fig. 1.6) y de la ventana redonda. La cadena de huesecillos (formada por el martillo, yunque y estribo, ver fig. 1.6) que atraviesa el oído medio va desde el tímpano hasta la ventana oval.

El tímpano es sensible a cualquier variación de la presión exterior: por ejemplo, a la llegada de una onda acústica. Esta presión se comunica al primero de los huesecillos, al martillo (fig. 1.7), que, al estar unido al tímpano, es sensible a todas sus variaciones. La cabeza del martillo se mueve sobre la superficie articulatoria del yunque (segundo huesecillo), el que en su parte inferior se prolonga por medio de la apófisis lenticular (así llamada por su forma), que es la que enlaza con la cabeza del estribo (tercer huesecillo). La base del estribo cierra la ventana oval del oído medio. De este modo, cualquier variación de presión sobre el tímpano se transmite por medio del martillo, del yunque y del estribo hasta el oído interno (figs. 1.6 y 1.8). La cadena de huesecillos convierte las vibraciones del tímpano en ondas hidráulicas, en el oído interno.

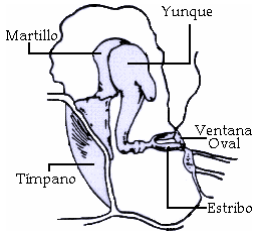


Figura 1.7 Cadena de huesecillos [Dis]

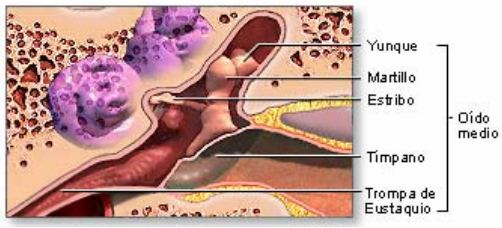


Figura 1.8 Oído medio [San06]

La caja del tímpano está cerrada por los huesos del cráneo, pero abierta hacia la faringe por medio de la *trompa de Eustaquio* (figs. 1.6 y 1.8); este conducto sirve para igualar la presión del aire contenido en el oído medio con la presión del aire exterior; sin esta condición, la membrana del tímpano no podría vibrar en perfectas condiciones.

El oído medio desempeña dos funciones:

a) Aumenta, gracias a la cadena de huesecillos, la energía acústica que desde el tímpano llega al oído interno. El tímpano, que es muy flexible entra en vibración en cuanto se produce la menor diferencia de presión entre el conducto auditivo externo y el oído medio. La cadena de huesecillos, que actúa como un conjunto de palancas, aumenta catorce veces la presión que llega a la ventana oval, con relación a la presión que tenían las ondas que llegaron al tímpano. Si consideramos, como hemos dicho antes, que entre los 2.500 y los 4.000 Hz, el conducto acústico externo multiplica de dos a cuatro veces la presión que entra a él hasta que alcanza el tímpano, resulta que para esas frecuencias la presión que llega a la ventana oval es entre veintiocho a

cincuenta y seis veces mayor, aproximadamente (2 x 14 a 4 x 14) ^{2}. Gracias a esta amplificación podemos oír sonidos que son mil veces más débiles, y no podríamos oír muchos de ellos de otra manera.

b) Protege el oído interno de los ruidos fuertes que llegan al tímpano. La ganancia de sensibilidad para los sonidos débiles que se logra por medio del oído medio puede ser perjudicial cuando los sonidos son fuertes. Para evitar este peligro, por un lado, existe el músculo tensor del tímpano, cuya finalidad es aumentar la rigidez de éste para que vibre menos; por otro lado, está el músculo del estribo, que modifica su disposición con relación al yunque; esto da como resultado la modificación del modo de vibrar el estribo; en lugar de apoyarse como un pistón sobre la ventana oval, oscila alrededor del eje longitudinal de esta última, neutralizando la vibración transmitida por la cadena de huesecillos. La acción de estos dos músculos es en parte refleja y en parte voluntaria: parece que esté unida a la previsión de la señal acústica.

El *oído interno* se llama también, a causa de su complicación, *laberinto*. Comprende dos partes: el óseo y el membranoso: este último dentro del anterior.

El laberinto óseo, de paredes óseas, comprende en su interior todas las estructuras membranosas y sensoriales que forman el oído interno. Consta de tres partes bien delimitadas:

a) El *vestíbulo*, que comunica con la caja del tímpano por medio de la ventana oval;
b) Los *tres canales semicirculares* o *canales vestibulares* (figs. 1.6, 1.9 y 1.10), óseos, de forma semicircular, situados en tres planos perpendiculares; están abiertos por sus dos extremidades a la parte posterior del vestíbulo; a estos tres canales llegan las ramificaciones del *nervio vestibular*.
c) El *caracol óseo* o *cóclea* (figs. 1.6 y 1.9). Está alojado en los huesos del cráneo. Recibe esta denominación por su forma helicoidal, que lo hace semejante a la concha del molusco. Está enrollado sobre sí mismo dos veces y media. Es la sede de las transformaciones de las vibraciones mecánicas en impulsos nerviosos.

El caracol está dividido en dos regiones (*rampas*) por medio de la *lámina espiral*. (vea en la fig. 1.10 un esquema del caracol desenrollado). La lámina espiral es ósea en su parte interna, y membranosa, en la externa. El interior de la lámina espiral forma una tercera región.

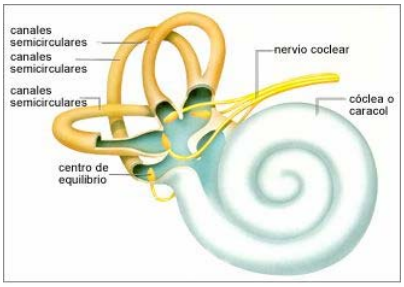


Figura 1.9 Oído interno [Dis]

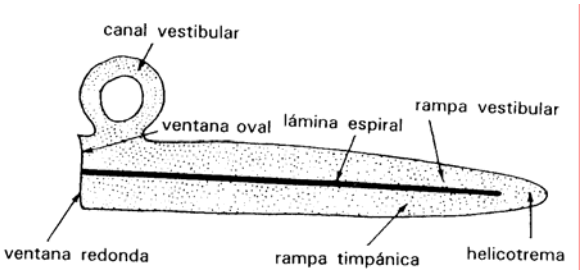


Figura 1.10 Esquema del caracol [Qui99]

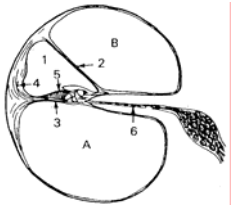


Figura 1.11 Sección del caracol: A: rampa timpánica; B: rampa vestibular;
1: conducto coclear; 2: membrana de Reissner; 3: membrana basilar;
4: ligamento espiral; 5: órgano de Corti; 6: nervio auditivo o coclear [Qui99]

La región superior se comunica libremente con el vestíbulo, y está en contacto con la ventana oval; recibe la denominación de *rampa vestibular* (fig. 1.10 y B de la fig. 1.11). La región inferior comunica con la ventana redonda; recibe el nombre de *rampa timpánica* (fig. 1.10 y A de la fig. 1.11). Estas dos rampas se comunican por el extremo del caracol, por la parte denominada *helicotrema* (fig. 1.10). Ambas están llenas de un líquido claro y viscoso denominado 'perilinfa'.

Las ondas sonoras provenientes del exterior ejercen cierta presión sobre el tímpano; esta presión se transmite, multiplicada, a través de la cadena de huesecillos a la ventana oval, la que se mueve hacia el interior presionando sobre el líquido, que, como no se puede comprimir, transforma las vibraciones mecánicas en ondas hidráulicas que se desplazan hacia la parte final del caracol, a través del helicotrema, poniendo también en movimiento la perilinfa que se encuentra en la rampa timpánica.

Si en el caracol hacemos un corte transversal, tal como muestra la fig. 1.11, nos encontramos con las siguientes partes: rampa vestibular en la parte superior (B); rampa timpánica en la parte inferior (A); la parte central está ocupada por el *conducto coclear* (1), que está lleno de un líquido muy viscoso llamado ' endolinfa'; la membrana de Reissner (2) separa la rampa vestibular del conducto coclear; la *membrana basilar* (3) separa el conducto coclear de la rampa timpánica; por un lado, se une al *ligamento espiral* (4) que envuelve la pared externa de la cóclea. La membrana basilar va ensanchándose a medida que se aproxima al helicotrema. Contiene unas 24.000 fibras elásticas, transversalmente colocadas, que son las que perciben los primeros cambios de presión.

Los movimientos mecánicos que recibe la membrana basilar se convierten en señales que se transmiten al cerebro. Esta conversión se realiza del siguiente modo: por encima de la membrana basilar se encuentra el *órgano de Corti* (5 de la fig. 1.11), que convierte la energía

hidráulica en energía bioeléctrica; está constituido por unas 25.000 células ciliadas; uno de los extremos de estas células ciliadas se encuentra en la membrana basilar. De estas células ciliadas parten después las fibras nerviosas que, reunidas en haz, dan lugar al *nervio auditivo* o *coclear* (6 de la fig. 1.11). El movimiento que percibe la membrana basilar se transmite a través de las células ciliadas hasta el nervio auditivo, compuesto de unas 30.000 neuronas. Bajo la acción de un estímulo, una neurona responda transmitiendo una serie de impulsos u ondas de actividad de naturaleza electroquímica.

El nervio auditivo es el que conduce los influjos recibidos en el oído interno hasta la zona auditiva cerebral. Atraviesa, por el conducto auditivo interno, el hueso que separa el oído interno de la cavidad craneana, para penetrar inmediatamente en los centros nerviosos al nivel del bulbo raquídeo. Las fibras de cada nervio auditivo suben hasta el cerebro; una parte de ellas redirige hacia el hemisferio cerebral situado en el mismo lado del oído de donde proceden, y la otra parte atraviesa el bulbo raquídeo y va a parar al otro hemisferio cerebral. En cada uno de los dos hemisferios, las fibras auditivas llegan a la *zona auditiva*, región localizada en la corteza cerebral. Así, cada uno de los dos hemisferios, por separado, recibe las sensaciones de cada uno de los oídos, de tal modo que la destrucción de una de esas zonas auditivas no impide la audición. Como es fácilmente comprensible, oímos siempre mejor con los dos oídos que con uno solo, hasta tal punto que se ha admitido que cuando un sonido llega a los dos oídos se percibe una sensación doble que si llega solamente a uno. En el momento actual no conocemos experimentalmente la transformación del fenómeno fisiológico en psíquico, ni cuáles puedan ser sus modalidades. Lo único que sabemos es que la zona auditiva del cerebro es el soporte material necesario para la audición.

2. FONÉTICA Y FONOLOGÍA

2.1 Lengua y habla

La *lengua* es un modelo general y constante que existe en la conciencia de todos los miembros de una comunidad lingüística determinada. Es el sistema supraindividual, una abstracción que determina el proceso de comunicación humana [Qui92].

El *habla* es la realización concreta de la lengua en un momento y en un lugar determinados en cada uno de los miembros de esa comunidad lingüística.

La *lengua*, por lo tanto, es un fenómeno social, mientras que el *habla* es individual.

Cuando dos individuos hablan, comunicándose sus pensamientos, sus ideas, comprendiéndose entre sí, es porque existe algo común a ellos y que está en un plano superior a ellos mismos; es decir, se entienden porque existe la *lengua*, el modelo lingüístico común a los dos, el sistema que establece ciertas reglas a las que someten cuando hablan; y en el momento que expresan sus ideas oralmente, están realizando, materializando la lengua en cada uno de ellos, están practicando un acto de *habla*.

El plano de la *lengua* y el plano del *habla* se suponen recíprocamente: sin actos concretos de habla, la lengua no existiría, y los actos concretos de habla no servirían para la comunicación, para entenderse, si no existiese la lengua, que establece las normas por las que ha de regirse el habla. Por lo tanto, los dos planos están unidos inseparablemente y constituyen los dos aspectos del fenómeno conocido con el nombre de *lenguaje*.

También debemos tener en cuenta que todo lo que pertenece al lenguaje, es decir, tanto al plano de la lengua como al del habla, tiene dos facetas: el *significante* (la expresión) y el *significado* (el contenido, el concepto, la idea): ambos constituyen el signo lingüístico. Es decir:

Significante + significado = signo lingüístico

Un signo lingüístico como mesa está formado por un *significante*, que sería: /m + é + s + a /, es decir, por la suma de unos elementos fónicos y por un *significado*, que sería la idea o el concepto que nosotros tenemos de lo que es una *mesa*.

Cada una de estas dos facetas del signo lingüístico tiene su función en el plano de la lengua y en el plano del habla.

El *significado* en el plano del habla es siempre una comunicación concreta, que tiene sentido únicamente en su totalidad. En el plano de la lengua, por el contrario, está representado por reglas abstractas (sintácticas, fraseológicas, morfológicas y lexicales).

El *significante*, en el plano del habla, es una corriente sonora concreta, un fenómeno físico capaz de ser percibido por el oído. En el plano de la lengua, es un sistema de reglas que ordenan el aspecto fónico del plano del habla.

El significado en la lengua consiste, pues, en un número limitado, finito de unidades, mientras que en el habla el número de unidades es infinito, ilimitado. Del mismo modo, el significante en el habla representa un número infinito de realizaciones articulatorias, pero en la lengua, sin embargo, este número es finito.

Cuando el hombre habla emite sonidos; pero hay que tener presente que los sonidos no son realizados de igual manera por todos los individuos de una misma colectividad, y que no todos los sonidos tienen en todo momento el mismo lugar articulatorio, sino que muchas veces se encuentran modificados por el contexto fónico que los rodea.

2.2 Fonética y fonología

En el plano del habla, el significante se ocupará de estudiar detalladamente la realización articulatoria y acústica de los sonidos que constituyen una lengua dada, mientras que en el plano de la lengua estudiará aquellos "sonidos" que tienen un valor diferenciador, distintivo en cuanto al significado. Del estudio del significante en el habla se ocupará la *fonética*, mientras que del estudio del significante en la lengua se ocupará la *Fonología*. Los elementos fónicos que estudia la fonética son los *sonidos*, y los elementos fónicos que estudia la fonología son los *fonemas*. Es de notar que el valor y desarrollo de la Fonología y de la Fonética se condicionan mutuamente [Qui92].

En este momento podemos entonces definir a un fonema como la unidad fonológica más pequeña en que puede dividirse un conjunto fónico recibe el nombre de fonema. Una palabra, como, por ejemplo, /páso/ *paso*, está formada por una serie de cuatro fonemas, ya que el máximo de unidades mínimas en que puede ser dividida es /p/ + /a/ + /s/ + /o/, sin que podamos fragmentar cada uno de estos fonemas en elementos más pequeños; tanto la /p/, como la /a/, como la /s/, como la /o/ son unidades completamente indivisibles.

En la ciencia del habla, el término fonema es empleado para denotar cualquiera de las unidades mínimas de los sonidos del habla en una lengua y que pueden servir para distinguir una palabra de otra. Convencionalmente se emplea el término *fono* para denotar la realización acústica de un fonema [Ace01].

Por otra parte, según se ha visto, un fonema puede tener diferentes realizaciones fonéticas, de acuerdo con el contexto en que se halle situado. Lo anterior se refiere a que la realización de un fonema dado, según las modificaciones que sufra por la acción de los sonidos que lo rodean; puede variar su lugar de articulación sin que por ello cambie el valor significativo de la palabra. Estos sonidos nuevos que resultan reciben el nombre de alófonos.

2.3 Sílaba

La sílaba es una unidad intermedia entre los fonos y el nivel de palabra, está definida como una unidad de sonido compuesta de la secuencia organizada de sonidos del habla y consta de:

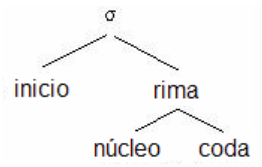
- Un pico central de sonoridad (usualmente una vocal), y
- las consonantes que se agrupan alrededor de este pico central.

Acero, Huang y Hon mencionan que la sílaba a veces está pensada como el dominio primario de coarticulación, es decir, que los sonidos dentro de una sílaba influyen más entre sí que los mismos sonidos separados por límites silábicos [Ace01].

La estructura general de una sílaba consta de las siguientes partes [Sil2]:

Partes	Descripción	Presencia
Comienzo (onset)	Segmento inicial de una sílaba	Opcional
Rima	Núcleo de una sílaba, consiste de un núcleo y coda (ver abajo)	Obligatorio
- Núcleo	Segmento central de una sílaba	Obligatorio
- Coda	Segmento de cierre de una sílaba	Opcional

El diagrama de la estructura de una sílaba es el siguiente:



Típicamente el núcleo de la sílaba es un sonido sonoro, usualmente una vocal. También es posible que el núcleo esté compuesto de un “diptongo o triptongo; o bien pueden ser consonantes sonoras como /l/ o /r/” [Wik].

Algunos tipos de sílabas son los siguientes [Sil2]:

Tipo	Descripción	Ejemplo
Pesada	Tiene una rima bifurcada. Todas las sílabas con un núcleo bifurcado (vocales largas) son consideradas pesadas. Algunos lenguajes tratan a las sílabas con una vocal corta (núcleo) seguida de una consonante (coda) como sílabas pesadas.	CV:C, CVCC, CVC
Ligera	Tiene una rima no bifurcada (vocal corta). Algunos lenguajes tratan a las sílabas con vocal corta (núcleo) seguida de una consonante (coda) como ligera.	CV, CVC
Cerrada	Termina con una consonante coda	CVC, CVCC, VC
Abierta	No tiene consonante final	CV

Fenómenos en la posición final de la sílaba

Existen fenómenos que se presentan en los sonidos en posición coda, son procesos comunes que parecen ser un tipo de debilitamiento en que ciertos rasgos se pierden o se disminuyen, o inclusive ganar otros. Lo anterior se presenta cuando el sonido adyacente de la sílaba siguiente es sordo. Por otra parte, cuando el sonido ubicado en posición coda está adyacente a un sonido sonoro de la sílaba siguiente no habrá pérdida de rasgos.

Algunos fenómenos son presentados por Marlett [Sil3] y se mencionan a continuación:

Ensordecimiento. Puede ser total o parcial. "En español se puede percibir un ensordecimiento de la vibrante en posición final del enunciado en algunos estilos de algunas variantes, como en la palabra *flor*" [Sil3].

Pérdida de otros rasgos. Es común que una consonante pierda algunos rasgos distintivos en posición coda. Por ejemplo, mientras que una oclusiva se vuelve oclusiva glotal por la pérdida de los rasgos de punto de articulación, un sonido [+continuo] se vuelve fricativa glotal en algunas lenguas. En algunas variantes de náhuatl, una aproximante en posición final de sílaba siempre se vuelve [h]. Entonces [h] es alófono de /j/ y también de /w/. (…) A modo similar, en ciertas variantes del español, la sibilante pierde sus rasgos de punto de articulación y se vuelve [h] en posición final de sílaba.

Velarización. Es bastante común encontrar un proceso de velarización en posición final. Por ejemplo, hay variantes del español en que una nasal se velariza en esta posición cuando no hay una consonante a cuyo punto de articulación puede asimilar.

Otros procesos. En español, una vibrante se hace un tipo de vibrante múltiple, optativamente, cuando se presenta en posición final de sílaba: *bárbaro*, *par*.

3. PROCESAMIENTO DIGITAL DE VOZ

En los capítulos anteriores se examinó la producción y percepción de la voz natural. Muchas aplicaciones del procesamiento de voz (p.e. codificación, síntesis, reconocimiento) explotan estas propiedades para llevar a cabo sus tareas.

Para el almacenamiento de voz o reconocimiento, la eliminación de información redundante y aspectos irrelevantes de la forma de onda de una señal de voz simplifica la manipulación de datos. Una representación eficiente para el reconocimiento de voz sería un juego de parámetros que sea consistente a través de los hablantes y que produzca valores similares para los mismos fonemas pronunciados por varios hablantes mientras exhiban variaciones confiables para diferentes fonemas.

Este capítulo presenta y describe técnicas para extraer propiedades o características de una señal de voz s(n) para poder analizarla. El propósito de estas técnicas es obtener representaciones más útiles de la señal de voz en términos de parámetros que contengan información relevante en un formato eficiente. Esto involucra la transformación de s(n) en otra señal, un conjunto de señales o un juego de parámetros, con el objetivo de simplificación y reducción de datos. Los análisis pueden realizarse tanto en el dominio del tiempo como de la frecuencia.

3.1 Análisis de voz en tiempo corto

La voz es dinámica o variante en el tiempo; parte de la variación está bajo el control del hablante, pero mucha de está variación es aleatoria. Algunos aspectos de la señal de voz directamente bajo el control del hablante (p.e. amplitud, sonoridad, frecuencia fundamental, forma del tracto vocal), por lo que existen métodos para extraer tales parámetros de la señal de voz.

Durante el habla lenta, la forma del tracto vocal y el tipo de excitación no se altera en duraciones de hasta 200 ms. Sin embargo, muchas veces cambian más rápidamente debido a que la duración de un fonema tiene un promedio de alrededor de 80 ms. La coarticulación y la frecuencia fundamental pueden hacer que cada periodo del pitch sea diferente al de su vecino. No obstante en el análisis de voz se asume que las propiedades de la señal cambian relativamente lento en el tiempo. Esto permite examinar una ventana de tiempo corto de la voz para extraer parámetros que presumiblemente permanecen fijos durante la duración de dicha ventana. Muchas técnicas producen parámetros promediados sobre el curso de la ventana en el tiempo. Así, para modelar parámetros dinámicos, se debe dividir la señal en ventanas sucesivas o *análisis de tramas* y así poder calcular los parámetros lo suficientemente seguido parea seguir cambios relevantes (p.e. debidos a las configuraciones dinámicas del tracto vocal).

3.1.1 Ventaneo

El ventaneo es la multiplicación de una señal de voz s(n) por una ventana w(n), la cual produce un conjunto de muestras de voz x(n) ponderadas por la forma de la ventana. w(n) puede tener una duración infinita, pero muchas ventanas prácticas tienen longitud finita para simplificar el cómputo. Recorriendo w(n) es posible examinar cualquier parte de s(n) a través de una ventana móvil (vea la figura 3.1).

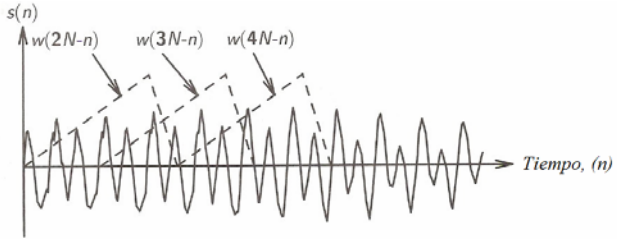


Figura 3.1: Señal de voz s(n) con tres ventanas superimpuestas (para ilustración se muestra una ventana asimétrica) [Osh00]

Muchas aplicaciones prefieren algunos promedios de la voz para producir un contorno como parámetro de salida (vs tiempo) que represente algunos aspectos fisiológicos de lenta variación de los movimientos del tracto vocal. La cantidad de suavización deseada conduce a un compromiso de tres factores en la selección del tamaño de la ventana: (1) w(n) lo suficientemente corta de tal manera que las propiedades de voz de interés cambien poco dentro de la ventana, (2) w(n) lo suficientemente larga para permitir el cálculo de los parámetros deseados (p.e. si estuviera presente un ruido aditivo, las ventanas largas pueden dar un promedio del ruido aleatorio), (3) ventanas sucesivas no tan cortas como para omitir secciones de s(n) como un análisis que se repite periódicamente. Esta última condición se refleja más en la tasa de ventanas (número de veces por segundo en que se desarrolla el análisis de voz, avanzando la ventana periódicamente en el tiempo) que en el tamaño de la ventana. Algunas de las propuestas en el traslape de ventanas, y que está relacionado con la tasa de ventanas, vienen por parte de Rabiner y O'Shaughnessy, el primero recomienda un traslape de dos terceras partes del tamaño de la ventana [Rab93], mientras que el segundo menciona un 50% de traslape [Osh00].

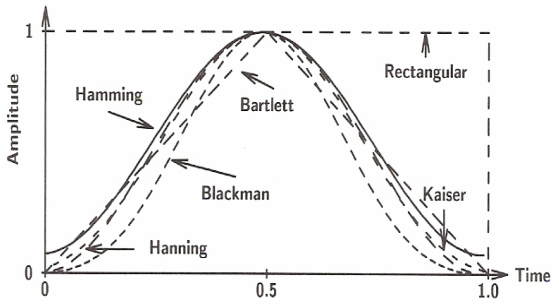
El tamaño y forma de w(n) depende de sus efectos en el análisis de la señal de voz. Típicamente w(n) es suave, porque sus valores determinan la ponderación de s(n) y todas las muestras son igualmente relevantes a priori. Salvo en sus extremos, w(n) rara vez tiene cambios súbitos; en particular, las ventanas rara vez contienen valores negativos o iguales a cero. La ventana más común y simple tiene una forma rectangular r(n):

$$w(n) = r(n) = \begin{cases} 1 & para\ 0 \le n \le N-1 \\ 0 & c.c. \end{cases} \qquad (3.1)$$

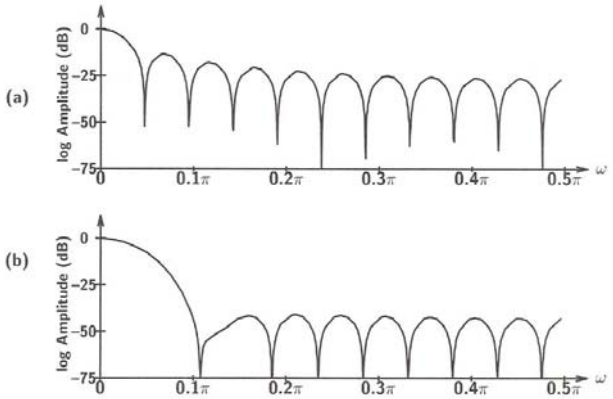
Algunas de las ventanas más comunes son: la ventana Rectangular, de Hamming, de Hanning, Blackman y Kaiser. Entre ellas, es la ventana de Hamming una de las más empleadas en el procesamiento digital de voz; su expresión es:

$$w(n) = \begin{cases} 0.54 - 0.46\cos\left(\dfrac{2\pi n}{N-1}\right) & para\ 0 \le n \le N-1 \\ 0 & c.c. \end{cases} \qquad (3.2)$$

La figura 3.2 muestra algunos tipos de ventanas en el dominio del tiempo y los espectros de magnitud de dos de ellas.



i) Ventanas en el dominio del tiempo



ii) Espectro de magnitud de ventanas: (a) rectangular; (b) Hamming

Figura 3.2. Tipos de ventanas [Osh00]

3.1.2 El efecto de la ventana [Opp00]

El principal propósito de la ventana en la transformada de Fourier de tiempo corto es limitar la extensión de la secuencia que se va a transformar de forma que las características espectrales sean razonablemente estacionarias en el intervalo de duración de la ventana. Cuanto más rápidamente cambien las características de la señal, más corta deberá ser la ventana.

A medida que la longitud de la ventana decrece, la resolución en frecuencia también decrece. Por otra parte, a medida que decrece la longitud de la ventana, aumenta la potencialidad de resolver cambios en el tiempo. Por lo anterior, se presenta un compromiso en la selección de la longitud de la ventana, entre resolución en el tiempo y en la frecuencia.

3.2 Parámetros en el dominio del tiempo

El objetivo del procesamiento de una señal de voz es obtener una representación más conveniente o útil de la información que posee. La precisión requerida para esta representación es dictada por la información particular de una señal de voz que se desea mantener o, en algunos casos, hacer más prominente. Por ejemplo, el propósito del procesamiento digital puede ser el facilitar la determinación de que una señal sea de voz o no. De manera similar pero de alguna manera más complicada, podemos hacer una clasificación en voz sonora, voz sorda, o silencio (ruido de fondo). En tales casos, se prefiere una representación que descarte información "irrelevante" y ponga claramente en evidencia las características deseadas, sobre una representación más detallada que retenga toda la información inherente. Otras situaciones (por ejemplo, transmisión digital) pueden requerir la representación más precisa que se pueda obtener de la señal voz, dadas una serie de limitantes.

Las técnicas de procesamiento que son llamadas conocidas como **métodos en el dominio del tiempo**, son métodos de procesamiento que involucran a la forma de onda de la señal directamente. Algunos ejemplos de la representación de una señal de voz en términos de las mediciones en el dominio del tiempo incluyen la *tasa promedio de cruces por cero, energía*, y la *función de auto correlación*. Tales representaciones son atractivas porque el procesamiento digital requerido es muy simple de implementar, y, a pesar de su simplicidad, las representaciones resultantes proveen una base útil para la estimación de parámetros importantes de la señal de voz.

La principal hipótesis en la mayoría de los esquemas de procesamiento de voz es que las propiedades de la señal de voz cambian de lentamente con el tiempo. Esta hipótesis lleva a una variedad de métodos de procesamiento en "tiempo corto" en los que se aíslan y procesan pequeños segmentos de la señal de voz como si fueran segmentos de un sonido con propiedades fijas. Esto se repite (usualmente de manera periódica) con tanta frecuencia como se desee. A menudo estos segmentos cortos, que son llamados algunas veces *tramas de análisis*, se traslapan uno con otro. El resultado del procesamiento en cada trama puede ser un número, o un conjunto de números. Por tanto tal procesamiento produce una nueva secuencia dependiente del tiempo que puede servir como representación de la señal de voz.

3.2.1 Energía y magnitud promedio en tiempo corto

La amplitud de la señal de voz varía apreciablemente en el tiempo. En particular, la amplitud de los segmentos de voz sordos es generalmente mucho menor que la amplitud de los segmentos de voz sonoros. La energía en tiempo corto de la señal de voz provee una representación conveniente que refleja estas variaciones en amplitud. En general, definimos la energía en tiempo corto como:

$$E_n = \sum_{m=0}^{N-1} \left| x(m) w(n-m) \right|^2 \qquad (3.3)$$

El mayor significado de E_n es que provee la base para distinguir entre segmentos de voz sonoros de segmentos de voz sordos. Los valores de E_n para segmentos sordos son significativamente menores que los de segmentos sonoros. La función de energía puede también ser usada para localizar aproximadamente el tiempo en el cual la voz sonora se vuelve sorda, y viceversa, y, para voz de alta calidad (alta relación señal a ruido), la energía puede ser usada para distinguir la voz del silencio.

Sin embargo, una dificultad con la función de energía en tiempo corto es que es muy sensible a niveles altos de señal (ya que se calcula el cuadrado), por tanto enfatiza las variaciones grandes entre muestra y muestra de x(n). Una manera simple de solucionar este problema es definir una función **magnitud promedio en tiempo corto**. La función de magnitud está dada por:

$$M_n = \sum_{m=0}^{N-1} |x(m)| w(n-m) \tag{3.4}$$

Gracias a la magnitud promedio en tiempo corto las diferencias en nivel entre señales sonoras y sordas no son tan pronunciadas como en la energía en tiempo corto.

3.2.2 Tasa promedio de cruces por cero en tiempo corto

En el contexto de las señales en tiempo discreto, se dice que un cruce por cero ocurre si muestras sucesivas tienen signos algebraicos diferentes. La frecuencia a la que ocurren los cruces por cero es una medida simple del contenido en frecuencia de la señal. La tasa de cruces por cero está dada por la siguiente expresión:

$$Z_n = \frac{\sum_{m=0}^{N-2} |sign[x(m+1)] - sign[x(m)]| w(n-m)}{2N} \tag{3.5}$$

donde

$$\text{sgn}[x(n)] = \begin{cases} 1 & para\ x(n) \geq 0 \\ -1 & c.c. \end{cases} \tag{3.6}$$

Veamos ahora como la tasa promedio de cruces por cero en tiempo corto se aplica a señales de voz. El modelo para la producción de voz sugiere que la energía de la voz sonora se encuentra concentrada por debajo de los 3 kHz debido a la caída del espectro introducida por la onda glotal, mientras que para la voz sorda, la mayor parte de la energía se encuentra a altas frecuencias. Debido a que las altas frecuencias implican altas tasas de cruces por cero, y bajas frecuencias implican bajas tasas de cruces por cero, existe una fuerte correlación entre la tasa de cruces por cero y la distribución de energía con la frecuencia. Una generalización razonable es

que si la tasa de cruces por cero es alta, la señal de voz es sorda, mientras que si la tasa de cruces por cero es baja, la señal de voz es sonora. Esto, sin embargo, es una afirmación muy imprecisa ya que debe establecerse qué es alto y qué es bajo, y por supuesto, no es posible ser preciso. A veces no es posible tomar una decisión sordo/sonoro inequívoca basada únicamente en la tasa promedio de cruces por cero. Sin embargo, este método por lo general es bastante útil.

3.2.3 Detección de inicio y fin de palabra

El problema de detección de inicio y fin de una palabra en presencia de ruido ambiental es complicado. Para grabaciones en habitaciones a prueba de ruido se puede recurrir al uso de la energía en tiempo corto para la detección, sin embargo en ambientes con ruido es necesario tomar otras consideraciones.

El poder determinar el inicio y el final de una palabra proporciona ciertas ventajas en los sistemas de reconocimiento para:

- Procesar menor número de información.
- Comparar únicamente los patrones de información.
- Evitar confusiones a causa del ruido o señales de fondo.

Sin embargo algunos de los problemas que se presentan en la detección son:

- Espurias de ruido que se pueden confundir con la señal.
- Silencios contenidos dentro de las palabras que tienen fonemas plosivos (p.e. /t/, /p/, /k/) que pueden confundirse con un falso principio o fin.
- Los fonemas fricativos (p.e. /s/, /h/), ya que tienen baja energía.
- Sonidos cortos (p.e. /t/, /p/, /k/).
- Detección de fonemas nasales al final de la palabra (baja energía y cruces por cero)
- Respiraciones del locutor, que pueden confundirse por su duración.
- Los micrófonos tienen resonancia después de de pronunciar una palabra (sobre todo en vocales).
- Los niveles de ruido pueden confundirse con la señal de voz.

Método para la detección de inicio - fin (Rabiner–Sambur) [Rab93]

Ante las dificultades para la detección del inicio y fin de palabra, Rabiner y Sambur desarrollaron un método basado en las características sonoras de los sonidos:

Sonidos sonoros	Tiene alto contenido en energía. Ocupan las frecuencias bajas del espectro de la voz humana.
Sonidos sordos	Tienen bajo contenido de energía. Ocupan las frecuencias superiores del espectro de la voz humana.

De esta forma se puede implementar un detector que incluya ambas características a través del análisis de frecuencia y energía; es decir, calculando sus valores de cruces por cero y magnitud promedio (o energía en tiempo corto) respectivamente.

Algoritmo de Detección de Inicio y Fin de palabra:

I. Detección de inicio.

1) Por cada trama de N muestras, se calculan las funciones de cruce por ceros, Z_n, y la magnitud promedio de la señal, M_n.
2) Para obtener las estadísticas del ruido ambiental, se considera que las primeras diez tramas son ruido. Es decir,

$$Ms_n = \{M_1, M_2, \cdots, M_{10}\}$$
$$Zs_n = \{Z_1, Z_2, \cdots, Z_{10}\} \quad (3.7)$$

3) Calcular la media y la desviación estándar para las características del ruido y obtener los siguientes umbrales:

Umbral	Nombre del umbral	Valor
UmbSupEnerg	Umbral Superior de Energía	$0.5 \max\{M_n\}$
UmbInfEnerg	Umbral Inferior de Energía	$\mu M_s + 2\,\sigma M_s$
UmbCruCero	Umbral de cruces por cero	$\mu Z_s + 2\,\sigma Z_s$

4) Recorrer la función M_n incrementando en una unidad a n de 11 hasta que $M_n >$ UmbSupEnerg. En este punto se garantiza la presencia de señal. A este punto se marca como ln.

5) Resulta lógico pensar que el inicio se la señal se encuentra en algún punto anterior a ln, por lo que ahora recorremos la función M_n desde n = ln hasta que $M_n <$ UmbInfEnerg. Este punto lo marcaremos como le y lo reconocemos tentativamente como el inicio de la señal, determinado por la función de magnitud.

6) Ahora se decrementa n desde n=le hasta n=le–25, o en su defecto n=11, verificando si sucede alguna de las siguientes condiciones en la función de cruces por cero, ya que lo que ahora se busca es la posibilidad de que un sonido no sonoro preceda a un sonido sonoro:
   a) Si { $Z_n <$ UmbCruCero } significa que no se encontró alguna porción de la señal con aumento importante de frecuencia en 25 ventanas anteriores, por lo tanto el inicio es le.
   b) Si se encuentra que { $Z_n >$ UmbCruCero } menos de tres veces seguidas, significa que sólo fue una espiga de ruido, el punto de inicio sigue siendo le.
   c) Si se encuentra que { $Z_n >$ UmbCruCero } al menos tres veces seguidas, se encontró un sonido no sonoro; entonces se busca el punto n para el cual { $Z_n >$ UmbCruCero } la primera de las más de tres veces, es decir, el punto para el

cual la función Z_n sobrepasa el umbral, indicando el comienzo del sonido no sonoro y se desplaza el inicio de la palabra a lz.

II. Detección de fin. Para la detección de fin de la palabra, se hace lo mismo que en el inicio de palabra pero en sentido inverso a partir del punto (4) de la sección anterior, como si se detectara un inicio con la señal invertida en el tiempo.

Recomendaciones adicionales

Finalmente, existen algunas recomendaciones que fueron adoptadas en este trabajo al programar el algoritmo:

1) Calcular nuevos umbrales de ruido para la detección de fin de palabra en base a las últimas 10 tramas de la señal. Esto es debido a que las condiciones de silencio antes y después de la pronunciación de una palabra pueden ser diferentes por diferentes razones: resonancia, respiraciones. Evidentemente estas razones resultan indeseables y deben evitarse en la medida de lo posible si no se cuenta con un estudio de grabación profesional.

2) Un ambiente de grabación donde no se puede tener demasiado control del mismo puede ocasionar que la forma de onda en las regiones de silencio manifieste muy baja intensidad pero cuente con una tasa de cruces por cero muy elevada, inclusive mayor que el de un sonido sordo. Lo anterior producirá falsas detecciones de inicio o fin de palabra. Una manera de lidiar con este hecho fue aplicar un offset a la señal de voz en base a la media del valor absoluto de las primeras 5N muestras de la señal, donde N es el tamaño de la trama (recuerde que las primeras y últimas 10 tramas de la señal se consideran ruido). Se calculó un offset distinto para la detección del inicio y la detección del final de una palabra y se consideró como el paso 0 del algoritmo de Rabiner-Sambur. Esta medida permitió una mejora notable en la detección.

3) En un principio, en el análisis de fin de palabra se obtuvieron falsas detecciones. Esto se presentó en palabras que terminan en la secuencia de fonos sonoro-sordo-sonoro y donde la magnitud promedio del primer fono sonoro era mayor que la del último fono sonoro (lo cual es sumamente común ya que en el náhuatl el acento se ubica siempre en la penúltima sílaba). La diferencia de magnitudes promedio ocasionaba que el punto le se estableciera al término del primer fono sonoro y que se procediera con el análisis de cruces por cero aún cuando había un fonema sonoro a final de palabra. La manera más sencilla para resolver este problema fue bajar el umbral superior de energía a $0.2 \max\{M_n\}$, con resultados muy positivos.

4) Por último, en el paso 6 de su algoritmo, Rabiner y Sambur proponen un análisis de cruces por cero en 25 tramas anteriores a le. Tenga en cuenta que si la palabra fue pronunciada de forma pausada, se debe pensar en un número de tramas diferente. En este trabajo funcionó bien con 25 tramas, pero considero que resulta conveniente analizar aunque sea un poco la señal a detectarle su inicio y fin de palabra, para observar si las 25 tramas son suficientes.

3.2.4 Función de autocorrelación en tiempo corto

La función de autocorrelación de una señal determinística en tiempo discreto se define como:

$$\phi(k) = \sum_{m=-\infty}^{\infty} x(m)x(m+k) \tag{3.8}$$

En la práctica, el rango de suma es usualmente limitado (i.e. ventaneado). La representación de una señal mediante la función de autocorrelación es una forma conveniente de mostrar ciertas propiedades de la señal. Por ejemplo, la función de autocorrelación de una señal periódica es también periódica con el mismo periodo. Otras propiedades importantes de la función de autocorrelación son:

1. Es una función par.
2. Posee un valor máximo en k=0.
3. La cantidad $\Phi(0)$ es igual a la energía para señales determinísticas, y a la potencia promedio para señales aleatorias o periódicas.

La función de autocorrelación en tiempo corto, para una señal aleatoria o periódica, con una trama de N muestras, es:

$$R_n(k) = \sum_{m=0}^{N-1-k} \left[s(n+m)w'(m)\right]\left[s(n+m+k)w'(k+m)\right] \qquad para\ k = 0,1,2,..., p \tag{3.9}$$

Esta ecuación puede ser interpretada como sigue: primero se selecciona un segmento de la voz por la multiplicación por una ventana (donde w'(n)=w(-n)); entonces se aplica la definición determinística de la autocorrelación al segmento de voz al que se aplicó el ventaneo. Usualmente, aunque dependiendo de la aplicación de interés, la ventana es rectangular.

La función de autocorrelación en tiempo corto tiene aplicaciones en el procesamiento de señales de voz, por ejemplo para la estimación de tono (pitch), determinación de rasgos sonoros y predicción lineal.

3.3 Análisis en el dominio de la frecuencia

En muchas áreas de la ciencia y la ingeniería, la representación de señales u otras funciones mediante sumas de sinosoides o exponenciales complejas conduce a soluciones convenientes a problemas y a menudo a una mayor penetración en fenómenos físicos que la que está disponible mediante otros medios. Tales representaciones – representaciones de Fourier como se llaman comúnmente- son útiles en el procesamiento de señales por dos razones básicas. La primera es que para sistemas lineales es muy conveniente determinar la respuesta a una superposición de sinusoides o exponenciales complejas. La segunda razón es que las representaciones de Fourier a menudo sirven para poner en evidencia algunas propiedades de la señal que podrían ser menos evidentes en la señal original.

La investigación y tecnología en comunicaciones de voz son áreas donde tradicionalmente el concepto de una representación de Fourier ha tomado un papel importante. Para observar el por qué de esto, es útil recordar que el modelo de producción de un sonido de voz de estado estable como una vocal o una fricativa, simplemente consiste en un sistema lineal excitado por una fuente que varía con el tiempo ya sea de manera periódica o aleatoria. En general, el espectro de la salida de tal modelo será el producto de la respuesta en frecuencia del sistema del tracto vocal y el espectro de la excitación. Por tanto, se esperaría que el espectro de la salida reflejara las propiedades tanto de la respuesta en frecuencia de la excitación como del tracto vocal. Hemos visto, sin embargo, que las ondas de voz son generalmente más complicadas que una simple vocal sostenida o un sonido fricativo. Por lo tanto las representaciones de Fourier estándares que son apropiadas para señales periódicas, transitorias o aleatorias estacionarias, no son directamente aplicables a la representación de señales de voz cuyas propiedades cambian notablemente en función del tiempo. Sin embargo, el principio del análisis en tiempo corto es una aproximación válida para el procesamiento de voz. “Las propiedades temporales como energía, cruces por cero y correlación pueden asumirse como fijas en intervalos de tiempo del orden de 10 a 30ms” [Rab93]. Las propiedades espectrales de la voz, de manera similar, cambian de manera relativamente lenta con el tiempo.

En contraste con los métodos en el tiempo, los métodos en el dominio de la frecuencia involucran (explícita o implícitamente) alguna forma de representación del espectro.

Una señal no estacionaria es una señal cuyas propiedades (amplitudes, frecuencias y fases) varían con el tiempo. Para el caso de señales de voz, la transformada de Fourier de tiempo corto (denominada también transformada de Fourier dependiente del tiempo [Opp00]) proporciona a menudo una descripción útil de cómo las propiedades de la señal cambian con el tiempo.

La voz es claramente una señal no estacionaria. Sin embargo, se puede suponer que las características de la señal permanecen esencialmente constantes en intervalos de tiempo del orden de 30 o 40 ms [Opp00]. Por otra parte, el contenido en frecuencia de la señal de voz puede abarcar hasta 15 kHz o más [Opp00], pero la voz es altamente inteligible incluso con bandas de frecuencia limitadas a unos 3 kHz. Los sistemas telefónicos comerciales, por ejemplo, limitan

típicamente a unos 3 kHz la frecuencia más alta transmitida. La frecuencia de muestreo estándar para sistemas digitales de comunicación telefónica es de 8000 muestras/s. Con esta frecuencia de muestreo, un intervalo de 40 ms tiene 320 muestras.

3.3.1 Preénfasis

La descripción del modelo fuente-filtro para la producción de voz mostrado el la figura 2.1 del capítulo 2, indica que el espectro de sonidos sonoros posee una pendiente de -6 dB/octava, mientras se incrementa la frecuencia. Esta es una combinación de una pendiente de -12 dB/octava debido a la fuente de excitación de la voz y una pendiente +6 dB/octava debida a la radiación producida por la boca. Esto significa que por cada vez que se duplica la frecuencia, la señal en amplitud, y por tanto la respuesta del tracto vocal medida, son reducidas por un factor de 16. Es por tanto deseable compensar la caída de -6 dB/octava pre-procesando la señal de voz para darle un levantamiento de +6 dB/octava en el rango apropiado de tal forma que el espectro medido tenga un rango dinámico similar a través de toda la banda de frecuencias. Esto se llama **preénfasis**. En un sistema digital de procesamiento de señales, el preénfasis puede ser implementado ya sea de primer orden, usando un filtro analógico paso-altas con una frecuencia de corte a 3 dB en algún lugar entre 100 Hz y 1 kHz (la posición exacta no es crítica) que precede al filtro de anti-aliasing y el convertidor A/D, o bien usando un filtro digital paso-altas que procesa la señal de voz digitalizada. El filtrado paso-alto puede ser logrado digitalmente usando la ecuación en diferencias:

$$y(n) = x(n) - ax(n-1) \tag{3.10}$$

Donde y(n) denota la muestra actual de salida del filtro de preénfasis, x(n) es la muestra actual de entrada, x(n-1) es la muestra anterior de entrada y a es una constante usualmente elegida entre 0.9 y 1. De nuevo, el valor elegido no es crítico.

En el caso de sonidos sordos no hay necesidad de aplicar preénfasis, ya que no hay ninguna pendiente que deba ser removida. Sin embargo, por simplicidad, se aplica preénfasis normalmente a sonidos sordos también.

3.3.2 Análisis de Fourier en tiempo corto

Como la técnica espectral tradicional, el análisis de Fourier brinda una representación de la señal de voz en términos de amplitud y fase como una función de la frecuencia. Observando el modelo tracto vocal como un sistema lineal, la transformada de Fourier de la voz es el producto de las transformadas de la excitación glotal (o de la fuente de ruido) y de la respuesta del tracto vocal. Para vocales de estado estable o sonidos fricativos, la transformada de Fourier básica (de tiempo infinito) puede emplearse extendiendo o repitiendo secciones o periodos de pitch de la voz ad infinitum. Sin embargo la señal de voz no es estacionaria y por lo tanto resulta necesario el análisis en tiempo corto de la señal usando ventanas.

La transformada de Fourier en tiempo corto de una señal s(n) se define como:

$$S_n(e^{j\omega}) = \sum_{m=-\infty}^{\infty} s(m)e^{-j\omega m} w(n-m) \qquad (3.11)$$

En la ecuación (3.11) w(n-m) es una ventana real que determina la porción de la señal de entrada que recibe énfasis en un determinado índice de tiempo *n*. La transformada de Fourier dependiente del tiempo es claramente función de dos variables: el índice de tiempo *n*, que es discreta, y la variable de frecuencia ω que es continua.

Para fines computacionales, la transformada discreta de Fourier (TDF, o DFT en inglés) es empleada en lugar de la transformada de Fourier estándar, de tal manera que la variable frecuencial ω sólo toma valores discretos sobre N (N = duración o tamaño de la ventana, de la TDF). Cabe mencionar que la transformada rápida de Fourier, o FFT, es usada para implementar la TDF, cuya expresión es:

$$S_n(k) = \sum_{m=0}^{N-1} s(m)e^{-j2\pi km/N} w(n-m) \qquad (3.12)$$

Interpretación de la transformada de Fourier

Consideremos a $X_n(e^{j\omega})$ como la transformada de Fourier normal de la secuencia w(n-m)x(m), $-\infty<m<\infty$, para una *n* fija. La transformada de Fourier dependiente del tiempo es una función del índice de tiempo *n*, que toma todos los valores enteros hasta "desplazar" la ventana w(n-m) sobre la secuencia x(m). Esto se muestra en la figura 3.3, que muestra a x(m) y w(n-m) como funciones de *m* para varios valores de *n*. (note que la señal y la ventana son graficadas por conveniencia como funciones continuas a pesar de que únicamente se encuentran definidas para valores enteros de *m* y *n-m*).

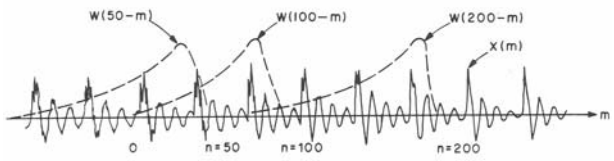


Figura 3.3: Gráficas de x(m) y w(n-m) para varios valores de n [Rab78]

Las condiciones de existencia de la representación mediante transformada de Fourier dependiente del tiempo son fácilmente obtenidas si recordamos que es condición suficiente para la existencia de la transformada de Fourier convencional que la secuencia x(m)w(n-m) sea absolutamente sumable para todos los valores de n. Si, como suele ser el caso, w(n-m) es de duración finita, entonces esta condición se satisface claramente.

Como en el caso de la transformada de Fourier normal de señales en tiempo discreto, la transformada de Fourier variante en el tiempo es periódica en ω con periodo 2π. Esto se observa fácilmente sustituyendo ω +2π en la ecuación (3.11). También note que es posible expresar la transformada de Fourier variante en el tiempo en términos de una variedad de variables de frecuencia. Por ejemplo, si ω=ΩT, donde T es el periodo de muestreo usado para obtener la secuencia x(m), entonces Ω es la frecuencia analógica en radianes. También, haciendo las sustituciones ω=2πf o ω=2πFT, podemos expresar la transformada de Fourier variante en el tiempo como función de la frecuencia cíclica normalizada (f) o la frecuencia analógica cíclica convencional (F, en Hertz) respectivamente.

El hecho de que, para un valor dado de n, $X_n(e^{j\omega})$ tiene las mismas propiedades que una transformada de Fourier normal, la secuencia de entrada x(m) puede ser recuperada exactamente a partir de la transformada de Fourier variante en el tiempo con el requerimiento de que w(0) sea diferente de cero.

Hay una expresión alterna de la transformada de Fourier de tiempo corto es la siguiente:

$$X_n(e^{j\varpi}) = \frac{1}{2\pi}\int_{-\pi}^{\pi} W(e^{j\theta})e^{j\theta n}X(e^{j(\varpi+\theta)})d\theta \qquad (3.13)$$

La forma de la secuencia de la ventana tiene un importante efecto en la naturaleza de la transformada de Fourier dependiente del tiempo. Uno de los resultados importantes que se desprende es que la transformada de la secuencia x(m), $-\infty<m<\infty$, es convolucionada con la transformada de Fourier de la ventana desplazada; por ende, las propiedades de la transformada de Fourier de la ventana, $W(e^{j\omega})$, se vuelven importantes. Por lo anterior, es importante que para una reproducción fiel de las propiedades de $X(e^{j\omega})$ en $X_n(e^{j\omega})$, la función $W(e^{j\theta})$ debe aparecer como un impulso con respecto a $X(e^{j\omega})$. En la sección pasada se discutieron las propiedades de las ventanas rectangular y de Hamming. Se mostró cómo el ancho del lóbulo principal de $W(e^{j\omega})$ es inversamente proporcional a la longitud de la ventana, mientras que los niveles de los lóbulos laterales son esencialmente independientes de la longitud de la ventana.

El análisis de los efectos de usar ventanas para el análisis espectral de la voz ha demostrado que el uso de la ventana de Hamming produce un espectro más suave de la señal de voz que el que produce una ventana rectangular. Esta última ventana produce un espectro ruidoso a consecuencia de sus grandes lóbulos laterales); este efecto indeseable es compensado por los beneficios del lóbulo principal angosto de la ventana rectangular. Sin embargo, las ventanas rectangulares son raramente usadas en el análisis espectral de la voz.

Por tanto, al respecto de la relación básica entre la duración en tiempo de la ventana y las propiedades de la transformada de Fourier de tiempo corto se concluye que la resolución en frecuencia varía inversamente con la longitud de la ventana. El propósito de la ventana es limitar el intervalo de tiempo a ser analizado para que las propiedades de la señal no cambien de manera apreciable, sin embargo se requiere un cierto compromiso: una buena resolución en el tiempo

requiere un ventana corta mientras que una buena resolución en frecuencia requiere una ventana larga.

3.3.2 Espectrogramas

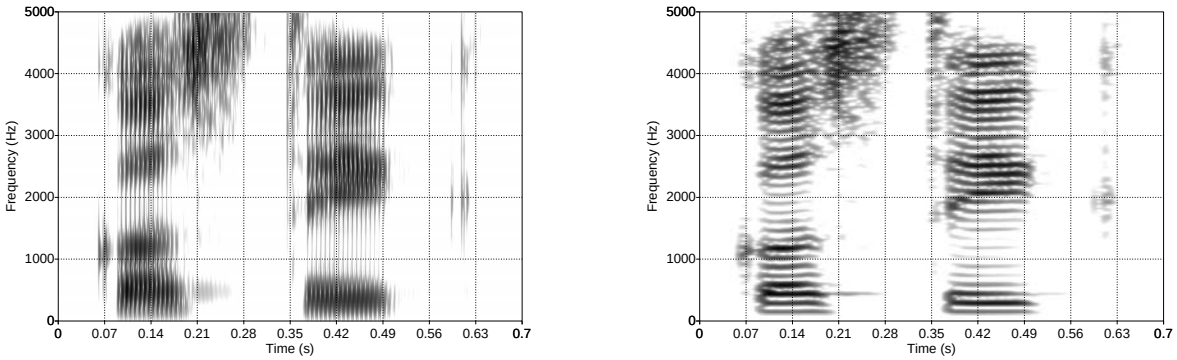
Es una herramienta básica para el análisis espectral, un espectrograma convierte una señal de voz bidimensional (amplitud/tiempo) en un patrón tridimensional (amplitud/frecuencia/tiempo). Con el tiempo y la frecuencia en los ejes horizontal y vertical, respectivamente, la amplitud es denotada por la oscuridad mostrada en el espectro. Picos en el espectro (por ejemplo, resonancias formantes) aparecen como bandas horizontales oscuras.

Los espectrogramas proporcionan mucha información relevante para la fonética acústica: las duraciones de segmentos acústicos, si la señal de voz es periódica, y el movimiento detallado de los formantes. Los espectrogramas presentan únicamente la amplitud espectral, ignorando información de la fase, pues se asume que la fase es relativamente poco importante para muchas aplicaciones de la voz.

Los espectrogramas de **banda ancha** se obtienen con ventanas de corta duración (típicamente de 10 ms). Los periodos sonoros aparecen como secuencias de bandas verticales oscuras. Estos espectrogramas resaltan el comportamiento espectral de la señal y por lo tanto hacen posible describir la evolución temporal de los formantes, por lo que son una herramienta invaluable para fonetistas. Incluso gente experta puede leer espectrogramas, es decir, la recuperación de la pronunciación a partir de su representación tiempo-frecuencia. Los espectrogramas de **banda angosta** son empleados con menor frecuencia; sin embargo, ellos ponen en evidencia la estructura espectral fina: los armónicos en secciones sonoras aparecen como bandas horizontales [Dut97].

La figura 3.4.i muestra un espectrograma de banda ancha, se caracteriza por una resolución pobre en el dominio de la frecuencia y una buena resolución en el dominio del tiempo. El espectrograma despliega el rango frecuencial de 0 Hz a 5 kHz ya que allí se presentan las características más importantes de la voz, la duración de la ventana es de 5 ms. Las barras gruesas y oscuras que se mueven horizontalmente a través del espectrograma corresponden a las frecuencias de resonancia del tracto vocal que, como vemos, cambian con el tiempo. La apariencia vertical y estriada del espectrograma se debe a la naturaleza cuasiperiódica de las porciones sonoras de la señal, como es evidente al comparar las variaciones en la señal desplegada y el espectrograma. Como la longitud de la ventana de análisis es del orden de la longitud de un periodo de la señal, mientras la ventana se desplaza en el tiempo, alternadamente cubre segmentos de alta energía de la señal y entre ellos segmentos de baja energía, produciendo las estriaciones verticales en la gráfica durante intervalos sonoros.

En un análisis de Fourier de banda angosta dependiente del tiempo, se usa una ventana más larga para proveer una mayor resolución infrecuencia con el correspondiente decremento de la resolución en el tiempo. Tal análisis de banda angosta de la voz es ilustrado en la figura 3.4.ii, en este caso la ventana tiene una duración de 30 ms.



i. Espectrograma de banda ancha ii. Espectrograma de banda angosta

Figura 3.4: Espectrogramas de banda ancha y angosta de la palabra [kóstik] (amarillo) en náhuatl

3.4 Análisis acústico del habla

Gracias al procesamiento digital de voz se cuenta con diversos métodos para el análisis acústico del habla. El habla es la representación audible del sistema del lenguaje y una facultad únicamente humana. Si estamos interesados en la facultad del lenguaje humano, entonces el estudio del habla es esencial. Ciertas áreas de la lingüística estudian las bases de esta manifestación tan común y natural en los seres humanos. La fonética es el estudio del rango completo de sonidos vocales que los seres humanos son capaces de realizar. La fonética lingüística es el estudio los sonidos que los seres humanos emplean cuando hablan una lengua. Finalmente, la fonología es el estudio del sistema subyacente en la selección y uso de sonidos en las lenguas del mundo.

Tradicionalmente, el campo de la fonética lingüística está dividido en dos áreas amplias: la fonética acústica y la fonética articulatoria. "La primera analiza las complejas ondas sonoras del habla para determinar los componentes que contienen información lingüística significativa; la segunda investiga la producción de los sonidos del habla por el aparato vocal. Desde un punto de vista puramente físico toda pronunciación es un continuo. Acústicamente es una onda sonora variante y continua, mientras que desde el punto de vista articulatorio es un flujo continuo de gestos y movimientos. No hay puntos obvios de demarcación en este continuo. No obstante, el habla es percibida y funciona lingüísticamente como una serie de unidades discretas llamadas sonidos. La meta general de la fonética lingüística es describir con precisión (tanto acústicamente como articulatoriamente) todos los tipos de sonidos del habla que funcionan en los lenguajes del mundo" [Ken79].

Existen varios métodos para el análisis acústico del habla. La selección depende de las características de los sonidos que se deseen observar. Los análisis más usuales mencionados por Joaquim Llisterri en [Lli05] son:

Oscilográfico. Para análisis de la sonoridad, la duración, las pausas, el acento y el ritmo.
Espectral (a través de FFT o LPC). Para análisis de la estructura formántica (timbre).
Espectrográfico. Para análisis de la sonoridad, la duración, la estructura formántica (timbre), la intensidad, las pausas, el acento y el ritmo.
Melódico. Para análisis de la melodía, el acento y la entonación.
De intensidad. Para análisis de la intensidad, el acento, el ritmo y las pausas.

Observe que es el análisis espectrográfico es muy completo, sin embargo eso no garantiza que despliegue todas las características del habla que se desean observar. Como se verá más adelante, este trabajo combina el análisis oscilográfico y espectrográfico.

Para denotar la importancia de los análisis acústicos, vale la pena mencionar el amplio rango de aplicaciones que enlista Llisterri [Lli05].

"Las aplicaciones del análisis acústico del habla:

- Descripción fonética de la lengua.
- Descripción de los elementos segmentales.
    - Correlatos acústicos de las propiedades fonéticas de los elementos segmentales:
        - Sonoridad.
        - Modo de articulación.
        - Lugar de articulación.
    - Manifestación acústica de procesos fonéticos y fonológicos:
        - Procesos de asimilación.
        - Procesos de debilitación y refuerzo.
        - Procesos que afectan la estructura silábica.
- Descripción de los elementos suprasegmentales.
    - Correlatos acústicos de los elementos supresegmentales:
        - Acento.
        - Melodía.
        - Ritmo.
    - Relación entre los elementos suprasegmentales y fenómenos gramaticales.
        - Modalidad oracional.
        - Estructura funcional de la oración: tematización, rematización, focalización.
- Estudio de la variación lingüística:
    - Variación geográfica. Caracterización de fenómenos fonéticos propios de la variación geográfica.
    - Variación social. Correlatos fonéticos de las variantes sociales.
    - Variación estilística. Caracterización fonética de los estilos de habla.
- Desarrollo de tecnologías del habla:
    - Conversión de texto a habla:
        - Incorporación de información fonética a los sistemas de síntesis:
        - Creación de diccionarios de unidades acústicas para la síntesis.
        - Definición de modelos prosódicos para la conversión de texto a habla.
        - Variaciones melódicas de los enunciados.
        - Duración de los sonidos.
        - Asignación automática de pausas.
    - Reconocimiento del habla.
        - Determinación de las unidades del reconocimiento.
        - Incorporación de información fonética segmental y suprasegmental para la mejora del reconocimiento.
    - Sistemas de diálogo.
        - Análisis de la especificidad de los elementos prosódicos en el diálogo.
- Análisis y rehabilitación de las patologías del habla:
    - Análisis de las características acústicas del habla patológica.
    - Estudio experimental de las consecuencias acústicas de las patologías de la producción del habla.
    - Estudio de los factores acústicos que inciden en la inteligibilidad del habla patológica.
    - Diagnóstico basado en el análisis acústico.

- Aplicación de las técnicas de análisis acústico al diseño de herramientas informáticas para la reeducación.
    - Análisis de la interferencia en la adquisición de lenguas extranjeras y corrección fonética.
- Identificación de hablantes: fonética forense.
    - Determinación de la semejanza fonética y acústica entre enunciados producidos por un mismo hablante en situaciones diferentes.
    - Determinación de la similitud fonética y acústica entre enunciados producidos por un hablante conocido y enunciados producidos por hablantes diferentes entre los cuales se puede encontrar el conocido".

4. RECONOCIMIENTO DE VOZ POR CUANTIZACION VECTORIAL (VQ)

La mayor meta del reconocimiento automático de voz es producir un sistema que pueda reconocer, con precisión humana, sin restricciones, expresiones continuas del habla no importando el hablante.

El reconocimiento de voz es el proceso de conversión de una señal acústica, capturada por un micrófono o un teléfono, a un juego de palabras. Las palabras reconocidas pueden ser los resultados finales, con aplicaciones tales como comandos y control, entrada de datos, y preparación de documentos. Ellas pueden servir también en procesos lingüísticos para la comprensión del habla. En la figura 4.1 se muestra el diagrama de bloques de un sistema que incorpora diversas tecnologías de voz.

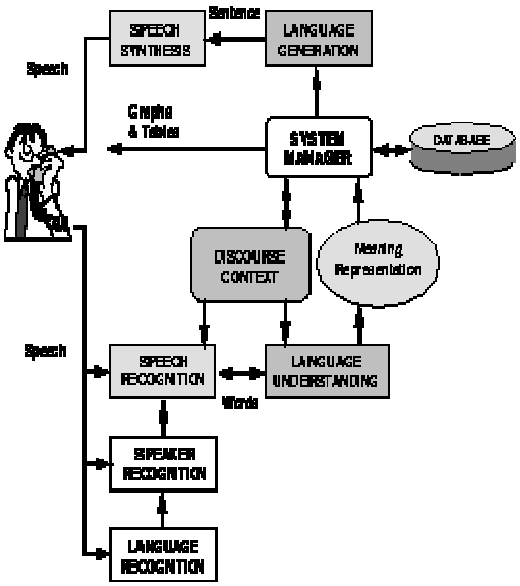


Figura 4.1: Tecnologías para interfaces de lengua hablada [Ogi]

De acuerdo a documentos sobre el estado del arte en tecnologías del lenguaje humano [Ogi], los sistemas de reconocimiento de voz pueden ser caracterizados por muchos parámetros, algunos de los más importantes se muestran en la tabla 4.1. Un sistema de reconocimiento de palabras aisladas requiere que el hablante haga breves pausas entre las palabras, mientras que en el caso de sistemas de reconocimiento de palabras continuas no lo requiere así. El habla espontánea, o generada extemporáneamente, es mucho más difícil de reconocer que el habla al leer un texto. Por otra parte, algunos sistemas requieren que un usuario (hablante) proporcione muestras de su voz antes de la utilización de ellos, mientras que hay sistemas independientes del hablante y no es necesario proporcionar tales muestras de voz. Algunos otros parámetros dependen de la tarea específica. El reconocimiento es generalmente más difícil cuando es grande

el vocabulario o tiene muchas palabras con sonidos similares. Cuando el habla es producida en una secuencia de palabras, modelos de lenguaje o gramática artificial son usados para restringir la combinación de palabras. La perplejidad está definida como la media geométrica del número de palabras que pueden seguir de una palabra después de que el modelo de lenguaje ha sido aplicado. Finalmente, hay algunos parámetros externos que pueden afectar el funcionamiento de los sistemas de reconocimiento de voz, incluyendo las características del ruido ambiental así como el tipo y ubicación del micrófono.

Parámetros	Rango
Modo de hablar	Palabras aisladas a voz continua
Estilo de hablar	Voz de lectura a voz espontánea
Inscripción	Dependiente del hablante a independiente del hablante
Vocabulario	Pequeño (< 20 palabras) a grande (> 20,000 palabras)
Modelo del lenguaje	Estado finito a sensible al contexto
Perplejidad	Pequeña (<10) a grande (> 100)
SNR	Alto (> 30 dB) a bajo (< 10 dB)
Transductor	Cancelación de voz de micrófono a teléfono

Tabla 4.1. Parámetros típicos usados para caracterizar la capacidad de sistemas de reconocimiento de voz [Ogi]

El reconocimiento de voz presenta problemas difíciles, en mayor parte debido a las muchas fuentes de variaciones asociadas con la señal. Primero, las realizaciones acústicas de los fonemas (las unidades más pequeñas de sonido de las que están compuestas las realizaciones acústicas de las palabras), son altamente dependientes del contexto en el cual ellas aparecen. En las fronteras de la palabra las variaciones contextuales pueden ser bastante dramáticas. Segundo, las variaciones acústicas pueden resultar de cambios en el ambiente así como en la posición y características del transductor. Tercero, existen variaciones del propio hablante debido a cambios de su estado físico y emocional, la tasa de habla, o calidad de la voz. Finalmente, diferencias en el fondo sociolingüístico, en el dialecto, y en el tamaño de extensión y la forma vocal pueden contribuir a variaciones a través del hablante [Ogi].

Dado este panorama, el problema del reconocimiento automático de voz manifiesta ser mucho más difícil de como fue originalmente apreciado y aún no está enteramente resuelto a pesar del gran esfuerzo de los investigadores. A pesar de esto, las investigaciones han fructificado en el desarrollo de sistemas de reconocimiento de voz limitada, las cuales funcionan razonablemente bien en algunas aplicaciones; además se han logrado progresos sustanciales en la disminución de las barreras de independencia del hablante, voz continua, y grandes vocabularios. Por otra parte, se ha hecho un gran esfuerzo en el desarrollo de grandes corpus orales para prueba, entrenamiento y desarrollo de sistemas. Algunos de estos corpus están diseñados para investigación fonético-acústica, mientras otros son para tareas altamente específicas. Actualmente no es insólito emplear decenas de miles de oraciones disponibles para prueba y entrenamiento del sistema.

En [Ogi] se ejemplifican algunos sistemas de reconocimiento, se mencionarán algunos de ellos a continuación.

Una de las más populares y potencialmente más usadas tareas con baja perplejidad (PP=11) es el reconocimiento de dígitos. Para el inglés estadounidense, el reconocimiento independiente del hablante de una serie de dígitos hablados continuamente y restringidos al ancho de banda del teléfono puede lograr una taza de error de 0.3% cuando la longitud de la serie de dígitos es conocida. Más recientemente, los investigadores han empezado a enfocarse en el reconocimiento de voz generada espontáneamente; por ejemplo, en el dominio del Servicio de Información de Viajes Aéreos (Air Travel Information Service, ATIS), una tasa de error de palabra menor al 3% ha sido reportada para un vocabulario de cerca de 2,000 palabras y un modelo de lenguaje con una perplejidad alrededor de 15.

Tareas de alta perplejidad con un vocabulario de miles de palabras son proyectadas primordialmente para aplicaciones de dictado. Después del trabajo durante muchos años en sistemas de palabras aisladas y dependientes del hablante, la comunidad se ha movido desde 1992 hacia sistemas de reconocimiento de voz continua, independiente del hablante, de alta perplejidad (PP 200) y con un vocabulario muy amplio (más de 20,000 palabras). El mejor sistema en 1994 logró una tasa de error de 7.2% de frases leídas de un periódico de negocios de Norte América [Pal94].

Con las constantes mejoras en el rendimiento de sistemas de reconocimiento de voz, están ahora siendo desplegados sistemas de reconocimiento dentro de teléfonos y redes celulares en muchos países. En marcación de voz, por ejemplo, los usuarios pueden marcar de diez a veinte números telefónicos mediante la voz (p.e. *call home*) después de haber registrado sus voces diciendo las palabras asociadas con números de teléfono. AT&T, por otra parte, ha instalado un sistema de ruteamiento de llamada usando una tecnología independiente del hablante que puede detectar algunas frases (p.e. persona a persona, tarjeta de llamada) en frases tales como: *I want to charge it to my calling card*.

En el presente, varios sistemas de dictado de vocabulario muy amplio están disponibles para la generación de documentos. Estos sistemas generalmente requieren a los hablantes pausas entre palabras. Sus rendimientos pueden ser aumentados si uno puede aplicarlos en dominios específicos tales como dictados de reportes médicos.

A pesar de que se ha hecho mucho progreso, las máquinas están lejos de reconocer voz del tipo de una conversación. Las tasas de reconocimiento de palabras en conversaciones telefónicas son alrededor del 50% [Coh94]. Aún pasarán varios años para que un sistema de dictado continuo, independiente del hablante y de vocabulario ilimitado sea realizado.

Las áreas de investigación sobre reconocimiento de voz se centran en:

- **Robustez.** De particular atención diferencias en características del canal y ruido ambiental.
- **Portabilidad.** Es decir, rapidez en el diseño, desarrollo y penetración de sistemas de reconocimiento para nuevas aplicaciones. En el presente, los sistemas tienden a sufrir degradaciones significativas cuando son movidos a una nueva tarea. Para regresar a un

máximo desempeño, los sistemas deben ser entrenados con ejemplos específicos de la nueva tarea, lo cual es consumo de tiempo y resulta costoso.

- **Adaptación.** Es decir, cómo pueden adaptarse continuamente los sistemas ante condiciones cambiantes. Por ejemplo, nuevos hablantes, micrófono, tareas. Tal adaptación puede ocurrir en muchos niveles del sistema, como modelo de subpalabras, pronunciación de palabras o modelos de lenguajes, por mencionar unos cuantos.
- **Modelado del lenguaje.** Los sistemas actuales usan modelos de lenguaje estadísticos para ayudar a reducir el espacio de búsqueda y resolver ambigüedades acústicas. Como el tamaño del vocabulario crece y otras restricciones están relajados para crear sistemas más habitables, será incrementalmente importante obtener tanta restricción como sea posible de modelos de lenguaje; quizás incorporando restricciones sintéticas y semánticas que no pueden ser capturadas por modelos puramente estadísticos.
- **Medición confiable.** Muchos sistemas de reconocimiento de voz asignan puntajes a hipótesis. Estos puntajes no dan una buena indicación de si una hipótesis es correcta o no, sólo que ésta es mejor que la otra hipótesis. Como se mueve en tareas que requieren acciones, se necesitan mejores métodos para evaluar la absoluta certeza de la hipótesis.
- **Palabras fuera de vocabulario.** Los sistemas están diseñados para usarse con un juego particular de palabras, pero los usuarios pudieran no saber exactamente qué palabras están en el vocabulario del sistema. Esto lleva a un cierto porcentaje de palabras fuera de vocabulario en condiciones naturales. Los sistemas deben tener algún método de detección de palabras fuera de vocabulario, o éstos terminarán mapeando una palabra desde el vocabulario sobre la palabra desconocida, causando un error.
- **Voz espontánea.** Los fenómenos en voz espontánea son falsos comienzos, dudas, construcciones no gramáticas y otros comportamientos no encontrados en la voz de lectura. Se han hecho desarrollos que han llevado a progresos en esta área, pero aún hay mucho trabajo por hacer.
- **Prosodia.** Se refiere a la estructura acústica que se extiende sobre varios segmentos de palabras. La acentuación, la entonación y el ritmo transportan información importante para reconocimiento de palabras y la intención del usuario (p.e. sarcasmo, ironía). Los sistemas actuales no capturan la estructura prosódica. Aún no ha sido respondido cómo integrar la información prosódica dentro de una arquitectura de reconocimiento, es una pregunta crítica.

La codificación por predicción lineal (LPC) es una de las técnicas más poderosas en el análisis de voz ya que brinda una representación precisa y económica de parámetros relevantes de la señal de voz que permiten reducir las tasas de transmisión en la codificación de dicha señal; además incrementa la precisión y reduce los cálculos en el reconocimiento de voz, así como genera sistemas eficientes de síntesis de voz [Osh00]. La técnica de entrenamiento por cuantización vectorial (VQ) es una de las más empleadas en el reconocimiento de voz.

También se han logrado grandes progresos debido al advenimiento de los modelos ocultos de Markov (HMM), que son una herramienta muy poderosa para modelar la estructura temporal y variaciones de la voz; HMM es una aproximación probabilística de concordancia de patrones el cual modela una secuencia de tiempo de patrones de voz como la salida de un proceso aleatorio o estocástico. El modelado estocástico de la señal de voz soluciona el problema que presentaba la

técnica de alineamiento de plantillas, proporcionando los mejores resultados hasta la fecha tanto para el reconocimiento del habla, tanto aislada como continua, y para independencia del locutor. Las redes neuronales han sido también usadas para integrarlas dentro de arquitecturas de sistemas basados en el HMM, esto ha llegado a ser conocido como sistemas híbridos [Zue90], [Fan95].

4.1 Análisis de predicción lineal

El análisis LPC es la técnica más común para la codificación de voz de baja tasa de bits y es una importante herramienta en el análisis de voz. Su popularidad se deriva de su represtación compacta y precisa de la magnitud espectral de la voz. El análisis LPC es empleado en la señal de voz para estimar: la frecuencia fundamental, funciones de área del tracto vocal, las frecuencias y anchos de banda de polos y ceros espectrales (p.e. formantes); sin embargo, fundamentalmente da un conjunto pequeño de parámetros de voz que representan la configuración del tracto vocal. El análisis LPC estima cada muestra de voz basado en la combinación lineal de sus anteriores *p* muestras; una larga *p* permite un modelo más preciso. Los factores de peso (o coeficientes LPC) en la combinación lineal pueden ser directamente usados en filtros digitales como coeficientes multiplicadores para síntesis o pueden almacenarse como plantillas en reconocedores de voz. Los coeficientes LPC pueden transformarse en otros conjuntos de parámetros para una codificación más eficiente. En esta sección se examinará cómo calcular los parámetros.

Cabe mencionar que el análisis LPC tienen inconvenientes: minimiza la complejidad del análisis, la señal de voz se asume usualmente proveniente de una fuente todo polo; es decir, que su espectro no tiene ceros. Dado que la voz real tiene ceros debido a usual fuente de excitación glotal y también a múltiples trayectorias acústicas en las nasales y sonidos sordos, tal modelo es una simplificación, la cual sin embargo no causa mayores dificultades en muchas aplicaciones. No obstante lo anterior, se han realizado esfuerzos para modificar el sistema LPC todo polo para modelar ceros también.

Como se ha dicho, la idea básica del modelo LPC es que dada una muestra de voz en el tiempo discreto n, s(n), ésta puede ser aproximada como una combinación lineal de las p muestras de voz pasadas, esto es:

$$s(n) \approx a_1 s(n-1) + a_2 s(n-2) + \cdots + a_p s(n-p) \qquad (4.1)$$

donde los coeficientes a_1, a_2, … , a_p son considerados constantes en la trama de análisis. Para convertir la ecuación (4.1) en una equivalencia se agrega un término de excitación, Gu(n), dando:

$$s(n) = \sum_{k=1}^{p} a_k s(n-k) + Gu(n), \qquad (4.2)$$

donde u(n) es una excitación normalizada y G es la ganancia de la excitación. Si expresamos a (4.2) en el dominio de z, se tiene:

$$S(z) = \sum_{k=1}^{p} a_k z^{-k} S(z) + GU(z), \tag{4.3}$$

que conduce a la función de transferencia:

$$H(z) = \frac{S(z)}{GU(z)} = \frac{1}{1 - \sum_{k=1}^{p} a_k z^{-k}} = \frac{1}{A(z)}, \tag{4.4}$$

La interpretación de la ecuación (4.4) se muestra en la figura 4.2, muestra una fuente de excitación normalizada, u(n), que es escalada por una ganancia G, y actúa como entrada a un sistema todo-polos, $H(z) = \frac{1}{A(z)}$ para producir una señal de voz, s(n). Basado en el conocimiento de que la función de excitación para la voz es esencialmente: un tren de impulsos cuasi-periódicos (para sonidos de voz sonoros) o una fuente de ruido aleatorio (para sonidos no sonoros o sordos).

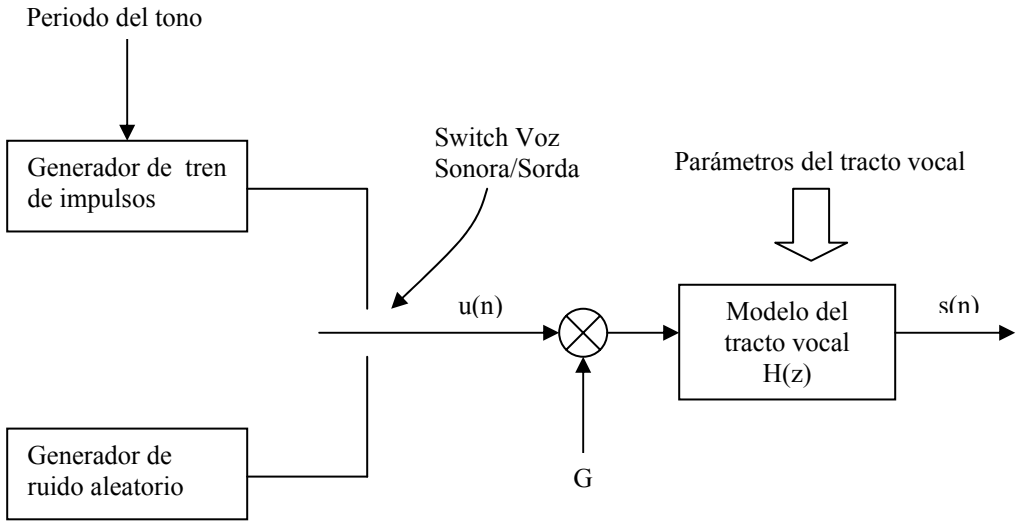


Figura 4.2: Modelo del tracto vocal basado en LPC

4.1.1 Ecuaciones del análisis LPC

Basado en el modelo de la figura 4.2, que representa la ecuación

$$s(n) = \sum_{k=1}^{p} a_k s(n-k) + Gu(n), \tag{4.5}$$

se considera la combinación lineal de las muestras pasadas de voz como una estimación $\tilde{s}(n)$, definida como:

$$\tilde{s}(n) = \sum_{k=1}^{p} a_k s(n-k), \tag{4.6}$$

El principal problema del análisis de predicción lineal es determinar el conjunto de coeficientes del predictor $\{a_k\}$, directamente de la señal de voz, de tal forma que las propiedades espectrales del filtro de la figura 1 se igualen, lo más posible, a la señal de voz dentro de una ventana.

Para obtener los coeficientes de predicción (a_k's) son determinados (calculados), minimizando la suma de diferencias cuadradas sobre un intervalo finito (error de predicción promedio), entre las muestras actuales de voz y las predecidas linealmente, esto es:

$$E_n = \sum_m e_n^2(m) = \sum_m (s_n(m) - \tilde{s}_n(m))^2 = \sum_m \left[s_n(m) - \sum_{k=1}^{P} a_k s_n(m-k) \right]^2 \tag{4.7}$$

Donde, E_n es el error de predicción promedio en tiempo corto, $e_n(m)$ es error de predicción, $s_n(m)$ es la señal actual multiplicada por una ventana y $\tilde{s}_n(m)$ es la muestra predecida. Con fines de simplificar la notación, se ha usado $s_n(m)$ para la señal en tiempo corto, determinada por:

$$s_n(m) = \begin{cases} s(m+n)w(m), & 0 \le m \le N-1 \\ 0, & \textit{para otro caso} \end{cases} \tag{4.8}$$

donde w(m) es una ventana de longitud N.

Por otro lado, para encontrar los valores de a_k que minimizan a E_n en (4.7), se calcula:

$$\frac{\partial E_n}{\partial a_i} = 0, \quad \textit{para} \quad i = 1,2,...,p \tag{4.9}$$

La solución de (4.9) se puede obtener por varios métodos, uno de los más usados es el **método de la autocorrelación**. La solución de este método da como resultado un sistema de ecuaciones lineales con p incógnitas y se pueden expresar de forma matricial de la forma:

$$\begin{bmatrix} R_n(0) & R_n(1) & . & . & R_n(p-1) \\ R_n(1) & R_n(0) & . & . & R_n(p-2) \\ . & . & . & . & . \\ . & . & . & . & . \\ R_n(p-1) & R_n(p-2) & . & . & R_n(0) \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ . \\ . \\ \alpha_p \end{bmatrix} = \begin{bmatrix} R_n(1) \\ R_n(2) \\ . \\ . \\ R_n(p) \end{bmatrix}$$

donde $R_n(\sigma)$ es la autocorrelación de la señal en tiempo corto, y a_k son los coeficientes de predicción lineal (LPC's) que resuelven el sistema.

La matriz p x p con los valores de autocorrelación es una matriz Toeplitz (es simétrica con todos los elementos de la diagonal iguales) y por lo tanto puede resolverse eficientemente mediante el empleo de procedimientos bien conocidos, como el algoritmo de Levinson-Durbin.

4.1.2 Algoritmo de Levinson-Durbin [Rab93]

Inicialización

$E^{(0)} = r(0)$

para $i = 1,2,..., p$

$$k_i = \frac{r(i) - \sum_{j=1}^{i-1} \alpha_j^{(i-1)} r(|i-j|)}{E^{(i-1)}}$$

$\alpha_i^{(i)} = k_i$

para $j = 1,..., i-1$

$\alpha_j^{(i)} = \alpha_j^{(i-1)} - k_i \alpha_{i-j}^{(i-1)}$

$E(i) = (1 - k_i^2) E^{(i-1)}$

La solución final está dada por:

$a_m = coeficientes\ LPC = \alpha_m^{(p)}$

El procedimiento para generar los coeficientes LPC's de tramas de la señal de voz s(n), es el siguiente:

- Aplicar el filtro de pre-énfasis a la señal s(n).
- Segmentación de la señal s(n) en tramas de N muestras.
- Ventaneo de cada trama de la señal s(n) (usualmente, ventana de Hamming).
- Cálculo de la función de autocorrelación para cada trama ventaneada. La función de autocorrelación para una trama de N muestras es:

$$r(k) = \sum_{m=0}^{N-1-k} s(m)s(m+k) \qquad para\ k = 0,1,2,..., p\ , \tag{4.10}$$

- Obtención de los vectores de coeficientes LPC's resolviendo el sistema de ecuaciones simultáneas ó empleando los procedimiento óptimos.

4.2 Cuantización Vectorial

La cuantización vectorial (VQ), resulta una generalización de la cuantización escalar, pero ahora aplicada a todo un vector. El cambio de una a varias dimensiones, trae consigo un gran número de ideas, conceptos, técnicas y aplicaciones nuevas. Mientras la cuantización escalar se utiliza, principalmente en la conversión analógico/digital, la cuantización vectorial se enfrenta a las sofisticadas técnicas del procesamiento digital de señales. En la mayoría de los casos, las características más relevantes de las señales de entrada tiene representación digital, por eso, la cuantización vectorial se utiliza usualmente en la compresión de datos. Sin embargo, existen ciertos paralelismos entre ambas cuantizaciones, lo que permite la utilización de varios métodos, en la cuantización vectorial, como una generalización [Ger97].

Un vector se puede utilizar para describir prácticamente cualquier tipo de patrón, como puede ser un segmento de una señal de voz o de una imagen, simplemente al formar un vector con las muestras de la señal de voz o de la imagen. La cuantización vectorial puede aplicarse al reconocimiento de patrones, ya que un patrón de entrada es comparado y aproximado a alguno de los patrones de referencia almacenados. El reconocimiento permite encontrar el patrón de referencia que más se acopla al patrón de entrada. Por lo tanto, la cuantización vectorial es más que una generalización de la cuantización escalar. En fechas recientes, se ha convertido en la principal herramienta del reconocimiento de voz, además de que se sigue utilizándose en la compresión de señales de voz e imágenes [Ger97].

4.2.1 Distancias y medidas de distorsión

Un componente clave en la mayoría de los algoritmos de comparación de patrones, es formular una medida de distorsión entre dos vectores característicos. Esta medida de distorsión, puede ser manejada con rigor matemático si los patrones son visualizados en un espacio vectorial.

Suponer que se tienen dos vectores característicos, **x** e **y**, definidos en un espacio vectorial χ. Se define una métrica o función de distancia, d, en el espacio vectorial χ, como una función de valor real, sobre el producto cartesiano χ x χ, que cumpla las siguientes condiciones:

1. $0 \le d(\boldsymbol{x}, \boldsymbol{y}) < \infty$, para $\boldsymbol{x}, \boldsymbol{y} \in \chi$ y $d(\boldsymbol{x},\boldsymbol{y}) = 0$ sí y solo sí $\boldsymbol{x} = \boldsymbol{y}$;
2. $d(\boldsymbol{x}, \boldsymbol{y}) = d(\boldsymbol{y}, \boldsymbol{x})$ para $\boldsymbol{x}, \boldsymbol{y} \in \chi$;
3. $d(\boldsymbol{x}, \boldsymbol{y}) \le d(\boldsymbol{x}, \boldsymbol{z}) + d(\boldsymbol{z}, \boldsymbol{y})$ para $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in \chi$.

Además, una función de distancia se denomina invariante sí

4. $d(\boldsymbol{x} + \boldsymbol{z}, \boldsymbol{y} + \boldsymbol{z}) = d(\boldsymbol{x},\boldsymbol{y})$

Las primeras tres propiedades comúnmente son conocidas como positividad (no negatividad), simetría y desigualdad del triángulo, respectivamente. Una métrica que contenga

estas propiedades, permite un alto grado de manejo matemático. Sí una medida de distancia, d, satisface solo la propiedad de positividad, se le denomina medida de distorsión, particularmente cuando los vectores son representaciones del espectro de la señal.

Para el procesamiento de voz es importante considerar que la definición (o elección), de la medida de distancia, es significativamente subjetiva. Una medida matemática de la distancia, para ser utilizada en el procesamiento de voz, debe tener una alta correlación entre su valor numérico y su distancia subjetiva aproximada, para evaluar una señal real de voz. Para el reconocimiento de voz, la consistencia psicofísica (los diferentes matices que se le pueden imprimir a una misma palabra o frase), que se desea medir con la distancia, obliga a que se encuentre una medida matemática ajustada por necesidad, a las características lingüísticas conocidas. Estos requisitos tan subjetivos no pueden ser satisfechos con medidas de distancia que proporcionen manejo matemático. Un ejemplo es la tradicional medida del Error Cuadrático, $d(\mathbf{x}, \mathbf{y}) = (\mathbf{x} - \mathbf{y})^2$.

Dado que existe una enorme dificultad al querer cumplir simultáneamente ambos objetivos (subjetividad y manejo matemático), algún compromiso es inevitable. Por consiguiente, y dado que se necesita manipular matemáticamente las propiedades de esa medida de distancia, se necesita probar que estas propiedades subjetivas son lo suficientemente buenas como para lograr el reconocimiento de voz. Por otra parte se hablará de "medidas de distorsión" en vez de "métricas" debido a que se relajan las condiciones de simetría y desigualdad del triángulo. No se debe utilizar el término distancia en sentido estricto, acorde a la definición de arriba; por otro lado, se debe mantener la costumbre de la literatura de voz, donde el término distancia es análogo, a las medidas de distorsión [Rab93].

Existen varios tipos de medidas de distorsión, cada una con sus características especiales. Entre ellas tenemos: Distancia Euclidiana Cuadrática, Distorsión del Error Cuadrático Medio, Distorsión del Error Cuadrático Ponderado, Distancia de Itakura, etc..

Distancia Euclidiana Cuadrática. La medida más conveniente y ampliamente usada para calcular distancias, es el Error Cuadrático o Distancia Euclidiana Cuadrática, entre dos vectores, definida como:

$$d(X_{1,}, X_2) = \left\| X_1 - X_2 \right\|^2 = \sum_{j=1}^{N} \left(X_{1j} - X_{2j} \right)^2 \qquad (4.11)$$

Distorsión del Error Cuadrático Medio. La distorsión del Error Cuadrático Medio (MSE) es otra de las medidas mas utilizadas y se define como:

$$\mathrm{d}(\mathrm{X}_1, \mathrm{X}_2) = \frac{1}{\mathrm{N}} (\mathrm{X}_1 - \mathrm{X}_2)^{\mathrm{T}} (\mathrm{X}_1 - \mathrm{X}_2) = \frac{1}{\mathrm{N}} \sum_{\mathrm{j}=1}^{\mathrm{N}} (\mathrm{X}_{1\mathrm{j}} - \mathrm{X}_{2\mathrm{j}})^2 \qquad (4.12)$$

en la cual la distorsión está definida por cada dimensión. La popularidad del MSE se basa en su simplicidad y seguimiento matemático.

Distorsión del Error Cuadrático Medio Ponderado. Otra medida de distorsión es el Error Cuadrático Medio Ponderado. En el MSE la medida asume que las distorsiones contribuyen cuantizando los diferentes parámetros $\{X_{1j}\}$ de igual forma. Y de manera general, se pueden introducir pesos diferentes con el fin de aportar ciertas contribuciones a la distorsión, dependiendo del parámetro. El MSE ponderado general se define como:

$$d(X_1, X_2) = (X_1 - X_2)^T W(X_1 - X_2) \qquad (4.13)$$

Donde W es una matriz de ponderación definida, simétrica y positiva, y los vectores X_1 y X_2 son tratados como vectores columna.

Cada una de las medidas de distorsión mencionadas anteriormente, resultan simétricas en sus argumentos X_1 y X_2 y pueden ser aplicadas a las características derivadas del análisis de producción lineal de la voz; el uso de algunas presenta ciertas desventajas, tal es el caso de la distancia euclidiana que aunque resulte fácil de calcular, no todas sus características tienen el mismo significado perceptible.

Por lo anterior, en ciertos casos resulta conveniente y efectivo escoger una matriz de ponderación $W(X_1)$ que dependa explícitamente del vector X_1, para así obtener una medida de distorsión perceptiblemente motivada. En este caso, la distorsión:

$$d(X_1, X_2) = (X_1 - X_2)^T W(X_1)(X_1 - X_2) \qquad (4.14)$$

es asimétrica.

Distancia de Itakura. En muchos casos del procesamiento de voz, es necesario tener otra medida de la distancia que existe entre dos vectores LPC. La distancia Euclidiana no es apropiada para medir los parámetros de dos LPC's individuales, en vectores que estén relacionados. Esto es debido a que los vectores LPC dependen del peso de la matriz de autocorrelación correspondiente a cada LPC.

La medida de distancia más comúnmente utilizada para este propósito es la propuesta por Itakura. Esta distancia se deriva utilizando una interpretación intuitiva del rango de predicción en el error de la energía. Fue obtenida originalmente, utilizando la máxima probabilidad existente entre argumentos similares. La distancia de Itakura es, probablemente, la medida de distorsión más empleada para encontrar la similitud entre dos vectores LPC [Del87].

Esta distancia se define de la siguiente forma: sean R_yx y R_yy las matrices de autocorrelación multiplicadas por las señales de voz de entrada y de comparación, respectivamente. Sean así mismo, xR_yx^T la energía de salida del filtro inverso, tomado como referencia con la entrada, y yR_yy^T la energía mínima posible de salida, del filtro LPC, con respecto a la entrada de la voz. Entonces tenemos que la distancia de Itakura se obtiene mediante la ecuación 5:

$$d(x,y) = \log\left(\frac{xR_y x^T}{yR_y y^T}\right) \tag{4.15}$$

donde

$$yR_y y^T = \begin{bmatrix} -1 & a_1 & a_2 & \cdots & a_p \end{bmatrix} \begin{bmatrix} r(0) & r(1) & r(2) & \cdots & r(p) \\ r(1) & r(0) & r(1) & \cdots & r(p-1) \\ r(2) & r(1) & r(0) & \cdots & r(p-2) \\ \vdots & \vdots & \vdots & & \vdots \\ r(p) & r(p-1) & r(p-2) & \cdots & r(0) \end{bmatrix} \begin{bmatrix} -1 \\ a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix} \tag{4.16}$$

o lo que es lo mismo:

$$yR_y y^T = \sum_{i=0}^{P} a_i \sum_{n=0}^{P} r_n(|i-n|) a_n \tag{4.17}$$

donde $y = \begin{bmatrix} -1 & a_1 & a_2 & \cdots & a_p \end{bmatrix}$

De esta forma, si *y* es el vector aumentado de coeficientes LPC's de "referencia" o de "plantilla" $\begin{bmatrix} -1 & a_1 & a_2 & \cdots & a_p \end{bmatrix}$, y sea *x* el vector aumentado de coeficientes LPC's "desconocido" u "observado" $\begin{bmatrix} -1 & a'_1 & a'_2 & \cdots & a'_p \end{bmatrix}$ entonces:

xR_yx^T es la energía del filtro inverso formado con la señal de entrada de voz.
yR_yy^T es la energía de salida mínima posible para el filtro de predicción lineal con la entrada de voz.

Por tanto (4.15) también se puede escribir como:

$$d(x,y) = \log\left(\frac{E_x}{E_y}\right) \tag{4.18}$$

4.2.2 Desarrollo de la cuantización vectorial (VQ)

Partimos de un conjunto de vectores que pertenecen a un espacio K-dimensional, asumiendo que x es un vector perteneciente a ese conjunto, cuyos componentes [x_i, $1 \le i \le K$] son variables aleatorias reales y de amplitud continua. Un cuantizador vectorial Q, de dimensión K y tamaño N, es una transformación de un vector x, del espacio euclidiano de dimensión R^K, en un conjunto finito C que contiene N salidas o puntos de reproducción, llamados code vectors (vectores de código):

$$Q : R^K \rightarrow C, \quad C = \{\mathbf{y}_1,...,\mathbf{y}_N\} \quad \mathbf{y}_i \in R^K \quad \forall \quad i \in I \equiv \{0,1,...,N\} \qquad (4.19)$$

El conjunto C es llamado code book (libro de códigos) y tiene un tamaño N, esto significa que tiene N distintos elementos, cada uno de ellos dentro del espacio R^k.

Asociado a cada cuantizador vectorial de N puntos, existe una partición de R^k en N regiones o celdas, R_i para $i \in I$. La i-ésima celda esta definida por:

$$R_i = \{\mathbf{x} \in R^K : Q(\mathbf{x}) = \mathbf{y}_i\} \qquad (4.20)$$

algunas veces llamada imagen inversa o pre-imagen de y_i dentro del mapeo Q y denotada de forma más consistente por $R_i = Q^{-1}(y_i)$.

De la definición de celdas, tenemos que:

$$\bigcup_i R_i = R^K \quad \text{y} \quad R_i \bigcap R_j = \varnothing \quad \text{para} \quad i \neq j \qquad (4.21)$$

donde las celdas forman una partición de R^k.

La tarea de codificación de un cuantizador vectorial es, examinar cada vector de entrada x e identificar a que celda, k-dimensional del espacio R^k, pertenece. El codificador vectorial simplemente identifica el índice i de la región y el decodificador vectorial genera el vector del código y_i que representa a esta región [Ger97].

El conjunto y es conocido como diccionario de reconstrucción o simplemente diccionario, donde N es el tamaño del diccionario y $\{y_i\}$ es el conjunto de vectores del código. Los vectores y_i son conocidos también en la literatura de reconocimiento de patrones, como los patrones de referencia o plantillas. El tamaño N del diccionario, también se conoce como número de niveles, término proveniente de la cuantización escalar. De esta forma, se puede hablar de un diccionario de N niveles. Al proceso de creación del diccionario, también se le conoce como entrenamiento o población del diccionario

El modelo de operación de este codificador, se define de forma similar al caso escalar, la función de selección, $S_i(x)$, como indicador o función miembro $1_{Ri}(x)$ para la celda R_i de la partición, esto es:

$$S_i(\mathbf{x}) = \begin{cases} 1 & \text{si} \quad \mathbf{x} \in R_i \\ 0 & \text{c.o.c.} \end{cases} \qquad (4.22)$$

La operación de un cuantizador vectorial puede ser representada como:

$$Q(\mathbf{x}) = \sum_{i=1}^{N} \mathbf{y}_i S_i(\mathbf{x}) \qquad (4.23)$$

Una descomposición estructural (como la mostrada anteriormente), es particularmente evaluable para encontrar un algoritmo efectivo, durante la implementación de la cuantización vectorial [Ger97].

Agrupamiento

El agrupamiento es la forma en que se realiza la cuantización vectorial; consiste en lo siguiente: a partir de un conjunto de N muestras $\chi = \{X_1, X_2, X_3, ..., X_N\}$, se intentan separar en K subconjuntos disjuntos $\chi_1, \chi_2, \chi_3, ..., \chi_K$. En donde cada subconjunto representa a un grupo (clúster) y en el cual, las muestras pertenecientes tienen una mayor similitud entre sí, en comparación a las muestras de cualquier otro grupo.

Existen varios algoritmos de agrupamiento, entre los que tenemos: simple, distancia máxima, K-Medias, ISODATA y LBG; estos dos últimos son variantes del agrupamiento K-Medias, pero con mayor complejidad.

Definición del algoritmo de K-medias

Se describe a continuación el algoritmo de agrupamiento K-Medias. Su criterio de función es:

$$J_e = \sum_{j=1}^{K} \sum_{x\, en\, \chi_j} d(\mathbf{x}, \mathbf{z}_j) \qquad (4.24)$$

donde:

K = numero de grupos
z_j = centro del grupo (centroide) para el grupo j
χ_j = subconjunto de muestras asignadas al grupo j
$d(x,z_j)$ = Es la distancia de Itakura entre el ventor x y el centroide z_j

Algoritmo de K-medias:

1) Escoger K centroides iniciales $z_1(1), z_2(1), ..., z_k(1)$.

2) En la iteración *l*, asignar las muestras a los grupos:
   Asignar:

$$\mathbf{x} \quad a \quad \chi_i(l) \quad si \quad d(\mathbf{x}, \mathbf{z}_i(l)) \quad \leq \quad d(\mathbf{x}, \mathbf{z}_j(l)) \quad j = 1,2,...,K \quad j \neq i$$

donde d(x, $z_j(l)$) es la distancia (o distorsión) de Itakura

3) Calcular los nuevos centros de grupo:

$$\mathbf{z}_i(l+1) = \frac{1}{N_i} \sum_{x \in \chi_i(l)} \mathbf{x} \qquad i = 1,2,...,K$$

donde N_i es el número de muestras asignadas a $\chi_j(l)$.

4) Si $z_i(l+1) = z_i(l)$ para i = 1, 2, ..., K, el algoritmo ha convergido y debe terminarse. En caso contrario, regresar al paso 2.

Una característica de este algoritmo es que los centroides o cuantizadores obtenidos, no son los óptimos globales; es decir, se obtienen cuantizadores óptimos locales. Estos dependen de varios factores, como son: asignación inicial de centroides (principalmente), orden de la toma de muestras, propiedades geométricas de los datos, tipo de distorsión empleada, etc.. Una forma de poder alcanzar los óptimos globales, es probar con una gran variedad de centroides iniciales y seleccionar los cuantizadores finales que tengan la menor distorsión, con respecto a los vectores a los cuales representan. Solución poco factible porque existen un gran número de combinaciones para los cuantizadores iniciales. Otra forma es, elegir los centroides iniciales de forma aleatoria, para buscar una distribución homogénea [Del87].

5. NÁHUATL DE SAN MIGUEL TZINACAPAN Y NOTAS PRELIMINARES AL ESTUDIO FONÉTICO-ACÚSTICO

Actualmente existen 7,000 lenguas en el mundo; sin embargo, aproximadamente en 100 años se estima que perdurarán menos de la mitad de ellas. Hasta este momento, 48% de la población en el mundo habla una de las 10 siguientes lenguas: Chino estándar (mandarín), inglés, español, bengalí, hindi, portugués, ruso, árabe, japonés y alemán. Por otro lado, resulta de gran contraste que 52% de todas las lenguas son habladas aproximadamente por 10,000 personas [Lad05-2].

Las lenguas usualmente comparten fonemas tales como las vocales cardinales /i, a, u/ y /p, t, k/, pero las formas del tracto vocal son usualmente ligeramente diferentes para estos sonidos en diferentes lenguas [Osh00]. Existen casos extremos como en el idioma ubykh caucásico con su sistema fonético compuesto de dos o tres vocales y casi 80 consonantes [Jak87].

Las lenguas usualmente difieren significativamente en el juego de vocales empleadas. Muchas tienen solamente unas pocas vocales; 25% de las lenguas del mundo (como el español y el japonés) tienen 5 vocales o inclusive menos [Osh00]. Muchas lenguas contrastan la duración de las vocales. Por ejemplo, el estonio tiene tres duraciones en sus vocales [Osh00].

En diversas lenguas, la nasalización de las vocales no constituye fonemas (pero sí alófonos); tal es el caso del español y del náhuatl. Sin embargo, lenguas como el francés poseen fonemas vocálicos basados en su nasalización.

Por otra parte, existen lenguas donde diferentes patrones del formante F0 (o frecuencia fundamental) brindan diferentes fonemas, a estas lenguas se les denomina tonales (como el mandarín, el tai, el sueco, el ebira –de Nigeria-). El náhuatl no es una lengua tonal pero sí existen lenguas indígenas en México con esta característica, así ocurre con aquellas del tronco otomangue (integradas por el zapoteco y mixteco, entre otros), la mayoría de estas lenguas son habladas en el estado de Oaxaca [Sil].

Por último, resulta elocuente la aseveración de Medina en [Med90] al respecto de la riqueza lingüística en México. “El conjunto de las lenguas amerindias que se hablan en México actualmente representa uno de los más preciosos legados de las culturas mesoamericanas y sintetiza una experiencia histórica de una complejidad tal que la investigación lingüística y antropológica apenas ha trabajo una pequeña parte de las numerosas e importantes cuestiones que plantea. El que no sepamos con precisión el número de lenguas que se hablan y no tengamos un registro sistemático de aquellas que se encuentran en proceso de extinción, constituyen manifestaciones evidentes de la magnitud de las dificultades presentadas por el acervo étnico-lingüístico. Son lenguas de una enorme complejidad que nos presentan como un reto para la teoría y la práctica de la lingüística”.

5.1 Presencia de la lengua náhuatl

Gabriela Cortés, en su artículo "los nahuas", [Cor90], presenta a este grupo tan importante de México. Dado que es un artículo muy ilustrativo, a continuación se condesa una parte de lo expuesto por la autora.

"El grupo étnico más extendido en México y aquel que tiene mayor población (más de un millón de habitantes) es el nahua o náhuatl. Su distribución a lo largo de México no es homogénea puesto que sólo se localizan en ciertas regiones, su rasgo principal de identidad es el hablar el idioma náhuatl.

Dado que es un grupo tan extendido se deben hacer algunas precisiones: 1) No todos los nahuas que ahora habitan en México son descendientes de los aztecas o mexicas. 2) No todos los nahuas tienen ni la misma cultura ni la misma lengua.

La segunda precisión obedece a cuatro razones: a) Aunque los mexicas dominaron a muchos grupos, nahuas o no nahuas, no les impusieron ni su cultura ni su lengua, por el contrario, más bien retomaron elementos culturales; b) los nahuas, fueran o no mexicas, se relacionaron comercial, cultural o militarmente con otros grupos. Esto dejó hondas huellas en su cultura, su religión y su lengua; c) después de la Conquista, algunos grupos no nahuas fueron nahuatizados, es decir, sustituyeron su lengua original por el náhuatl, manteniendo sus ragos culturales propios. Esto fue por políticas españolas para congregar varias comunidades en un solo pueblo y emplear el náhuatl como lengua oficial indígena; d) finalmente, entre los mismos nahuas hubo (y hay) diversidad lingüística."

Al respecto de que la lengua náhuatl cuenta con una distribución geográfica muy amplia, en su estudio etnocientífico de los colores en Cuetzalan, Castillo [Cas00] señala:

"Varios investigadores han mencionado que desde los tiempos prehispánicos la lengua náhuatl se extendió más allá del territorio nacional. Heath (...) señala que el náhuatl se convirtió en la lengua "oficial" antes de la llegada de los españoles. También Lastra (...) apunta que las lenguas de la familia yutoazteca se hablaban desde la meseta de la Gran Cuenca del Oeste de los Estados Unidos hasta algunas regiones de Nicaragua, y dice que la familia más importante de esta familia es el náhuatl, debido a la influencia y extensión que alcanzó el imperio azteca sobre el territorio mexicano. En un trabajo más reciente, Dakin (...) comenta que "la familia yutoazteca es la que cubre el área más amplia geográficamente en las Américas, sin tomar en cuenta clasificaciones de relaciones remotas. Desde el territorio del payute norteño, hablado en los estados norteamericanos de Nevada, Oregon e Idaho, este último en la frontera con Canadá, se extiende hasta la región de los nawates o pipiles, que hablan la variedad más sureña del náhuatl", tal región se localiza en El Salvador.

Dada su extensa distribución geográfica, el náhuatl es una lengua que presenta una rica variación dialectal tanto desde el punto de vista de los dialectos modernos como de su trayectoria evolutiva. Por dialecto se entiende una variante local de la lengua. Castillo cita a García de León

en [Cas00], "es la lengua más dispersa de México y Centroamérica, y las características socioculturales de sus hablantes tienen que ver con la región en donde estén y con la lengua, o lenguas, con que convivan, y esto ha aumentado la variación dialectal"

Los estudios sobre dialectología náhuatl realizados por Lastra [Las86] y Canger nos permiten conocer la agrupación de los dialectos modernos y ubicar las zonas donde se hablan. Lastra señala que "el criterio principal que se empleó para constituir las áreas fue que hubiera rasgos comunes en una zona continua y que éstos no se dieran en zonas adyacentes" [Las86]. Los rasgos a los que se refiere son fonológicos, gramaticales y léxicos. Esta investigadora además advierte que no es una clasificación satisfactoria, aunque es de gran ayuda para reconocer cada una de las variantes.

A continuación se presentan las clasificaciones, con sus áreas y subáreas, determinadas por Lastra y Canger, mencionadas en [Las86] y [Cas00]. Es de notar las similitudes entre ellas. Conforme a estas clasificaciones, el dialecto náhuatl estudiado en esta tesis corresponde al área de la Periferia Oriental, subárea: Sierra de Puebla. El dialecto es el de San Miguel Tzinacapan, una comunidad perteneciente al municipio de Cuetzalan, en la Sierra Norte del estado de Puebla, la cual forma parte de la Sierra Madre Oriental.

Lastra

Periferia Occidental
- Costa Occidental
- Occidente del Estado de México
- Durango-Nayarit

Periferia Oriental
- Sierra de Puebla
- Istmo
- Pipil

Huasteca

Centro
- Subárea Nuclear
- Puebla-Tlaxcala
- Xochiltepec-Huatlatlauca
- Sureste de Puebla
- Guerrero Central
- Sur de Guerrero

Canger

Área Central
- Valle de México
- Tlaxcala
- Morelos
- Guerrero Central
- Puebla Central
- Norte de Puebla
- La Huasteca

Periferia

Occidental
- Durango
- Nayarit
- Jalisco
- Colima
- Michoacán
- Almomoloa
- Norte de Guerrero

Oriental
- Sierra de Puebla
- Istmo
- Pipil

Área Nuclear
- Náhuatl Clásico
- San Martín de las Pirámides

Hablantes de lengua indígena en México y en la zona de Cuetzalan

De acuerdo con el XII censo de población y vivienda del INEGI [Ine] realizado en el año 2000, la población hablante de lengua indígena mayor de 5 años es de 6 044 547 personas (es decir, el 7.2% del total de la población en el país). De ella, el 83.1% es bilingüe y el 16.9% es monolingüe, siendo este último más alto en Chiapas, Guerrero y Oaxaca. Las entidades del país donde se concentra la mayor población hablante de lengua indígena son Oaxaca (18.5%), Chiapas (13.4%), Veracruz (10.5%), Puebla (9.4%) y Yucatán (9.1%). De las 66 lenguas captadas por el censo, las lenguas más numerosas son: náhuatl (24.0%), maya (13.2%), mixteco (7.2%), zapoteco (7.0%) y tzotzil (5.0%). Todos estos porcentajes se calculan con respecto al total de población hablante de lengua indígena.

Siguiendo los datos del XII censo, en el estado de Puebla el 73.7% de los hablantes de lengua indígenas hablan náhuatl. Puebla ocupa el primer lugar en número de hablantes de lengua náhuatl con 28.8%; le sigue Veracruz, 23.4%; después Hidalgo, 15.3%; San Luis Potosí, 9.6%, y Guerrero, 9.4%.

Los datos anteriores conviene considerarlos como números nos cerrados. Como indica Medina en su artículo sobre la etnografía de México [Med90], a pesar de que en nuestros días se tienen ideas generales sobre las características socioeconómicas de la población indígena, aún se desconocen muchas de sus particularidades culturales y lingüísticas. Tal es el caso de su magnitud demográfica, pues se carece de un dato preciso, Medina expresa que los datos censales presentan (en 1990) serias limitaciones; una razón, pero no la única, son las dificultades para cubrir una población extremadamente dispersa en selvas, montañas, pantanos y lejanos valles. Por otro lado, la filiación étnica no es lo mismo que filiación lingüística (un ejemplo son los pueblos que abandonan su lengua original y ahora son hispanohablantes), y es una clara señal de extinción de lenguas. Finalmente, a veces los datos censales pudieran registrar como lengua a una familia de lenguas (p.e. el zapoteco), o no consignar en la información censal lenguas habladas por poblaciones reducidas (p.e. el tuzanteco, motozintleco, mochó, chuj, jacalteco, kanjobal, todas ellas lenguas mayanses habladas en Chiapas). Medina concluye que los datos ofrecidos por los censos de población hay que verlos más como aproximaciones que como referencias seguras.

Ahora bien, de acuerdo con el XI Censo de Población y Vivienda de 1990, la Sierra Norte de Puebla contaba con una población total de 1 003 160 habitantes, de los cuales 457 849 eran indígenas: 362 991 nahuas o mexicanos, 86 793 totonacos, 7 688 otomíes y 377 tepehuas [Ine]. Estas lenguas pertenecen a tres familias lingüísticas distintas: la yutoazteca, la totonaca y la otopame. Efectivamente, los hablantes de náhuatl se encuentran distribuidos en todo el territorio poblano y su mayor concentración está en la Sierra Norte.

Volviendo a los datos del XII censo, y refiriéndose exclusivamente a la lengua náhuatl, los municipios que cuentan con más hablantes son Cuetzalan del Progreso (o simplemente Cuetzalan, como se le suele llamar), Zacapoaxtla, Huauchinango, Zautla, Tlatlauquitepec y Tlaola. De éstos, Cuetzalan ocupa el primer lugar con 27 785 hablantes, y equivale al 4.29% de los 217 municipios de todo el estado.

En particular, la Sierra Norte está conformada por 68 municipios y 1 430 comunidades. El municipio de Cuetzalan está conformado por ocho juntas auxiliares, según la organización social y política establecida por las autoridades municipales y comunitarias. El poblado de San Miguel Tzinacapan es una de esas juntas.

Las variantes del náhuatl de Cuetzalan presentan una fuerte vitalidad al interior de las comunidades, no obstante que existe un alto grado de bilingüismo en lengua indígena y español. El náhuatl cumple la función principal en las relaciones intraétnicas y el español se convierte en el vínculo interétnico con la población totonaca y mestiza. Las comunidades conservan con vitalidad la lengua náhuatl así como buena parte de sus costumbres y tradiciones antiguas.

Castillo [Cas00] señala que, a pesar del alto grado de bilingüismo en la región, también podemos encontrar un importante número de adultos monolingües en lengua indígena. Quienes conforman esta población son en su mayor parte mujeres y ancianos, las primeras debido, entre otras cosas, a que rara vez salen de la comunidad o tienen poco contacto con el exterior. Añade que, de acuerdo a los datos censales más recientes, el municipio de Cuetzalan es la región del país donde se concentra el mayor número de hablantes de lengua náhuatl.

5.2 Metodología

Como se ha mencionado, la lengua náhuatl posee una gran variedad de dialectos. Los trabajos realizados por Lastra [Las86] y Canger mencionados en la anterior sección constituyen fuentes de consulta esenciales para conocer las agrupaciones dialectales, ubicar las zonas donde se hablan y determinar la variante a estudiar, como en el presente trabajo.

Por otra parte, las diferencias dialectales pueden presentarse entre comunidades distantes a pocos kilómetros unas de otras. Horcasitas [Hor92] señala que los dialectos modernos "varían de región en región y frecuentemente de pueblo en pueblo". Dado lo anterior, para realizar el estudio fonético de un dialecto náhuatl es importante enfocarse en una sola comunidad.

Para este trabajo de tesis me he concentrado en el municipio de Cuetzalan por ser una zona donde se ha estudiado mucho la lengua náhuatl, de tal manera que es posible contar con valioso material de consulta. Sin embargo, como cabecera del municipio, Cuetzalan es una población donde confluyen hablantes de diferentes dialectos náhuatl, por lo que consideré más conveniente enfocarme en una comunidad más pequeña. Gracias a contactos desde la misma ciudad de Puebla, hallé en San Miguel Tzinacapan el lugar donde grabar, ubicado a 4 kilómetros de Cuetzalan.

Al realizar trabajo de campo resulta adecuado presentarse con la autoridad local (en mi caso, con el presidente de la junta auxiliar) para, acompañado de una carta por parte de la universidad, notificarle sobre la naturaleza del trabajo en cuestión.

Debido a las características de la base de datos para la tecnología de reconocimiento de voz, fue necesario hacer grabaciones de campo para este trabajo de investigación.

La metodología constó de:

- *Adquisición del sistema fonético del dialecto a analizar*. Esta información la consulté en [Cas00], [Bañ] y [Tou84], que son obras del náhuatl de la región de Cuetzalan y de Tzinacapan. Particularmente es Castillo, [Cas00], quien presenta el sistema fonético del náhuatl de la región de Cuetzalan. Consultas con habitantes de Tzinacapan y el examen de las grabaciones no detectaron la presencia o ausencia de otros fonemas (aunque hay un caso especial mencionado más adelante)
- *Diseño del cuestionario*. Éste consta de alrededor de 100 palabras. Como en la fase de elaboración del cuestionario carecía de nociones sobre lengua náhuatl, me apoyé en trabajos publicados por [Hor92], [Las86], [Lau92], [Bañ], [Tou84]. Integré algunas palabras del cuestionario de Lastra [Las86] y de [Bañ], de tal manera que también añadí palabras que permitieran conformar un vocabulario muy básico de la lengua. Las observaciones y sugerencias por parte de Leopoldo Valiñas fueron de enorme ayuda para estructurar un cuestionario coherente y evitar palabras ambiguas o no existentes. Finalmente, las observaciones a partir

del dialecto en específico por parte de los hablantes, en el trabajo de campo, dieron la forma definitiva al cuestionario.

- *Selección de hablantes*. Se consideraron tres aspectos:
  - Que fueran hablantes del dialecto de San Miguel Tzinacapan.
  - Que fueran nativos del lugar o habitaran en él desde hace muchos años, de tal manera que su grado de conocimiento de la lengua fuera reconocido por los otros miembros de la comunidad.
  - Hombres y mujeres, jóvenes y adultos. El rango de los hablantes va de los 18 a los 80 años. Se descartaron adolescentes y niños.
- *Equipo de grabación*. Se empleó:
  - MiniDisc Sony MZ-NH700, con frecuencia de muestreo de 44.1 kHz y resolución de 16 bits. Se grabó sin compresión, en formato WAV y con codificación PCM. Este equipo resultó el más adecuado por su sobresaliente calidad de grabación, portabilidad y accesible precio.
  - Micrófono Behringer modelo XM1800S. Dinámico, super cardioide.
- *Grabaciones*. Que constituyen la materia prima de este trabajo, se grabaron al menos 20 repeticiones de cada palabra del cuestionario (se usarán 20 repeticiones para el sistema de reconocimiento, pero se grabaron algunas repeticiones más por si se presentaba alguna mala pronunciación o algún ruido ambiental inesperado). Los hablantes fueron 7 hombres y 5 mujeres, aunque de una persona sólo se grabó una parte. El tiempo neto de grabación por persona -pues se hacían pausas para descansar, tomar agua o aclarar dudas- oscila entre 2 y 2.5 horas.

Como he mencionado, no tenía bases en lengua náhuatl, por lo que me pareció pertinente atravesar por una fase de sensibilización de esta lengua, para ello asistí a un seminario de lengua náhuatl enfocado exclusivamente en su producción oral, fue impartido por Leopoldo Valiñas en el IIA y fue de considerable valía, un aporte concreto es haber podido hacer aclaraciones con los hablantes sobre las palabras del cuestionario a grabar durante el trabajo de campo.

Aunque muy generales, es conveniente mencionar algunos rasgos fonológicos del náhuatl que fueron tomados en cuenta para la elaboración del cuestionario.

- El náhuatl es una lengua aglutinante (como el alemán), dominar esta característica involucra un alto grado de conocimiento de la lengua.
- El parentesco y las partes del cuerpo son siempre poseídos, conformando siempre una sola palabra el sustantivo y su posesión.
- No hay forma infinitiva de los verbos.

En el último punto, *grabaciones*, se justifica por sí misma la razón de realizar este trabajo de campo. A pesar de sus ricos acervos, ninguna institución contactada hasta aquel momento (Instituto Nacional Indigenista, Museo Nacional de Antropología, Instituto de Investigaciones Antropológicas de la UNAM) poseía una base de datos con esas características.

El hecho de haber grabado alrededor de 10 hablantes es debido a que, para describir un lenguaje, existe la necesidad de contar con una muestra representativa. Ladefoged señala que "idealmente uno quiere cerca de media docena de hablantes de cada sexo. Pudieran haber diferencias sistemáticas entre las voces del hombre y de la mujer" [Lad05-01]. Se revisan las pronunciaciones de cada hablante para extraer las tendencias generales de cada sonido lingüístico.

Además, Ladefoged [Lad05-01] sugiere realizar grabaciones de voz que incluyan todas las frecuencias hasta 11 kHz; de hecho un análisis por debajo de los 5 kHz es suficiente para estudiar las vocales y las consonantes sonoras. Desafortunadamente el equipo de grabación no está diseñado para variar la frecuencia de muestreo y así haber cubierto únicamente el rango frecuencial de los sonidos lingüísticos, el cual va de los 50 Hz a los 13 kHz.

En los apéndices se encuentran el cuestionario e información de las personas que accedieron a ser grabadas.

5.3 Los sonidos lingüísticos del dialecto de San Miguel Tzinacapan

A continuación se muestra el sistema fonológico del dialecto hablado en San Miguel Tzinacapan, está integrado por 14 fonemas consonánticos y 8 fonemas vocálicos, se incluyen ejemplos. La fuente es [Cas00]. Se ha añadido una columna más, "Grafía en español", únicamente para que el lector no familiarizado con la fonética pueda vincular estos sonidos con el español y tener más claro la pronunciación, aunque en varios casos ésta es aproximada.

Consonantes:

		Fonema	Grafía en español	Punto de articulación	Ejemplo	
sordas	oclusivas	/p/	p	bilabial	[pokti]	humo
		/t/	t	alveolar	[tiltik]	negro
		/k/	k, qu	velar	[kostik]	amarillo
		/k^w/	ku	oclusiva velar labializada	[k^wali]	bueno
	africadas	/¢/	ts	alveolar	[¢onti]	cabello
		/č/	ch	palatal	[čičiltik]	rojo
	fricativas	/s/	s	alveolar	[siwat]	mujer
		/š/	x	alveopalatal	[šošoktik]	verde
		/h/	j	fricativa glotal	[nehwa]	yo
sonoras	nasales	/m/	m	bilabial	[me¢ti]	luna
		/n/	n	alveolar	[nakat]	carne
	lateral	/l/	l	dental	[ta:l]	tierra
	semivocal	/w/	gü	velar	[weyi]	grande
		/y/	ll	palatal	[yolot]	corazón

Vocales:

	Fonema	Clasificación	Ejemplo	
cortas	/i/	anterior alta	[istak]	blanco
	/e/	anterior media	[esti]	sangre
	/a/	central baja	[amo]	no
	/o/	posterior media	[omit]	hueso
largas	/i:/	anterior alta	[i:šyolke]	loco
	/e:/	anterior media	[e:wat]	cáscara
	/a:/	central baja	[a:ltia]	bañar
	/o:/	posterior media	[o:lini]	doler

Note que la fricativa glotal /h/ no ha sido clasificada con sonora o sorda, esto será explicado en el estudio fonético acústico.

Adviértase la presencia del fonema /t/ en vez del /λ/ (por ejemplo, "at" y "atl") que predomina en otras regiones de México. Esto quiere decir que se emplea el término "nahuat" en San Miguel Tzinacapan. Sin embargo seguiremos empleando el término náhuatl por ser el más conocido, sabiendo siempre que estamos estudiando una variante en particular.

Algo característico del náhuatl es la presencia del saltillo. En náhuatl el saltillo es un elemento demarcativo, no fonémico. Se presenta comúnmente al final de palabra que termine en vocal; aunque puede ser variable y presentarse al final de sílaba. Una de las funciones del saltillo es separar palabras de una oración. En este trabajo el saltillo será representado por un apóstrofe: / ' /.

Aunque este trabajo está dedicado exclusivamente a la fonética, cabe mencionar que la escritura de esta lengua fluctúa; es decir, "la ortografía náhuatl nunca ha sido fijada realmente" [Lau92]. Por eso no será sorprendente hallar textos en náhuatl que emplean diferentes grafías.

5.3.1 Silabación del náhuatl

Dada la influencia del fenómeno silábico en los sonidos de una lengua, es conveniente explicar la silabación de la lengua náhuatl.

Tuggy, en [Sil1], menciona que "el náhuatl generalmente no tiene diptongos así que cuando dos vocales están escritas juntas, pertenecen a dos sílabas distintas".

Además Tuggy declara que, "generalmente, en el náhuatl se permiten sílabas con una sola vocal (V) opcionalmente precedida de una sola consonante (C) y también opcionalmente seguida de una sola consonante. Es decir, el patrón silábico es (C)V(C). De las combinaciones permitidas dentro de este patrón, la más común (y preferida) es CV, pero también hay muchas sílabas CVC, y algunas VC y V. No existen combinaciones de consonantes en una sílaba; es decir, en el náhuatl no hay patrones CCV ni VCC al principio ni al final de una sílaba. Sólo dentro de una palabra puede haber combinaciones CC, y eso sólo porque las dos consonantes pertenecen a diferentes sílabas, es decir cuando el patrón es (C)VC.CV(C). Nunca se presentan combinaciones de tres consonantes (CCC).

Decir que la estructura silábica CV es la preferida es equivalente a decir que las consonantes se colocan preferentemente al principio y no al final de la sílaba. Así, cuando hay una secuencia CVCVCV, lo normal es que se silabice CV.CV.CV y no CVC.VC.V o en alguna otra configuración posible.

El náhuatl generalmente no tiene diptongos, de tal manera que cuando dos vocales se escriben juntas, pertenecen a sílabas diferentes. Por ejemplo, kitlālia 'él lo pone', a pesar de que se escribe como la palabra española 'Italia', /i.tál.^{i}a/, no se pronuncia como ésta, puesto que tiene la estructura silábica ki.t^{l}ā.li.a, y se pronuncia como la palabra española 'valía'.

A veces hay evidencia de que la silabación respeta las fronteras morfémicas no teniendo en cuenta esta preferencia. P.ej.: la palabra inohwi 'su camino de ellos' puede silabizarse in.oh.wi VC.VC.CV (y no en la forma esperada V.CVC.CV) porque su estructura morfémica es in-ohwi (de.ellos-camino)"

Tuggy sugiere tener presente que estas reglas se consideran con respecto a fonemas y no necesariamente a letras del alfabeto ya que varios de los fonemas del náhuatl se representan ortográficamente con un par de letras.

5.4 Consideraciones para el estudio fonético acústico

Las relaciones entre fonemas y sus realizaciones acústicas forman las bases para muchas aplicaciones de voz (como reconocimiento y síntesis) así como una comprensión de la producción y percepción de la voz. Esta sección investiga la forma de onda y las propiedades espectrales de los sonidos del habla náhuatl de la variante de San Miguel Tzinacapan.

La fonética acústica considera la señal de voz como la salida del proceso de producción del habla, relacionando dicha señal con su entrada lingüística (por ejemplo, en una oración). Considera la diferenciación de sonidos en una base acústica (no articulatoria). Debido a que cada fonema puede ser articulado de diferentes maneras y por diferentes aparatos tracto-vocales, hay mucha variabilidad en las señales de voz para el mismo fonema.

Muchos rasgos acústico-fonéticos son más claros espectralmente que en el dominio del tiempo. Examinando la forma de onda y/o el espectrograma de cada sonido nos permite caracterizar sus cualidades acústicas y determinar cómo compararlos y distinguirlos de otros sonidos del habla. El análisis de formantes es especialmente útil para distinguir entre sonidos debido a que muchos fonemas tienen un rango específico de frecuencias de formantes, las cuales definen sus características acústicas.

Conviene hacer algunas observaciones sobre la medición de sonidos en las formas de onda. Ladefoged, en [Lad05-1], hace notar que medir las duraciones de segmentos en una señal de voz no es una labor directa. Hay decisiones difíciles de tomar en este respecto cuando se realiza un análisis acústico, pues a veces no están claramente marcados los principios y finales de los sonidos lingüísticos en sus formas de onda. La recomendación de Ladefoged es “ser consistente con los puntos de medición seleccionados, y reportar la duración de cada sonido de la misma manera”. En este trabajo se establecen puntos de medición en los cruces por cero de las formas de onda, se darán mayores sobre dónde establecerlos en las secciones del estudio fonético acústico. El uso conjunto de espectrogramas con las formas de onda es muy recomendable para estos propósitos.

En muchos casos los dos primeros formantes son suficientes para caracterizaciones acústicas, pero ocasionalmente el tercer formante también es útil para la descripción. A pesar de que identificaremos valores de formantes como representativos de fonemas en particular, debe tomarse en cuenta que estos valores son mostrados para indicar tendencias, no valores absolutos. Los valores de los formantes varían a través de los hablantes y dependen de muchas variables. Aún para un hablante dado, los formantes pueden varias de acuerdo a contextos fonéticos, manera y velocidad del habla. De hecho, debe recalcarse que no hay descripciones absolutamente rígidas de fonemas.

Se empleó el programa Praat [Pra] para el estudio de esta sección. Es un programa libre para análisis de voz creado por especialistas del Institute of Phonetic Sciences de la Universidad de Amsterdam, Holanda, y es ampliamente empleado por lingüistas.

Como se ha indicado líneas arriba, el rango de los sonidos lingüísticos va de los 50 Hz a los 13 kHz, los sonidos sonoros se estudian por debajo de los 5 kHz. Por ello, el rango frecuencial para desplegar los espectrogramas en este estudio acústico se ha considerado de 0 Hz a 5 kHz.

De acuerdo a Ladefoged [Lad05-1], las mejores imágenes para observar formantes son aquellas en las que el ancho de banda es lo suficientemente ancho para no mostrar los armónicos individualmente (como ocurre en los espectrogramas de banda angosta). Esto corresponde con la aseveración de Dutoit [Dut97] donde indica que los espectrogramas de bancha angosta son menos frecuentemente utilizados, siendo los de banda ancha una herramienta invaluable para fonetistas y acústicos.

Praat ofrece una función bastante útil en la visualización de espectrogramas, se trata del **rango dinámico**. Este rango se explica de la siguiente manera "todos los valores fuera del rango dinámico y por debajo del máximo serán dibujados en blanco. Los valores intermedios tendrán tonos de gris. De tal manera que si el pico más alto en el espectrograma tiene una altura de 30 dB/Hz, y el rango dinámico es de 50 dB (el valor estándar), entonces los valores por debajo de -20 dB/Hz serán dibujados en blanco, y los valores entre -20 dB/Hz y 30 dB/Hz serán dibujados en varios tonos de gris"[Pra]. En general, los espectrogramas mostrados en este trabajo constaron de rangos dinámicos entre 35 dB y 50 dB.

Es conveniente mencionar que el realce en los todos los espectrogramas con preénfasis es de 6dB /oct.

Al presentar las imágenes de esta sección se han revisado aquellas por cada hablante con el fin de tratar de obtener una muestra representativa de los sonidos lingüísticos del náhuatl de Tzinacapan.

Debe advertirse que no hay criterios absolutos que definan unívocamente a un fonema en todas las circunstancias fonéticas, incluso para algún caso particular. Es decir, las diferencias acústicas de un fonema en particular puede ser atribuido a variabilidad contextual; sin embargo, el mismo fonema puede mostrar variación aún en el mismo contexto. Evidentemente conviene realizar mediciones para analizar acústicamente los sonidos de una lengua, pero resulta igualmente necesario observar y describir sus tendencias acústicas aún cuando no los criterios que los definan no puedan ser absolutos al medirlos.

A pesar de que el corpus consta de palabras aisladas debido a que éste se empleará para el sistema de reconocimiento de comandos, se grabaron algunas frases. Las frases son pocas, pero se han tomado en cuenta para extraer de ellas información útil para este estudio.

6. PROPIEDADES ESTÁTICAS DE LOS SONIDOS DEL HABLA

6.1 Consonantes

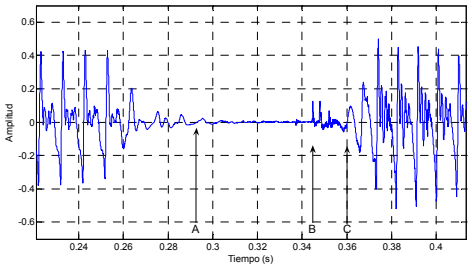
En esta sección agruparemos los fonemas de acuerdo al modo de articulación. Esto resulta útil porque los fonemas con esta agrupación exhiben muchas propiedades en común.

6.1.1 Oclusivas

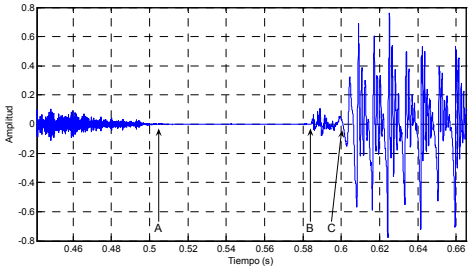
Como se ha visto en el cuadro fonético, las consonantes oclusivas del náhuatl de San Miguel Tzinacapan son /p/, /t/, /k/ y /k^w/; todas son sordas.

Las oclusivas pueden ser producidas en tres lugares del tracto vocal: en los labios (oclusiva bilabial /p/), en los alvéolos (oclusiva alveolar /t/) y en el velo (oclusiva velar /k/). Tenemos un caso muy particular en náhuatl: una oclusiva velar labializada /k^w/.

Un sonido oclusivo se produce cuando hay un cierre completo de los órganos articulatorios, deteniendo el flujo de aire en el tracto vocal, seguido de una liberación o explosión de aire. Esto distingue dos segmentos: cierre y explosión, los cuales son claramente identificados en la figura 6.1 donde se muestran segmentos de las realizaciones [tiotaki] (buenas tardes) e [istak] (blanco).



i. [ota]



ii. [sta]

Figura 6.1: Oclusivas alveolares sordas

El intervalo de cierre está mostrado entre los puntos A y B, la liberación ocurre entre B y C. En estos ejemplos, la oclusiva está seguida de una vocal, la señal después del punto C. Cuando una oclusiva está seguida de un sonido sonoro, p.e. una vocal, el tiempo en el punto C es a menudo denominado **comienzo sonoro** (**voice onset**) y la duración de la región entre B y C como **tiempo de comienzo sonoro (voice onset time**, **VOT**).

Ahora observemos los intervalos de cierre. En la región AB de la figura 6.1.i, la señal prácticamente no contiene energía, su duración es de alrededor de 40 ms. La energía anterior a la región AB corresponde a un segmento de la vocal que precede a la oclusiva. En la figura 6.1.ii podemos observar parte de un fono sordo anterior a la región AB; de hecho es una consonante fricativa, este fonema lo veremos más adelante. También notamos que ambos ejemplos tienen casi el mismo VOT, poco menos de 20 ms.

Cabe mencionar que -salvo para la fricativa, donde se determina A al extinguirse dicho fono- se establecieron los puntos A y C de la figura 1 al final del último pulso glótico del fono que antecede a la oclusiva y al comienzo del primer pulso glótico del fono que le sigue, respectivamente.

Como se ha comentado, no existen fonemas oclusivos sonoros en el náhuatl. Sin embargo, una diferencia importante entre oclusivas sordas y sonoras es que éstas últimas muestran un comportamiento cuasi-periódico durante la región de cierre, cosa que no ocurre con sus contrapartes sordas. “La vibración periódica en el cierre de una oclusiva sorda es debido al hecho de que las cuerdas vocales continúan vibrando aún cuando el tracto vocal está completamente cerrado; esta vibración es transmitida a través del cuello y mejillas. Teóricamente, por lo tanto, un examen de la forma de onda del intervalo de cierre de una oclusiva debe revelar si es sorda o sonora. Sin embargo, no siempre se cumple. En ciertas circunstancias el intervalo de cierre de una oclusiva sonora puede parecerse al de una oclusiva sorda y viceversa” [Oli93].

Existe un fenómeno que generalmente se presenta en el fonema /k/, se trata de una doble explosión. La figura 6.2 muestra el inicio de la realización [kali'] (casa), este extracto contiene la oclusiva y parte de la vocal que le sigue, la doble explosión se señala con flechas. Oliver, Greenwood y Coleman advirtieron este fenómeno, indicando que “la sección de explosión de la velar sorda a veces muestra una doble explosión; debido a que la doble explosión no es vista en ninguna otra oclusiva, su presencia indica una /k/” [Oli93].

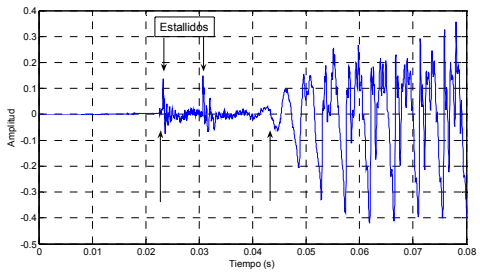
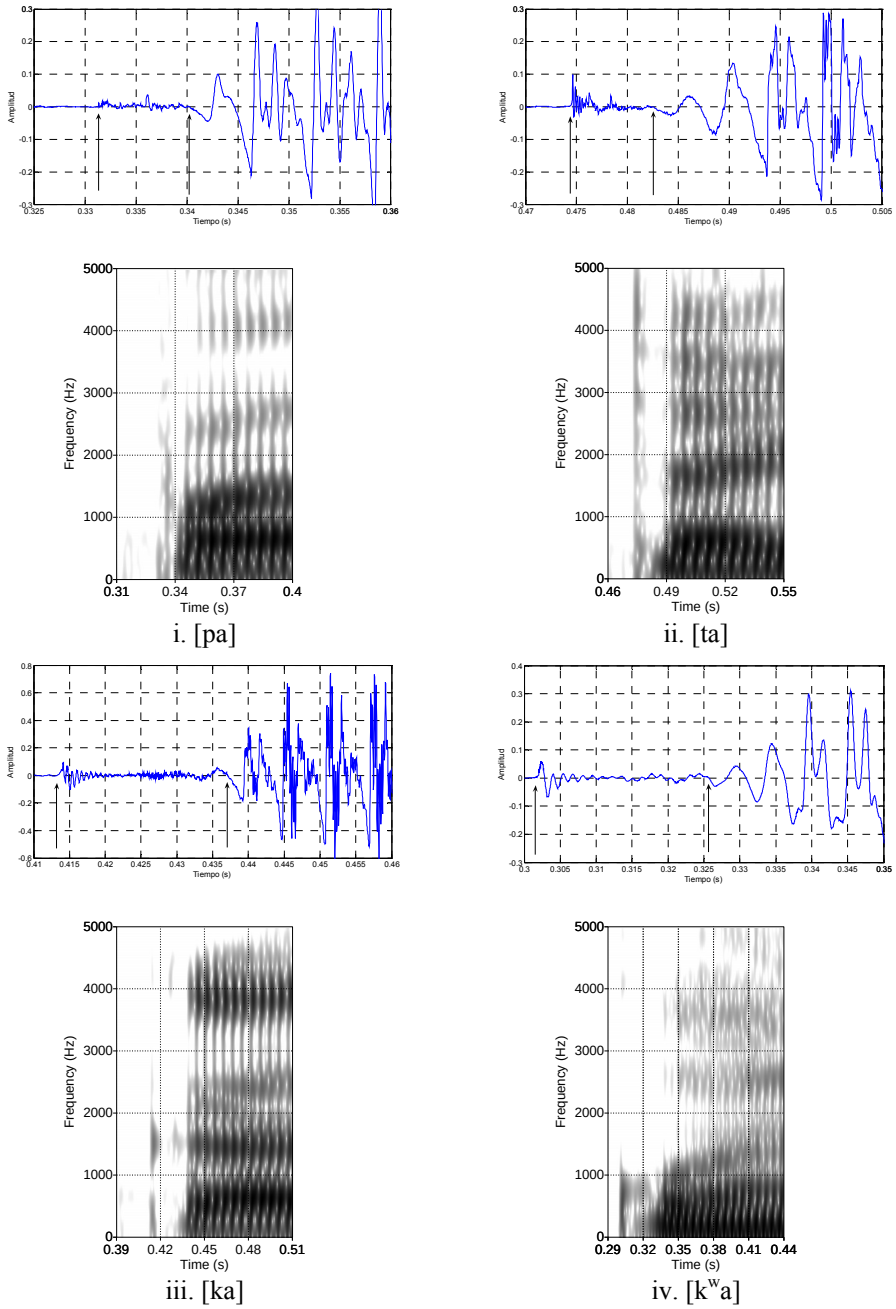


Figura 6.2: Forma de onda de [k] en inicio de palabra, note la doble explosión de la oclusiva velar sorda

En la figura 6.3 se presentan las formas de onda y espectrogramas de los sonidos oclusivos del náhuatl, son segmentos de las realizaciones [takat] (hombre), [tapalol] (comida), [tiotaki] (buenas tardes) y [tak^{w}al] (comida, en una segunda variante). Se muestra una parte de silencio y el estallido de cada oclusiva, así como el inicio del fono siguiente.



i. [pa] ii. [ta]

iii. [ka] iv. [k^{w}a]

Figura 6.3: Realizaciones de las oclusivas /p/, /t/, /k/, /k^{w}/

Los espectrogramas muestran algunas propiedades en común. El cierre se caracteriza por su baja energía; el comienzo del estallido es claramente visible debido al contraste entre el área clara del cierre y la línea vertical oscura que indica el estallido.

A pesar de estas propiedades en común, hallamos diferencias. La primera de ellas se refleja en el VOT; la oclusiva velar tiene un VOT de 24 ms, el de la alveolar es de 9 ms, el de la bilabial es de 8 ms, el de la velar labializada es de 24 ms. Se ha observado una tendencia en que el VOT de la alveolar es ligeramente mayor que el de la bilabial; sin embargo es contrastante el VOT velar, el cual es mucho más largo que el de las otras oclusivas. Otra diferencia está en la fuerza de cada estallido, en los correspondientes a /t/ y /k/ hay una considerable intensidad en la señal, que indica mayor energía que en /p/. Una tendencia que se observa en la forma de onda es que la energía de la explosión de la alveolar /t/ es la más alta de las tres. Una tercera diferencia se observa en la variación de la forma de onda durante el estallido, lo que refleja su contenido frecuencial; en /p/, /k/ y /k^w/ hay menor variación, lo que significa que contiene bajas frecuencias, mientras que /t/ tiene un contenido frecuencial más alto, como lo revelan los espectrogramas.

Existe una cuarta diferencia, de acuerdo a [Oli93], la posición de F2 entre el estallido y la vocal adyacente ayuda a diferenciar entre las oclusivas; de hecho el factor que más contribuye a dicha identificación es la transición o movimiento formántico de los sonidos adyacentes hacia/a partir de la oclusiva. La importancia de la transición dentro y fuera de las oclusivas ha creado el concepto de **loci**, definido como el valor de los formantes en el estallido. Debido a que durante el estallido la estructura formántica no es clara, el loci se infiere a partir del movimiento formántico hacia el siguiente sonido. El loci por sí mismo no puede ser determinado con precisión debido a que son muy variables; por eso es mejor describir a las oclusivas discutiendo el movimiento formántico que determinar el loci. El segundo formante es el que muestra mayor movimiento, por lo que en base a él nos ayudamos a identificar el tipo de oclusiva en base a su loci. En conclusión, la oclusiva velar labializada presentan el loci más bajo, seguida de la labial, la alveolar y la velar, esta última presenta el loci más alto de todas.

/k^w/ es una oclusiva velar labializada, es decir, se redondean los labios en la oclusión. Es necesario mencionar que /k^w/ además está integrada por un segmento similar al sonido aproximante /w/, el cual se revisará más adelante. Por lo tanto, la oclusiva /k^w/ está conformada por la región de cierre, el estallido y una estructura formántica semejante a /w/, donde F1 y F2 tienen valores muy bajos (por eso el loci resulta muy bajo), sin embargo F2 se eleva rápidamente para alcanzar el valor de la vocal siguiente (el fono [a], en el ejemplo mostrado). Note que los formantes superiores a F2 poseen mayor energía en la vocal que en la oclusiva /k^w/.

En resumen, las consonantes oclusivas pueden ser identificadas en una forma de onda o espectrograma si dos segmentos de tiempo distintos, el cierre y el estallido, son vistos. Teóricamente, en su representación mejor formada, una oclusiva no sonora no muestra ninguna periodicidad en el intervalo de cierre. El intervalo después del estallido, si la oclusiva es seguida de un sonido sonoro, es referido como *tiempo de comienzo sonoro* (voice onset time, VOT). Respecto al VOT de las consonantes oclusivas, la bilabial y la alveolar tienen el menor VOT, mientras que las velares el mayor. La sección de estallido de las bilabiales es la más débil y muestra la mayor energía en bajas frecuencias. La energía de la /t/ es mayor que las otras

oclusivas. La sección de estallido de la velar a veces muestra un doble estallido; dado que este doble estallido no es visto en ninguna otra oclusiva, su presencia indica una /k/. Por la variación de la forma de onda en el estallido, /t/ presenta un contenido frecuencial más alto que en /p/, /k/ y /k^w/. El loci de F2 para la oclusiva velar labializada es el más bajo, mientras que el de la velar es el más alto.

6.1.2 Fricativas

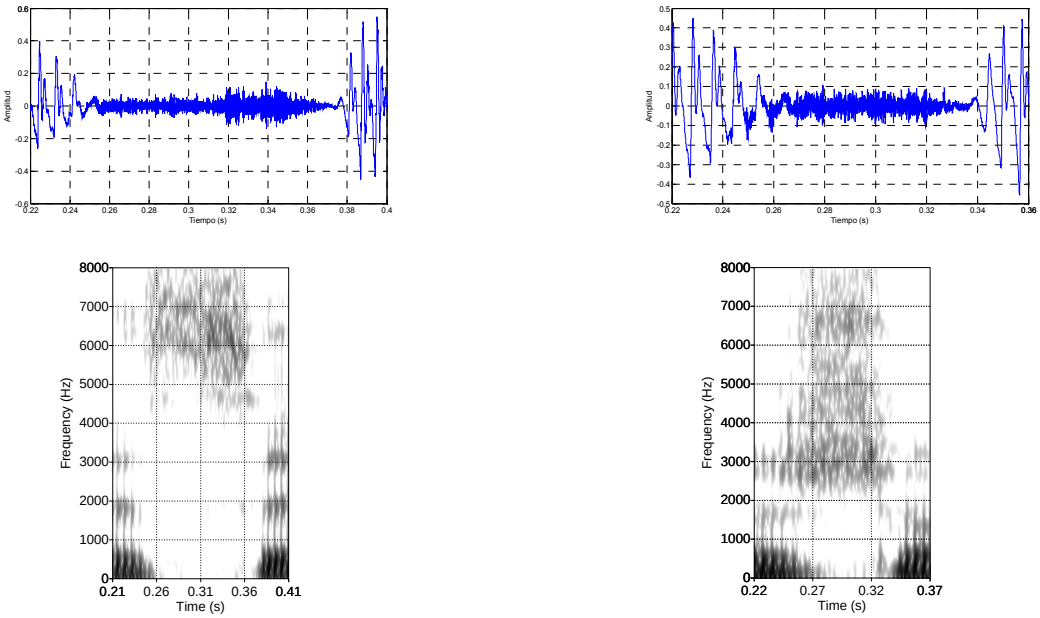
Las fricativas del náhuatl de la región de Cuetzalan se pueden distinguir por su punto de articulación. Las fricativas sordas alveolar y alveopalatal (/s/ y /š/, respectivamente) son articuladas con un flujo continuo de aire y se producen cuando el tracto-vocal se estrecha lo suficiente para causar turbulencia, pero no demasiado como para cerrar por completo el flujo de aire. Por eso se dice que un sonido fricativo tiene una apertura mínima del tracto vocal.

Por otra parte, la fricativa /h/ se produce tensando las cuerdas vocales para producir una fricción turbulenta. Lo anterior indica que la constricción se presenta en la glotis y no propiamente en el tracto vocal como en el caso de las otras fricativas, donde toman una participación los articuladores (como la lengua y los dientes). Para producir la /h/, ante la ausencia de articulación, el tracto-vocal actúa enteramente como resonador.

Una característica que diferencia /h/ de las otras fricativas es su rasgo de sonoridad. A pesar de que las cuerdas vocales forman una constricción para /h/, aún estarán habilitadas para vibrar. Por lo tanto, la fricativa /h/ puede ser sonora o sorda; sin embargo, la sonoridad no es fonémica. Es decir, la presencia o ausencia de sonoridad para /h/ está determinada en parte por el contexto fonético. La calidad sonora está influenciada por el rasgo sonoro de los segmentos de sonidos adyacentes. [Oli93] advierte que "hasta donde se sabe, no hay una regla predecible que dictamine la cualidad sonora de /h/. /h/ es generalmente sorda; en un ambiente sonoro, puede llegar a ser sonora" [Oli93]

En la figura 6.4 se presentan las formas de onda y los espectrogramas de las fricativas supraglotales; corresponden a las fricativas intervocálicas de las realizaciones [sesek] (frío) y [šošoktik] (verde); se incluyen parte de los sonidos adyacentes. La escala se ajustó de 0 Hz a 8 kHz para una apreciación más cómoda.

Como puede observarse, en las formas de onda de ambas fricativas hay mucha energía y variación de la señal. El espectro de [s] muestra una distribución relativamente uniforme de energía por arriba de los 4000 Hz. El espectro de [š] muestra una distribución relativamente uniforme de energía por arriba de los 2000 Hz. Las anteriores observaciones nos indican que las fricativas sordas pueden ser diferenciadas por su distribución de energía en los espectrogramas, siendo la alveolar más aguda que la alveopalatal. La duración de la fricativa alveolar es de 130 ms, mientras que el de la alveopalatal es de 85 ms.



i. [ese] ii. [ošo]

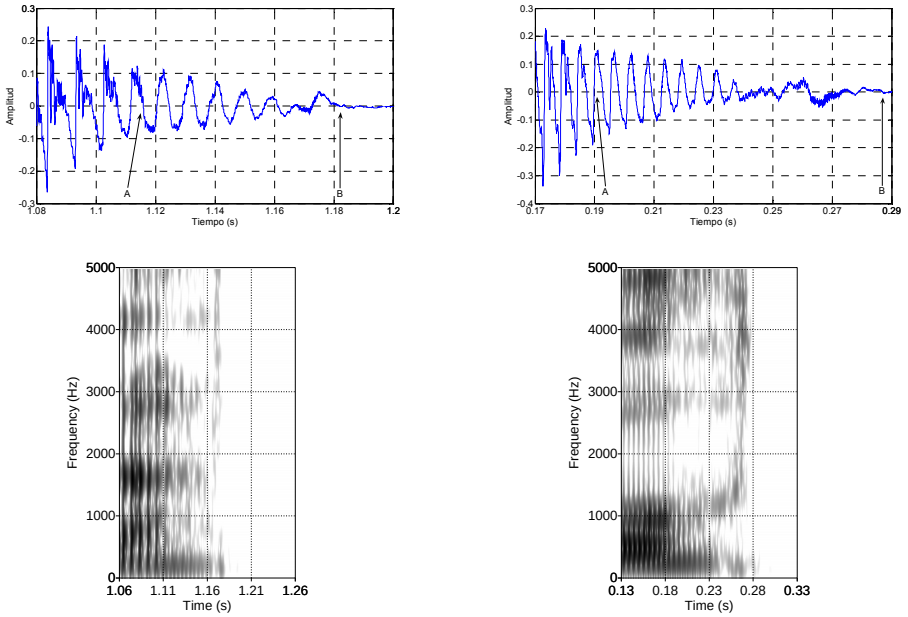
Figura 6.4: Realizaciones de las fricativas /s/ y /š/

Ahora se analizará la consonante /h/. Este sonido se produce poniendo rígidas las cuerdas vocales para producir una fricción turbulenta. El fonema /h/ se diferencia de las demás fricativas debido a que la constricción se presenta en la glotis y no propiamente en el tracto vocal. La constricción de las otras fricativas involucran a los articuladores (lengua, dientes y labios); la producción de /h/, por el contrario, deja libres a estos articuladotes. Además, debido a que la constricción de /h/ está en la glotis, todo el tracto vocal se comporta como el resonador, mientras que para las otras fricativas, sólo la parte del tracto vocal enfrente de la constricción sirve como la principal cavidad de resonancia.

Otra característica que diferencia a /h/ de las otras fricativas es su rasgo de sonoridad. A pesar de que las cuerdas vocales forman una constricción estrecha, aún pueden vibrar. Por lo tanto /h/ puede ser sonora o no; por esta razón en el sistema fonológico presentado no se clasificó a /h/ como sonido sordo o sonoro. El rasgo de sonoridad de /h/ puede preverse conforme a la sonoridad de los sonidos adyacentes.

La figura 6.5 muestra la producción de /h/ entre vocal y oclusiva, son segmentos de las realizaciones [koyotahtol] (español) y [ohti'] (camino). En las formas de onda se señalan el inicio y fin del fono fricativo Se ha aplicado preénfasis de 6dB/oct a los espectrogramas. Realzando las frecuencias superiores podemos observar con mejor detalle el comportamiento de la fricativa. La fricativa se señala con las flechas A, B; es evidente un comportamiento sonoro en buena parte de ella por influencia de la vocal que la precede. Los espectrogramas revelan un comportamiento formántico; sin embargo éstos, a comparación de los de las vocales, no poseen mucha energía. Al final de cada fricativa es evidente el segmento de cierre de las oclusivas.

Haremos observaciones más detalladas en las formas de ondas para revelar por qué /h/ es un sonido fricativo. A partir de la flecha A, sobre todo en la figura 6.5.i, se observa un comportamiento ruidoso tanto en las crestas como cerca de B. En la figura 5.ii un comportamiento ruidoso es notorio cerca de B.



i. [aht] ii. [oht]

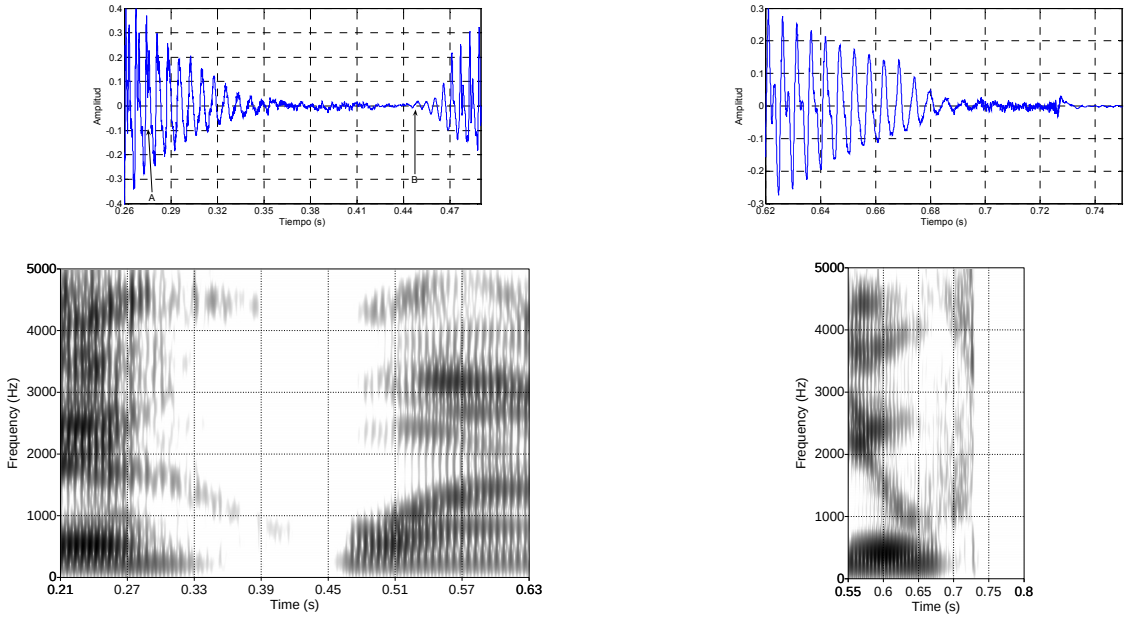
Figura 6.5: Realizaciones de la fricativa glótica /h/ entre vocal y oclusiva

En ocasiones es posible confundir un sonido aspirado con una fricativa glotal; ante ello, la estructura silábica del náhuatl nos permite identificar estos sonidos. La figura 6.6.i muestra a /h/ dentro del segmento [ehwa]. Por otro lado la figura 6.6.ii muestra el segmento [iwt]. Son segmentos de las realizaciones [tehwan] (nosotros) y [nisiwtok] (estoy cansado). Nuevamente se aplicó preénfasis en los espectrogramas.

Se describe a continuación la figura 6.6.i. Al estar determinada por el contexto fonético, /h/ es periódica cuando viene precedida de sonidos sonoros. Sin embargo, cuando la sonoridad del sonido precedente deja de influir, /h/ manifiesta con mayor claridad su naturaleza fricativa. Gracias a un marcado descenso de energía en el espectrograma, en la forma de onda de la figura 6.6i se señala el inicio y fin de la fricativa a través de las flechas A y B, la fricativa es sorda alrededor de 100 ms; después comienzan los pulsos glóticos de la aproximante, donde por supuesto no hay efectos de la fricativa. Los formantes se mueven durante [h] (vea F2) de la misma manera que lo harían en una transición entre vocales (o aproximantes, según sea el caso). Este movimiento en lo formantes refleja la habilidad de la lengua para moverse libremente durante la producción de /h/ ya que la producción de este sonido no tiene ninguna posición particular de la lengua asociada a ella.

Por otro lado, la figura 6.6.ii muestra un comportamiento muy semejante al de la figura que la precede, se nota en la forma de onda un comportamiento fricativo; sin embargo se trata de

una aproximante aspirada. La razón de lo anterior es que la estructura silábica del náhuatl no permite dos consonantes juntas en ella. Además note en el espectrograma que la región de la aspiración muestra un comportamiento más bien ruidoso, no hay claramente una estructura formántica como ocurre en los ejemplos de fricativa glotal.



i. [ehwa] ii. [iwt]

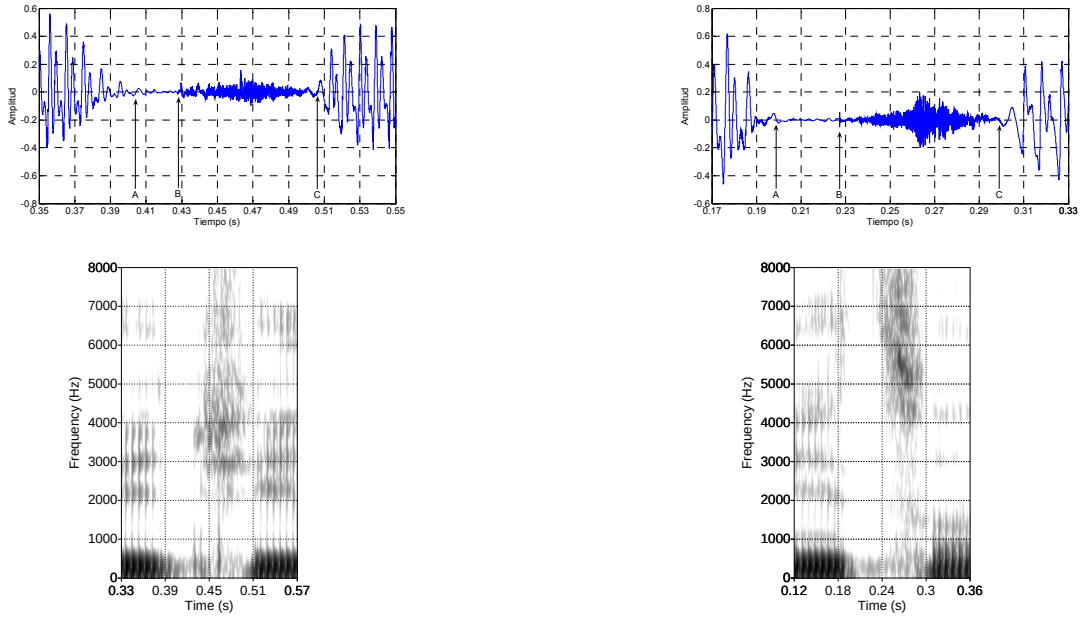
Figura 6.6: Ejemplo de presencia y ausencia de fricativa glótica /h/

6.1.3 Africadas

Las africadas son una categoría intermedia entre oclusivas simples y la secuencia de una oclusiva y una fricativa. Como en muchos tipos de sonidos, las africadas no tienen fronteras bien claras. Consideraciones fonológicas deben jugar una parte para decidir si una oclusiva y una fricativa homorgánica -es decir, "la zona de articulación de la fricativa es igual o casi completamente igual que la de la oclusiva" [Sil]-, que le preceda debe ser vista como una africada, la cual es una sola unidad, o como dos segmentos que forman una secuencia de una oclusiva y una fricativa.

La producción de africadas puede involucrar solamente un ligero ensanchamiento de la constricción articulatoria de la oclusiva, de tal manera que las componentes oclusiva y fricativa tengan idéntico punto de articulación. Algunas africadas, sin embargo, involucran un pequeño ajuste hacia adelante o hacia atrás de la posición del articulador activo.

En el náhuatl tenemos dos africadas sordas, la alveolar /ȼ/ y la palatal /č/. En la figura 6.7 se muestran las gráficas de estos sonidos consonánticos a través de las realizaciones [čičiltik] (rojo) e [iȼonteko] (su cabeza).



i. [ičі] ii. [i¢o]

Figura 6.7: Realizaciones y detalles de las africadas /č/ y /¢/ intervocálicas

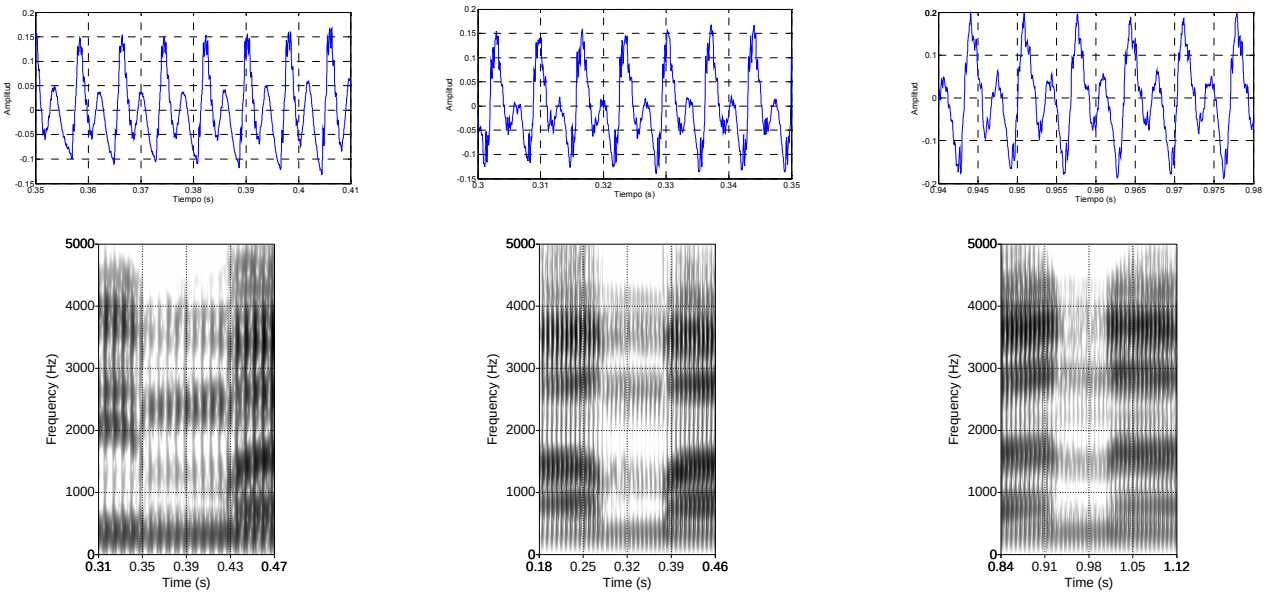
Las africadas están señaladas en la forma de onda mediante las flechas A y C. Observe que debido a la combinación oclusiva-fricativa las africadas contienen una sección de silencio. En dicha sección, las formas de onda presentan variaciones muy lentas, lo cual se refleja en energía a baja frecuencia, esto es revelado por los espectrogramas. Pareciera que la africada asimila un poco la sonoridad de la vocal precedente. De todos modos en los espectrogramas la secciones de silencio no revelan una estructura formántica, por lo que no se puede decir que existe sonoridad en la africada. Como ocurre con las oclusivas, existe un estallido en ambas africadas, se aprecia que existe un repentino cambio de amplitud. Es usual que al final de la región de cierre existe un estallido corto, observe las formas de onda, donde se ha señalado dicho fenómeno con la flecha B. El estallido de la oclusiva es visible en el espectrograma como una línea vertical justo antes de la región fricativa. La sección fricativa de la africada es evidente al observar que la energía se va incrementando después del estallido y gradualmente decrece. El VOT de la africada palatal es de 78 ms, mientras que el de la alveolar es de 73ms.

6.1.4 Nasales

Los sonidos nasales, al igual que las vocales, laterales y aproximantes, son considerados fonemas sonoros. Las nasales presentes en el náhuatl son la bilabial /m/ y la velar /n/. Los sonidos nasales y oclusivos se producen con el cierre completo de la cavidad oral; la diferencia entre estos sonidos radica en el velo o puerto nasal. El velo está abierto para la producción de las nasales, permitiendo que el flujo de aire escape a través de la cavidad nasal. Cuando el flujo de aire está restringido en la cavidad oral debido al cierre completo de los labios, la apertura del velo permite que el aire fluya a través de la cavidad nasal. Debido a que el aire es liberado a través de la naríz,

no hay presión acumulada. Por lo tanto, no ocurre una explosión, como en el caso de las oclusivas, cuando el cierre oral es liberado.

La figura 6.8 presenta las formas de onda (del sonido nasal) y los espectrogramas (con parte de los sonidos adyacentes) de las nasales en posición intervocálica. Se ha aplicado preénfasis de 6dB/oct a los espectrogramas. Corresponden a los segmentos [ima], [ana] (de [nimayana], tengo hambre) y [ama] (de [aman], hoy). No hay mucha diferencia en las formas de onda de las nasales; ambas muestran una variación lenta, una característica asociada con señales que tienen poca energía en las regiones de alta frecuencia. Los espectrogramas muestran que las nasales tienen un F1 muy por debajo de los 1000 Hz, denominándosele ***formante nasal***. A pesar de que la energía arriba de F1 es débil, alguna estructura resonante es observable. Es posible identificar F2 y F3; sin embargo hay aún resonancias arriba de estos formantes; vea por ejemplo el espectrograma del fono [n]. Los sonidos sonoros orales son producidos con el tracto oral configurado como un simple tubo. Los sonidos nasales se producen con el tubo configurado desde la glotis hasta el tracto nasal. La geometría del tubo es complicada por la adición de la cavidad oral cerrada. Las resonancias de la cavidad oral interactúan con las resonancias del tubo de una manera compleja; por lo tanto, el espectrograma muestra una estructura resonante múltiple, lo que hace difícil etiquetar formantes altos.



i. [ima] ii. [ama] ii. [ana]

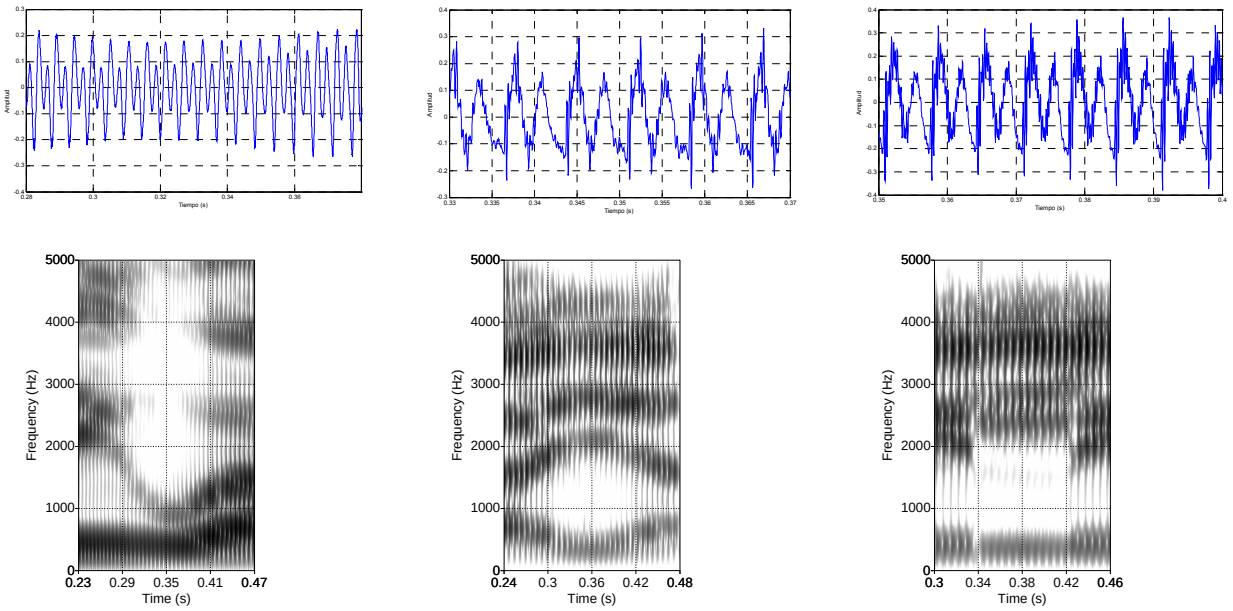
Figura 6.8 Realizaciones de /m/ y /n/

Un indicador de un sonido nasal es una clara y marcada discontinuidad entre los formantes de la nasal y los sonidos adyacentes. El fono [m] tiene un bajo F2 y por tanto puede ser identificado por un movimiento hacia abajo del F2 del sonido precedente y por el movimiento hacia arriba del sonido siguiente. El segmento [ima] muestra de manera muy clara este fenómeno. El fono precedente [i] posee un elevado F2, es notoria la discontinuidad con la F2 de [m].

6.1.5 Aproximantes y lateral

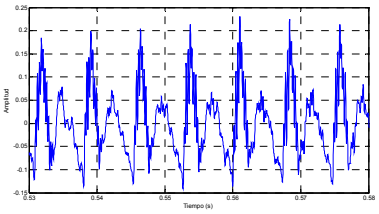
Los sonidos aproximantes tienen formas de onda semejantes a las vocales; como tales, no revelan mucha información acerca de sus características, por lo que resulta necesario estudiar sus espectrogramas.

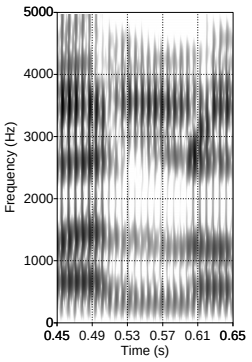
La figura 6.9 muestra las formas de onda (del sonido aproximante y lateral) y los espectrogramas (con parte de los sonidos adyacentes) de las dos aproximantes y la lateral en contextos intervocálicos. Nuevamente se ha aplicado preénfasis de 6dB/oct a los espectrogramas. Son segmentos de las realizaciones [siwat] (mujer), [kaya'] (todavía no), [pili'] (niño) y [tapalol] (comida). Se observa que el fono [w] tiene valores bajos en F1 y F2. El fono [y] tiene un F1 bajo y un F2 extremadamente alto, muy cercano a F3. El sonido lateral y los aproximantes son diferentes a las otras consonantes; estos sonidos no exhiben discontinuidades de formantes en sus transiciones de la vocal precedente a la vocal siguiente. Razón de ello es que la lateral y las aproximantes no son altamente constrictivas, los formantes son más continuos en las uniones con vocales. Note que [l] muestra una ligera discontinuidad; sin embargo ésta no es parecida a la discontinuidad presente en [n]. Esto es debido a que a pesar de que [l] y [n], son alveolares, la posición de la lengua para producir [l] es muy diferente a la posición correspondiente para [n].



i. [iwa] ii. [aya] iii. [ili]

Figura 6.9 (continúa): Realizaciones de /w/, /y/, /l/





iv. [alo]

Figura 6.9 (continuación): Otra realización de /l/

6.2 Vocales

Los sonidos vocálicos tienen una característica principal: la existencia de estructuras de formantes claras. Esto es fruto de la emisión del flujo de aire por el conducto bucal sin apenas resistencia, con las cavidades resonadoras potenciando los armónicos distintivos de cada vocal. En suma, todas las vocales son sonoras.

Como se aprecia en el sistema fonológico, existen vocales de corta y larga duración. La duración de las vocales es un rasgo importante del náhuatl pues el significado de las palabras puede ser muy diferente. Un ejemplo es el par mínimo "/čiči/ (perro) y /či:či/ (seno)" [Cas00].

Las vocales son formadas con diferentes posiciones de la lengua, labios y mandíbula. "A pesar de que las vocales pueden ser identificadas refiriéndose a la altura de la lengua, la posición de la lengua y el redondeo de los labios, la descripción es relativa e inespecífica" [Oli93]. Por lo tanto, describirlas en términos de las articulaciones involucradas resulta poco preciso. Dado que las vocales presentan estructuras de formantes bien definidas, resulta más conveniente describirlas en términos de sus propiedades acústicas.

Las vocales están ampliamente separadas en el espacio de los formantes F1 y F2 para minimizar confusiones perceptuales. La situación exacta de los formantes varía según el hablante y la realización del habla.

En la figura 6.10 se muestran las formas de onda y espectrogramas de los sonidos vocálicos. Corresponden a las palabras [a:t] (agua), [se'] (uno), [tit] (fuego) y [totonik] (caliente).

Se ha aplicado preénfasis de 6dB/oct a los espectrogramas de la figura 6.10 con el fin de observar mejor los formantes. Se presentan las gráficas de las vocales del náhuatl pronunciadas por un hombre y una mujer. Observe que los formantes de la mujer están ubicados a frecuencias más altas que las del hombre; esto es debido a la diferencia anatómica del tracto-vocal entre ambos sexos. Los formantes F1 y F2 son suficientes para caracterizar a las vocales. El formante F3 resulta útil para distinguir un sonido vocálico de un consonántico, observe que prácticamente conserva la misma frecuencia para las diferentes vocales.

Resulta importante hacer notar que hay una correlación entre la altura y posterioridad de las vocales con su estructura formántica. La altura de la vocal y la ubicación de F1 son inversamente proporcionales. Respecto a F2, más alta su ubicación, la vocal es más anterior; y si es más baja, la vocal es más posterior.

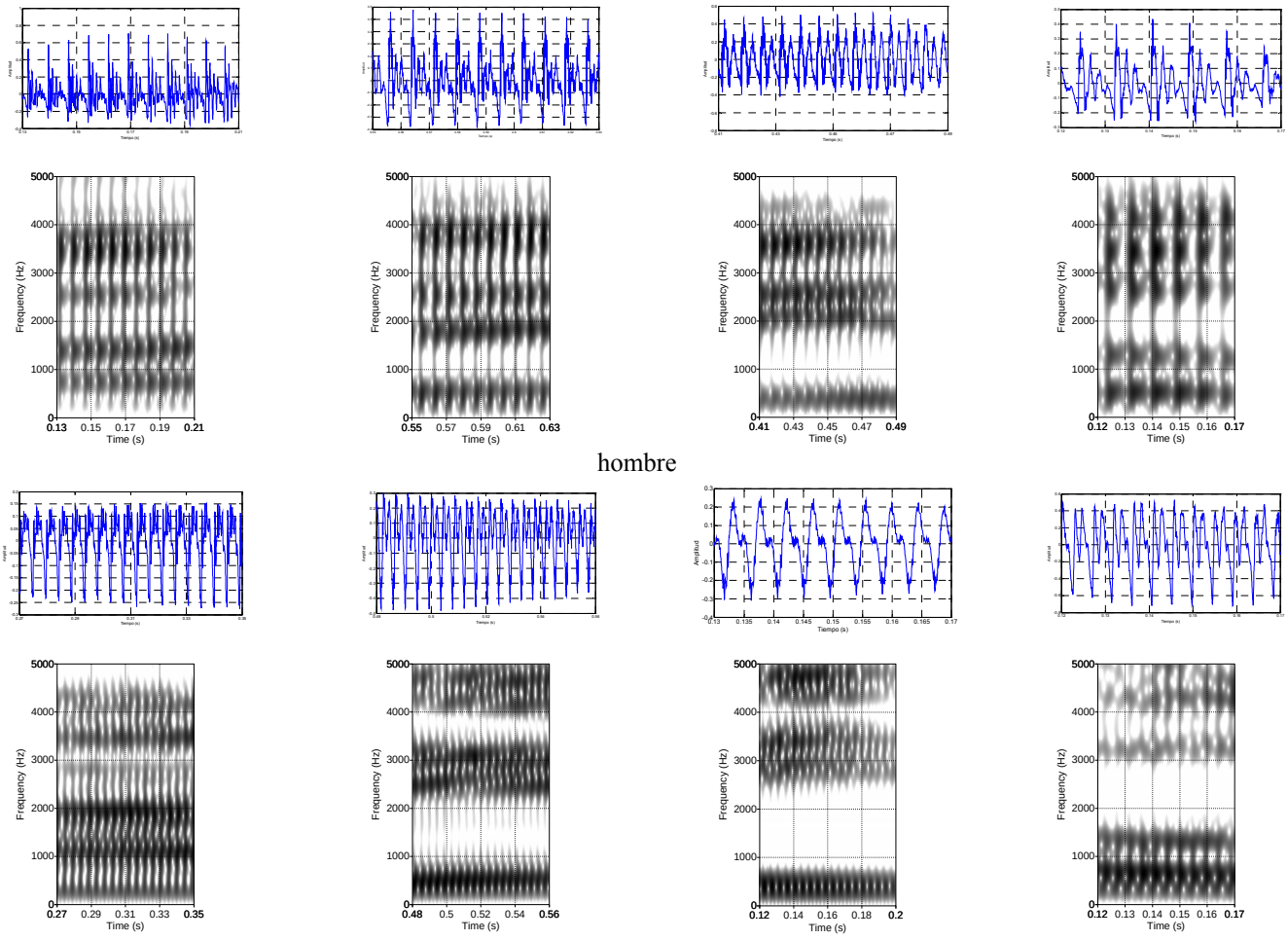


Figura 6.10: Realizaciones de /a/, /e,/ /i/, /o/

En la tabla 6.1 se presentan los valores de los formantes F1, F2 y F3 por vocal y por sexo. Estos valores son aproximados, y aunque brindan una guía y no valores absolutos, son bastante típicos de dichas vocales.

Hombre						Mujer				
Vocal	[i]	[e]	[o]	[a]		Vocal	[i]	[e]	[o]	[a]
F1	341	476	484	740		F1	396	473	569	1053
F2	2046	1798	1290	1442		F2	2805	2519	1275	1921
F3	2546	2562	2693	2575		F3	3408	3054	3227	3345

Tabla 6.1: Valores de F1, F2 y F3 de las vocales de la figura 10 (valores típicos, en Hertz)

Para determinar si una lengua cuenta en su sistema fonológico con vocales de larga duración, como en el caso del náhuatl, es suficiente si posee una duración temporal de al menos el doble de la vocal de corta duración. En las figuras 6.11 y 6.12 presentamos evidencia de esto al presentar segmentos de las realizaciones [tapalol] (comida) y [a:t] (agua) con el fin de contrastar

duraciones. En la forma de onda de la figura 6.11 se señalan las dos vocales, sus duraciones son de 52 ms y 82 ms.

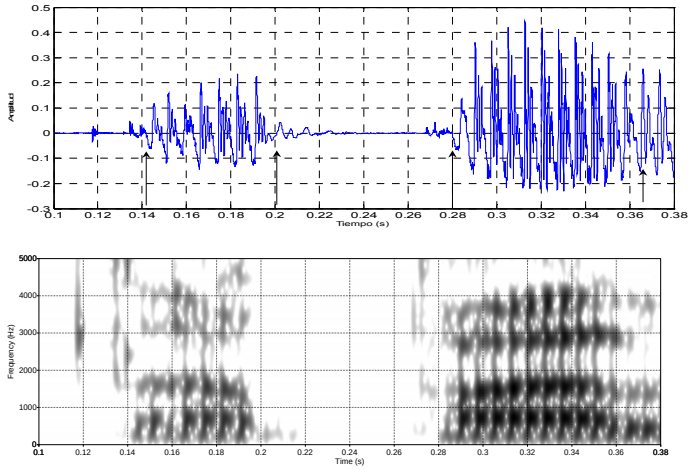


Figura 6.11: Segmento [tapal], palabra [tapalol]

La figura 6.12 muestra tres realizaciones de la palabra [a:t] por tres hablantes masculinos distintos. Los espectrogramas superior y central pertenecen a [a:t] dentro de una oración: [titayis tepiɕin a:t] (quieres tomar poquita agua). El espectrograma inferior corresponde a la misma palabra pero grabada aisladamente, sin formar parte de una oración. Las duraciones son, en orden superior a inferior, de 180 ms, 194 ms y 160 ms. Claramente la duración de la vocal larga es al menos del doble de la corta.

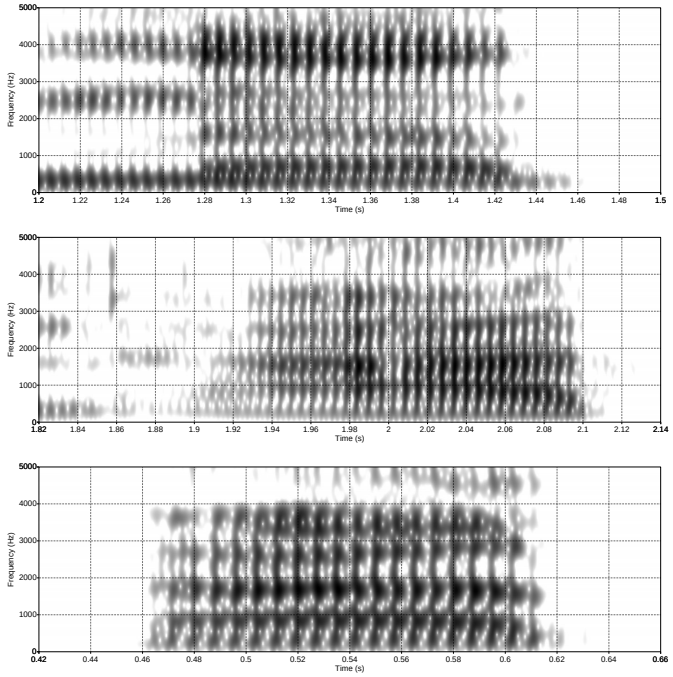


Figura 6.12: Diferentes realizaciones del segmento vocálico de la palabra [a:t]

6.3 Fonética acústica estática a final de sílaba

Examinaremos ahora los sonidos fonéticos en posición coda, cabe mencionar que no todos los fonemas se encuentran en esta posición.

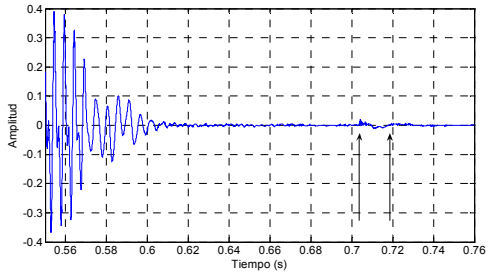
6.3.1 Oclusivas

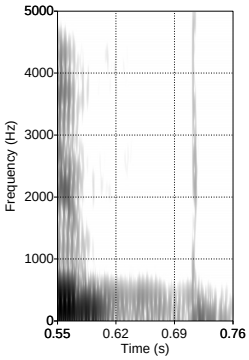
Uno de los rasgos observados en los sonidos oclusivos alveolares y velares a final de sílaba es una mayor duración del estallido, resultando muy notoria una aspiración de las oclusivas a final de palabra. En general, la versión aspirada de una oclusiva es más larga que la variante no aspirada de la misma consonante.

Por otra parte, la bilabial /p/ en posición coda sigue siendo el sonido oclusivo que muestra una menor intensidad de su estallido; sin embargo su duración presenta la mayor variación. Se presentarán dos ejemplos de ello.

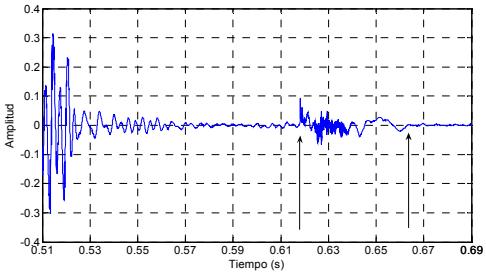
La figura 6.13 muestra las formas de onda y espectrogramas de las oclusivas labial (la misma palabra pronunciada por diferentes hablantes), alveolar y velar a final de sílaba, fueron extraídas de las realizaciones [yawipta'] (antier), [takat] (hombre) y [pokti'] (humo) respectivamente. Las duraciones de los estallidos de las bilabiales presentadas son de 15 ms y 45 ms, mientras que la de la alveolar en fin de palabra se extiende hasta los 113 ms, es notoria la aspiración. La duración de la velar en fin de sílaba es de 55 ms.

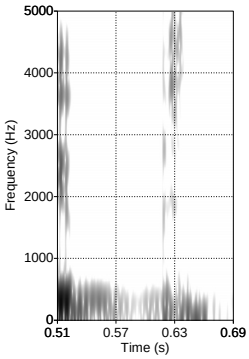
No se contaba en el corpus con la velar labializada /k^w/ en posición coda, aunque en las fuentes consultadas no se halló alguna palabra con esta característica.



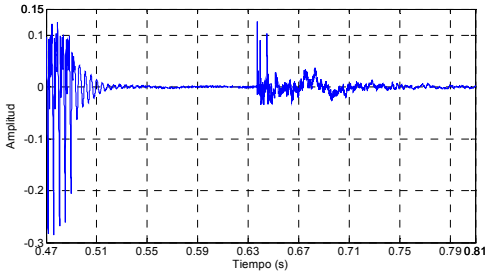


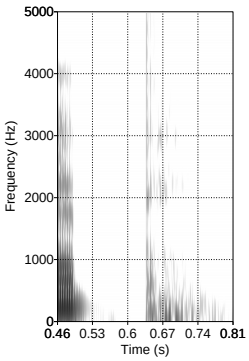
i. [ip]



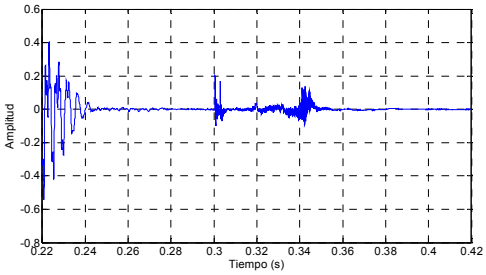


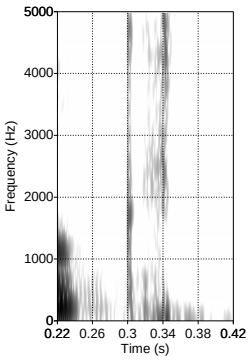
ii. [ip]





iii. [at]





iv. [ok]

Figura 6.13: Oclusivas en posición coda

6.3.2 Fricativas

En la figura 6.14 se presentan la forma de onda y espectrograma de la fricativa alveolar /s/ en posición coda. Es un ejemplo de la realización [esti'] (sangre). Se observaron las realizaciones de los hablantes, sin embargo no se estableció alguna diferencia del fonema si éste estaba en posición silábica onset o coda.

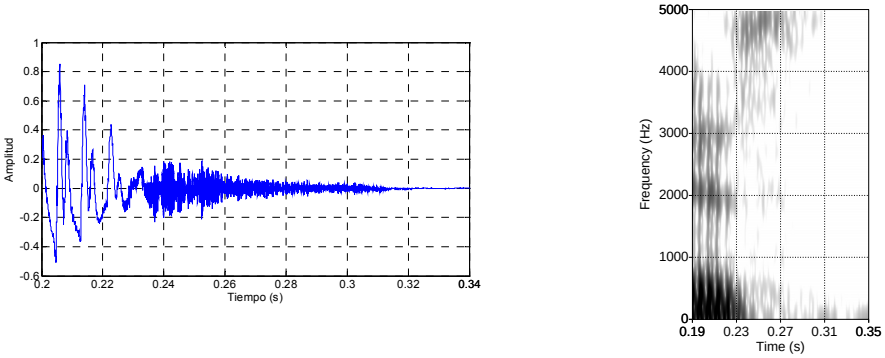


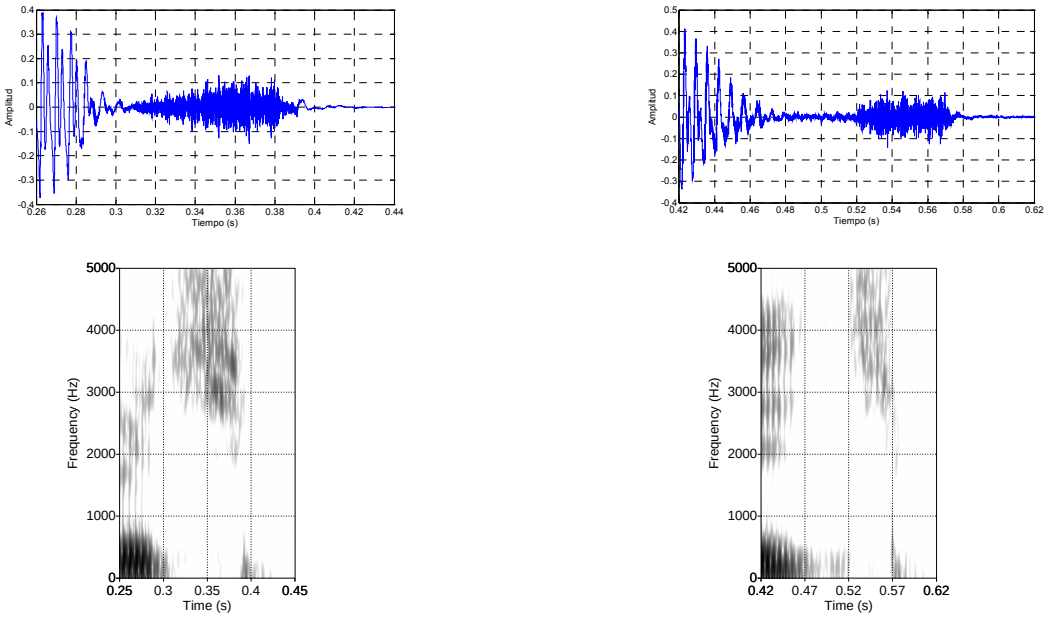
Figura 6.14: Fricativa [s] en posición coda

En el corpus no se disponía de la fricativa alveopalatal /š/ en posición coda. Por otra parte, la fricativa /h/ se presenta en posición coda y ya ha sido descrita en la sección anterior.

6.3.3 Africadas

En la figura 6.15 se presentan las formas de onda y espectrogramas de la africada palatal en posición coda, son dos realizaciones de la palabra [okičpil] (joven) pronunciadas por distintos hablantes. A diferencia de la ubicación de la africada cuando está en posición onset, se observa la ausencia del estallido y del silencio que lo antecede. El cambio de vocal a africada no es tan abrupto.

A pesar de lo anterior, también se observaron pronunciaciones de otros hablantes donde el sonido africado conservaba sus características de silencio y estallido. Aún así, en todos los casos, la duración de la africada (es decir, descartando la sección de silencio cuando éste existía) se encontraba entre los 90 ms (como en el caso de la figura 6.15i) y los 110ms, implicando una duración mayor al de la africada en posición onset. La duración de la africada en la figura 6.15ii, incluyendo la sección de fricción de baja amplitud, es de 104 ms.



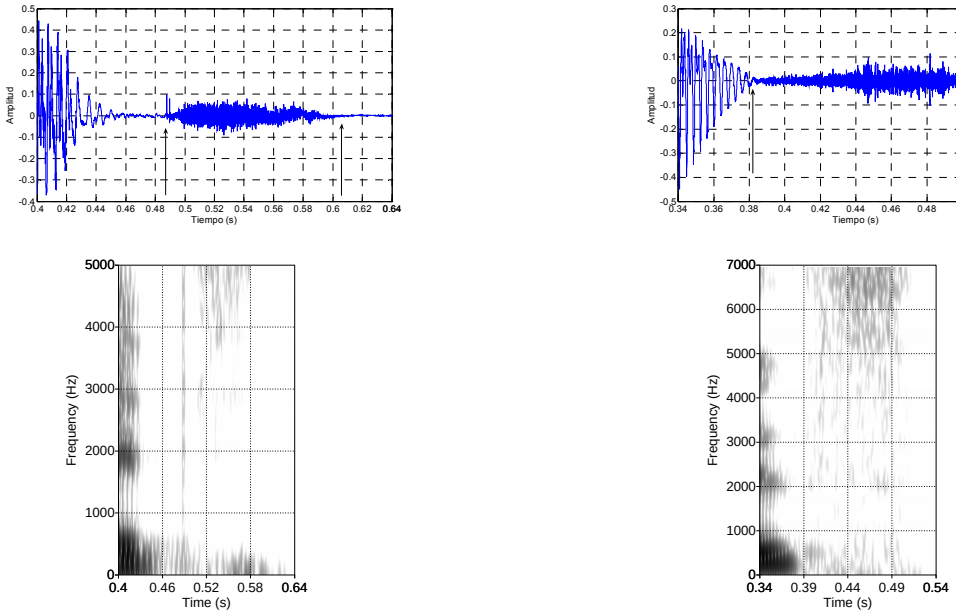
i. [ič] ii. [ič]

Figura 6.15: Africada [č] en posición coda de la misma palabra por diferentes hablantes

Para revisar el comportamiento de la africada alveolar /¢/ en posición coda, la figura 6.16 muestra dos realizaciones de la palabra [me¢ti'] (luna) pronunciada por distintos hablantes. Tres de los doce hablantes grabados omitieron la parte de oclusión (y por ende, de estallido) al pronunciar la palabra, vea la columna de la derecha en dicha figura.

La duración de las africadas (sin considerar la región de silencio, cuando lo había) mostradas son de 118 ms y 135 ms, corresponden a las columnas izquierda y derecha,

respectivamente. En general, la duración variaba de los 95 ms a los 135 ms. Esta duración es mucho mayor que en los casos donde la africada alveolar estaba en posición onset.



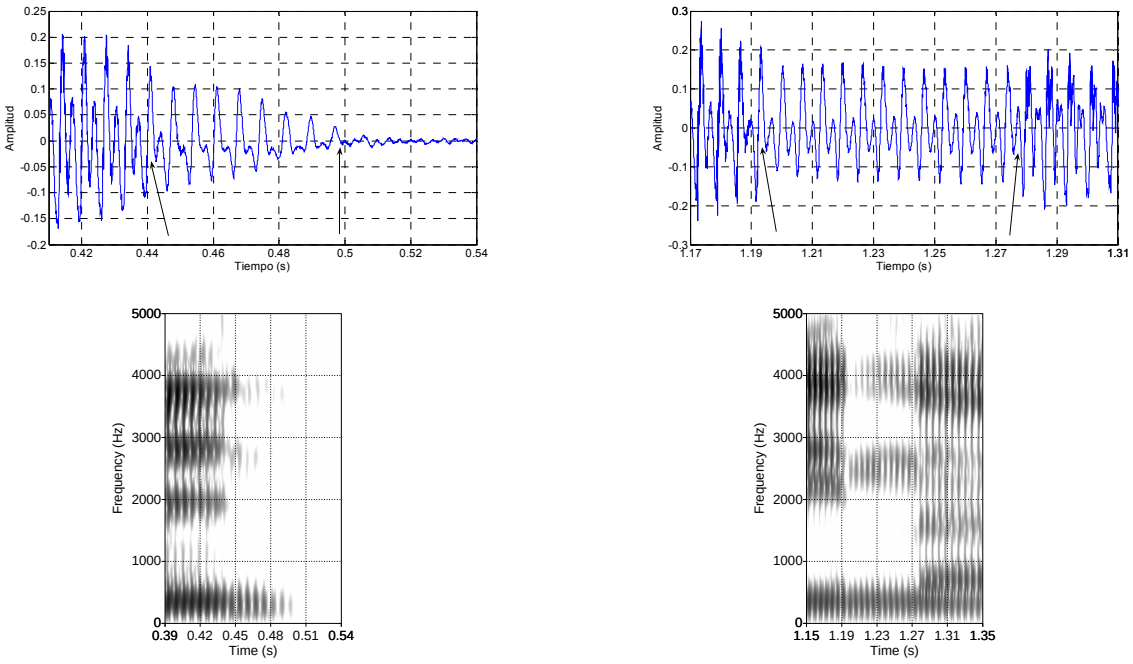
i. [eɕ] ii. [eɕ]

Figura 6.16: Africada [ɕ] en posición coda de la misma palabra y diferentes hablantes

6.3.4 Nasales

La figura 6.17 muestra segmentos de dos realizaciones de la nasal /n/ a final de sílaba. Se ha aplicado preénfasis a los espectrogramas. Estos ejemplos fueron tomados de la palabra [kanintinemi] (donde vives) y de la oración [titayis tepiʧin a:t] (quieres tomar poquita agua), ubicadas en las columnas izquierda y derecha respectivamente.

Es claro apreciar que el sonido adyacente posterior a la nasal influye en ella. La nasal pierde buena parte de su sonoridad, aunque conserva muy bien su formante nasal, cuando precede a un sonido sordo. Por el contrario, la nasal que precede al sonido sonoro recupera los rasgos que la caracterizan como tal: sonoridad, y por ende, una estructura formántica bien definida. El corpus no contenía ejemplos con la nasal /m/ a final de sílaba.



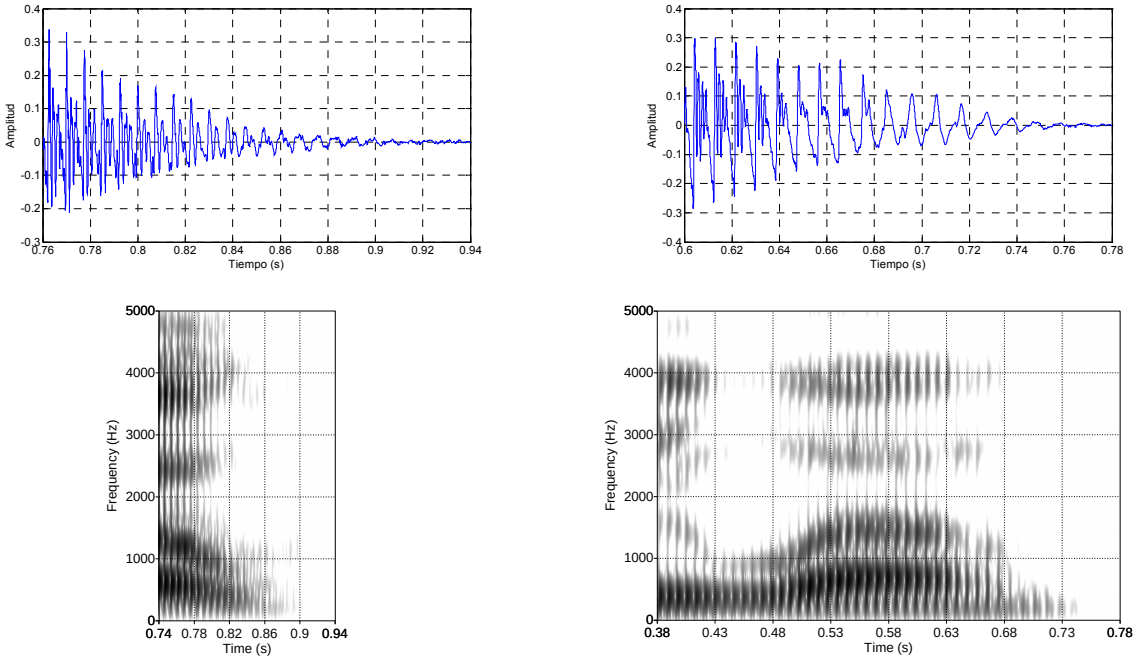
i. [in] ii. [ina]

Figura 6.17: Aproximante [n] en posición coda

6.3.5 Aproximantes y lateral

La figura 6.18 muestra los segmentos de dos realizaciones de la aproximante /w/ a final de sílaba (y de palabra). Se ha aplicado preénfasis a los espectrogramas. Los segmentos provienen de [notakaw] (mi esposo) y [nosiwaw] (mi esposa), fueron pronunciadas por diferentes hablantes.

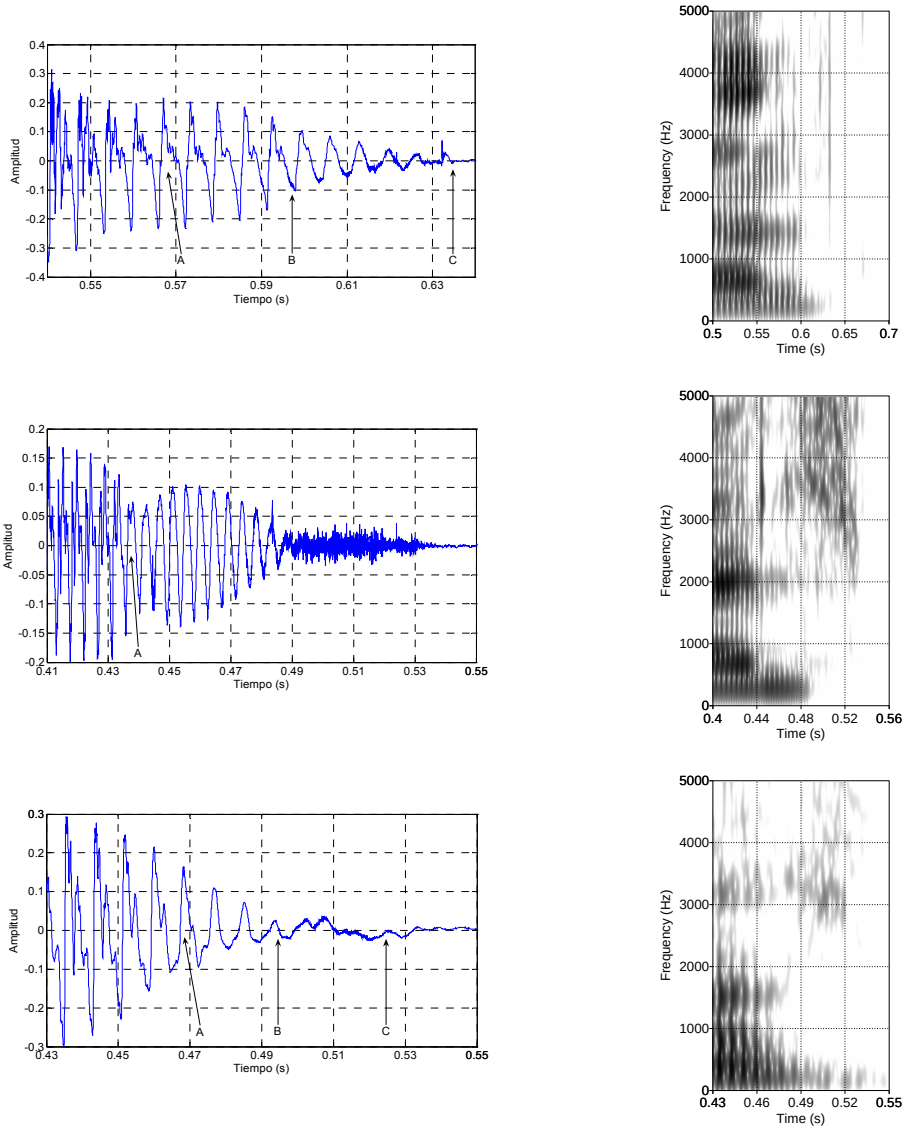
En estos ejemplos se aprecia un ensordecimiento de la aproximante, se observa que los formantes F1 y F2 de la aproximante a final de palabra tienden a extinguirse mientras alcanzan su posición más baja. El segmento [iwaw] nos permite observar en la misma realización dos aproximantes en posiciones silábicas diferentes.



i. [aw] ii. [iwaw]

Figura 6.18: Aproximante [w] en posición coda

Por otra parte, en la figura 6.19 se presentan segmentos de realizaciones por tres diferentes hablantes donde la lateral /l/ se localiza a final de sílaba, provienen de la palabra [tonal¢in] (solecito). El inicio de la lateral ha sido señalado con la flecha A. En todos los casos la lateral adopta una característica ruidosa al final de su producción, el grado de ruido es variable; ya que en los incisos i y iii el ruido no es tan notorio ha sido conveniente señalarlo entre las flechas B y C, donde el ruido comienza a manifestarse en los valles de la forma de onda. En todos los casos los espectrogramas nos permiten apreciar la presencia del fono [l] debido a una discontinuidad de F1 con respecto al F1 del fono vocálico que lo antecede; además se observa una menor intensidad de los formantes de [l]. Al respecto de la aproximante /y/ ubicada a final de sílaba, el corpus no cuenta con este caso.



i.

ii.

iii.

Figura 6.18: Laterales [l] en posición coda, segmento [al]

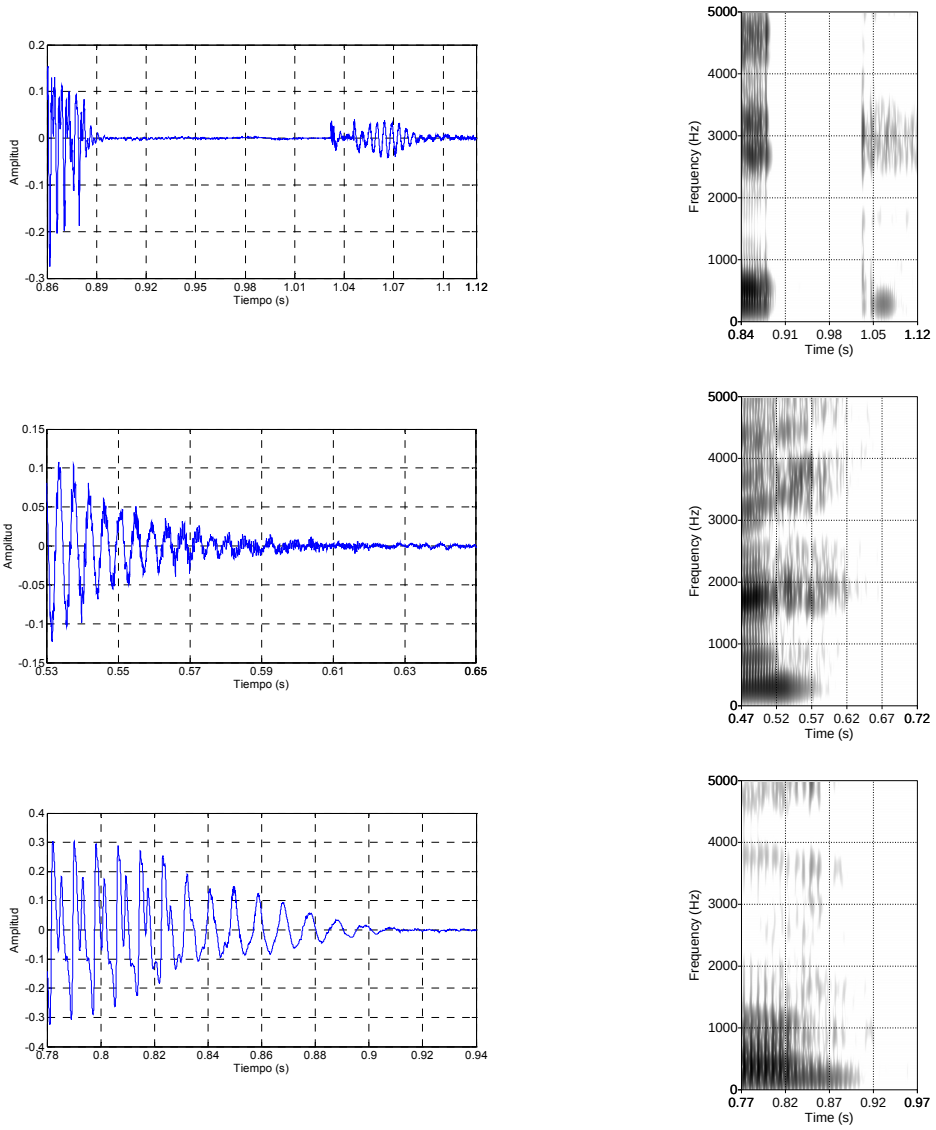
6.3.6 Vocales

Debido a que en lengua náhuatl las vocales se encuentran en el núcleo de la sílaba, no es posible hablar de una vocal en posición coda; sin embargo revisaremos casos donde las vocales se encuentran a final de palabra pues usualmente se presentan tres comportamientos: presencia de saltillo, aspiración y debilitamiento. La figura 6.20 muestra tales casos, en orden superior a inferior respectivamente. Todos los espectrogramas tienen preénfasis.

La figura 6.20i es un extracto de la realización [nopili'] (mi hijo), la realización tiene un saltillo bastante marcado, note en el espectrograma que se manifiesta una estructura cuasi-formántica en esa región, la cual coincide con la de la vocal que lo precede. Esto es debido a que, para producir el saltillo el hablante aprieta la glotis para luego relajarla, durante este proceso el resto de los articuladores se mantuvieron en la misma posición, por eso al exhalar se produce dicha estructura cuasi-formántica. Esto explica que la forma de onda de [i] no disminuya gradual y lentamente, también observe en el espectrograma un marcado cambio de energía entre el silencio que precede al saltillo y la vocal. Por último, observe que los formantes de la vocal están claramente marcados, indicando que no hubo un debilitamiento de este fono.

La figura 6.20ii presenta un extracto de [kema] (sí), observe que gradual y lentamente el fono [i] va terminando su realización pero al mismo tiempo se va contaminando de ruido y que corresponde a la aspiración; esto es observable en la forma de onda. El espectrograma muestra una estructura cuasi-formántica y ruidosa en la región de la aspiración.

Al respecto del último ejemplo de la figura 6.20iii, se presentan la forma de onda y espectrograma de la vocal final de la realización [iɬonteko] (su cabeza). Observe la ausencia de ruido, es una vocal no aspirada. El fono [o] se presenta en toda su duración, observe en el espectrograma el notorio debilitamiento de los formantes superiores a F1.



i. [i']

ii. [a]

iii. [o]

Figura 6.20: Vocales con posición en fin de palabra

Por último, para que se tenga una visión más clara del debilitamiento de las vocales en fin de palabra, vea la figura 6.21. Se presenta la realización completa de [ohti] (camino). Se ha señalado con las flechas A y B en la forma de onda a la vocal [i]. Note la diferencia de amplitud con la vocal [o]. Esto es reflejado en la energía de los formantes.

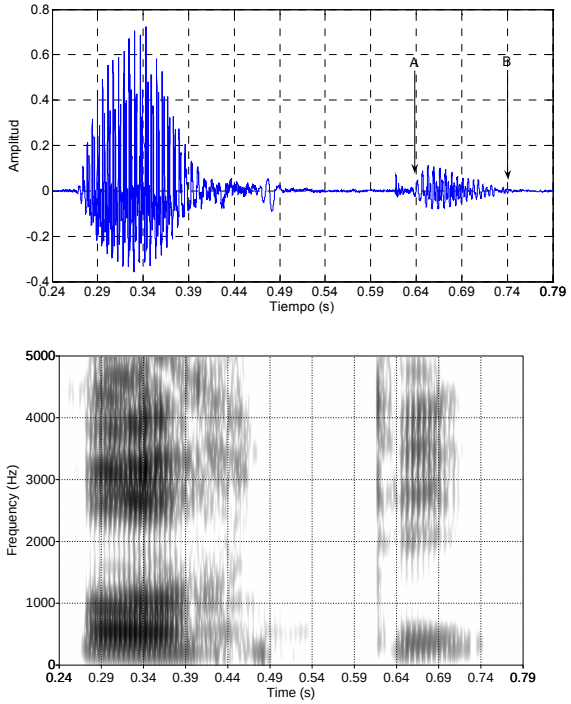


Figura 6.21: Realización [ohti] (camino)

7 TRANSICIONES VOCÁLICAS

En las secciones anteriores se discutieron las características de los fonemas del náhuatl de San Miguel Tzinacapan como aparecen en un estado idealizado. Empezaremos ahora a analizar el los sonidos de esta lengua más dinámicamente, observando lo que ocurre cuando el tracto vocal se mueve de una configuración a otra.

Se analizará el movimiento del tracto vocal y los valores de los formantes que los acompañan cuando una vocal, o un sonido similar a vocal, cambia a otra. Dado el patrón silábico del náhuatl, las transiciones de vocal a vocal serán entre sílabas. Al respecto de las aproximantes, éstas interactuarán con las vocales dentro de una sílaba o con las de sílabas contiguas.

7.1 Vocales y aproximantes

Las aproximantes (o semivocales) tienen muchas similitudes con las vocales porque ambas clases de sonidos se forman con una muy pequeña constricción de los articuladores. Los formantes de las aproximantes están claramente definidos y son predecibles porque el tracto vocal actúa como un tubo no uniforme. Puede pensarse de las aproximantes como vocales transitorias y rápidas, pero que no conservan un estado estable. Una aproximante que sigue a una vocal es llamada off-glide; una que preceda a la vocal, on-glide (…) las aproximantes están siempre adheridas a las vocales dentro de una sílaba. Cuando una aproximante precede a una vocal, la aproximante usualmente funciona como un fonema independiente. Cuando una aproximante sigue de una vocal, y se combina con ella, se forma un diptongo" [Oli93].

7.1.1 Transiciones vocal-aproximante

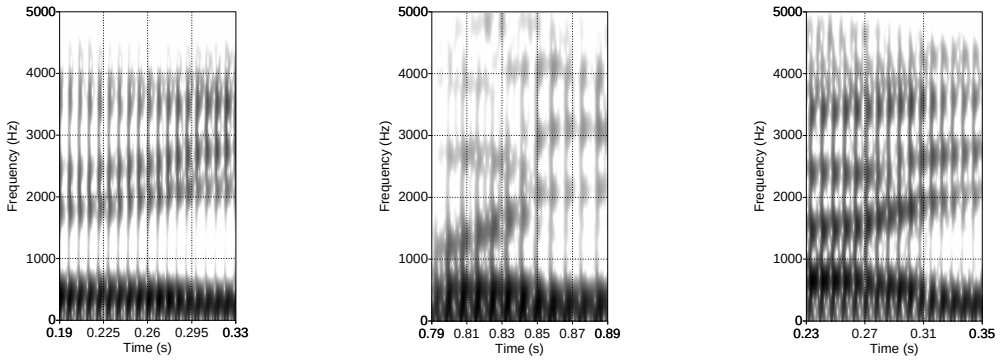
En idiomas como el inglés, las combinaciones vocal-aproximante en una misma sílaba conforman diptongos, funcionando como un solo fonema [Oli93]. En sus diptongos, entre otros rasgos, la porción de la aproximante no es tan estable como cuando es articulada como un fonema independiente, además el sonido de la aproximante es sugerido más bien a un escucha por la dirección del movimiento de sus formantes que por sus valores absolutos.

Como lo ha descrito [Oli93]: "La influencia entre vocal y aproximante es un tanto diferente que cuando estos sonidos funcionan como un solo fonema. Una diferencia básica es que la porción de transición del diptongo es más lenta y más gradual que en las transiciones vocal-aproximante. Otra diferencia es que las aproximantes independientes tienen una mayor duración que las aproximantes que funcionan como parte de diptongos".

Como se ha señalado, en el náhuatl el patrón silábico es (C)V(C), siendo la combinación más común CV. Por tanto, las transiciones vocal a aproximante se presentan a través de sílabas o fronteras de palabras. Al estudiar estas transiciones se esperará encontrar que la aproximante

funcione como un fonema independiente y no con las características acústicas propias de un diptongo.

La figura 7.1 muestra las transiciones de las cuatro vocales a la aproximante /y/, extraídas de las realizaciones [keye] (por qué), [koyotahtol] (español) y [kaya'] (todavía no).



i. [ey] ii. [oy] iii. [ay]

Figura 7.1: Transiciones [ey], [oy], [ay]

Observe en dicha figura que la vocal media anterior [e] tiene la estructura formántica más cercana a [y] por su bajo F1 y alto F2. La vocal media posterior [o] también tiene un F1 bajo, pero a diferencia de /e/ tiene un F2 muy bajo. Aunque con un F2 no tan bajo como el de [o], la vocal baja central [a] es la que presenta mayor diferencia en su estructura formántica con [y], esta vocal presenta un F1 más alto y un F2 más bajo que la aproximante.

En todos los casos F1 se mueve hacia abajo y F2 hacia arriba conforme cada vocal se mueve a la aproximante. En [o], la subida de F2 es muy grande porque el tracto vocal tiene que cambiar más para articular la aproximante. Se observa que, independientemente de cuánto tengan que moverse los articuladores, la duración de la transición es aproximadamente la misma (unos 40 ms). Por lo tanto las transiciones entre sonidos tienen que ocurrir más rápido en aquellos casos donde los articuladores se tienen que mover a una mayor distancia (como en el caso de [o] y [a]) donde la forma del tracto se somete a un cambio más gradual.

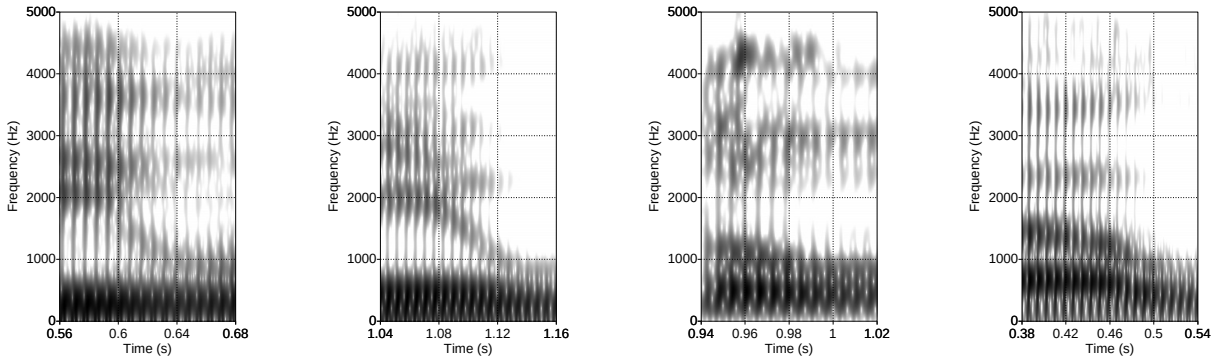
Además de observar el movimiento y dirección de los formantes durante las transiciones, es útil examinar los valores de los formantes de las vocales antes de la transición y de las aproximantes una vez que la transición ha terminado. Esto nos permitirá apreciar los efectos de las aproximantes en las vocales. La tabla 7.1 muestra los valores de los formantes de las vocales cuyos espectrogramas se mostraron en la figura 7.1; los valores F1 y F2 se muestran en la columna etiquetada *actual*. Estos valores se comparan con los valores de F1 y F2 de las mismas vocales previamente mostradas en la tabla 6.1; se anexan a la tabla 7.1 en una columna adicional con la etiqueta *esperado*. Cabe mencionar que estos datos son tomados del mismo hablante masculino; excepto en [oy], pronunciado por otro hablante masculino.

	actual				esperado	
	vocal		aproximante [y]		vocal	
	F1	F2	F1	F2	F1	F2
[e]	414	1855	295	2124	476	1798
[o]	454	1238	305	2103	484	1290
[a]	678	1537	318	2035	740	1442

Tabla 7.1 Valores de los formantes de tres vocales hacia [y]

Observamos que las vocales difícilmente están influenciadas por la aproximante; la diferencia en los valores de los formantes entre las columnas es pequeña. Por otro lado, en los estudios de [Oli93] se observó que el F2 de la parte aproximante de los diptongos eran afectados por los valores de sus kernel (el estado estable de un diptongo, localizado en la región inicial de la vocal, antes de la transición), sus valores variaban entre 1900 Hz y 2250 Hz; esto era muestra de la dependencia de las aproximantes al formar parte de un diptongo. En contraste, los valores mostrados de los formantes de la aproximante /y/ no están afectados por las vocales que la preceden. Durante la aproximante, F1 se encuentra por debajo de los 318 Hz. F2, adquiere valores entre 2035 Hz y 2124 Hz. Aunque F2 varía, permanece dentro del rango de valores de la aproximante independiente /y/.

La figura 7.2 muestra los espectrogramas de las transiciones de las cuatro vocales del náhuatl hacia la aproximante /w/, correspondientes a las realizaciones [siwat] (mujer), [šinečpalewi'] (ayúdeme), [nitahtowa] (yo hablo) y [kawehka] (cerca). El mayor cambio se encuentra en la transición de /w/ a /i/ debido a que la estructura formántica de [w] está más lejos de la estructura de [i]. Observando F1 y F2 independientemente, vemos que F1 es bastante estable en todas las transiciones pero desciende después de la vocal [a]. Debido a que el valor de F2 para la aproximante [w] es más baja que para cualquier vocal, F2 desciende en todas las transiciones. F2 desciende ligeramente desde [o], un poco más desde [a] y aproximadamente 1000 Hz desde [i]. Nuevamente, a pesar del hecho de que los formantes se mueven a diferentes distancias desde las cuatro vocales, las transiciones parecen tener la misma duración, aproximadamente unos 40 ms.



i. [iw] ii. [ew] iii. [ow] iv. [aw]

Figura 7.2: Transiciones [iw], [ew], [ow], [aw]

Los valores de los formantes de las vocales y los aproximantes de la figura 7.2 se muestran en la tabla 7.2. Los espectrogramas de [aw], [iw], corresponden al mismo hablante de la tabla 6.1. Los espectrogramas de [ew] y [ow] fueron tomados de otro hablante masculino. Previamente se constató que los dos hablantes poseen formantes vocálicos en alturas muy similares para comparar adecuadamente los resultados.

	actual				esperado	
	vocal		aproximante [w]		vocal	
	F1	F2	F1	F2	F1	F2
[i]	388	2056	341	938	341	2046
[e]	432	1967	331	756	476	1798
[o]	491	1200	378	770	484	1290
[a]	671	1450	415	834	740	1442

Tabla 7.2 Valores de los formantes de las cuatro vocales hacia [w]

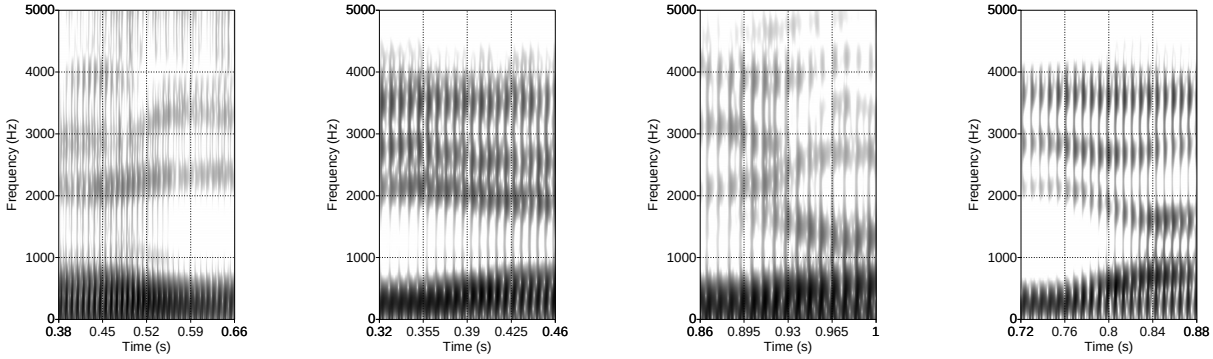
Observamos que la aproximante /w/, a diferencia de /y/, presenta una variación más grande en sus formantes. La variación de F1 va de 331 Hz a 415 Hz, lo cual no es grande. Sin embargo el rango de F2 es de 168 Hz. Aún así, a pesar de que F2 varía, permanece dentro del rango de valores para /w/. En la misma tabla observamos que los valores de las vocales no varían mucho en el contexto de /w/; sin embargo sí se aprecia una influencia de la aproximante en la vocal [e], hay una variación de 169 Hz, acercándose mucho a los valores de [i], pero la percepción en un escucha es aún de una /e/. La independencia de /w/ es evidente.

7.1.2 Transiciones aproximante-vocal

Las transiciones que se muestran en la figura 7.3 corresponden a [eyi'] (tres), [keye] (por qué), [koyotahtol] (español) y [nimayana] (tengo hambre); las transiciones se presentan dentro de sílabas. Comparando la figura 7.3 con la figura 7.1 vemos que estos espectrogramas son casi imágenes espejo. Esto indica que no hay diferencia en el movimiento articulatorio para las aproximantes si los fonemas están dentro de la misma sílaba o abarcan dos sílabas. Nuevamente vemos que las transiciones entre la aproximante /y/ y las vocales posterior y central requieren mayor movimiento de sus formantes.

El primer espectrograma de la figura 7.3 muestra el segmento [eyi]; después de los 450 ms se presenta la transición de [e] hacia la aproximante mientras que después de los 560 ms se presenta el segmento [yi]. Note que es prácticamente indistinguible el movimiento de [y] hacia [i]; hay tan poco cambio en F1 y F2 que es posible confundir esta combinación como una vocal de larga duración; perceptiblemente un escucha puede distinguir a la aproximante. Un examen más cercano del espectrograma mostraría que F2 del aproximante se vuelve más oscuro conforme [y] se mueve a [i], significando que la aproximante, teniendo mayor constricción que la vocal, tiene menor energía que ésta; esto puede apreciarse mejor al variar el rango dinámico del espectrograma en el programa Praat. Cabe mencionar que varios de los hablantes solían omitir la aproximante al pronunciar alguna palabra con la terminación /yi/. En los espectrogramas de las

vocales posteriores podemos apreciar que el F2 de la aproximante desciende en la transición; sin embargo F1 permanece muy abajo para [o], en el caso de la transición [ya] observamos que su F1 asciende al mismo tiempo que F2 desciende.



i. [eyi] ii. [ye] iii. [yo] iv. [ya]

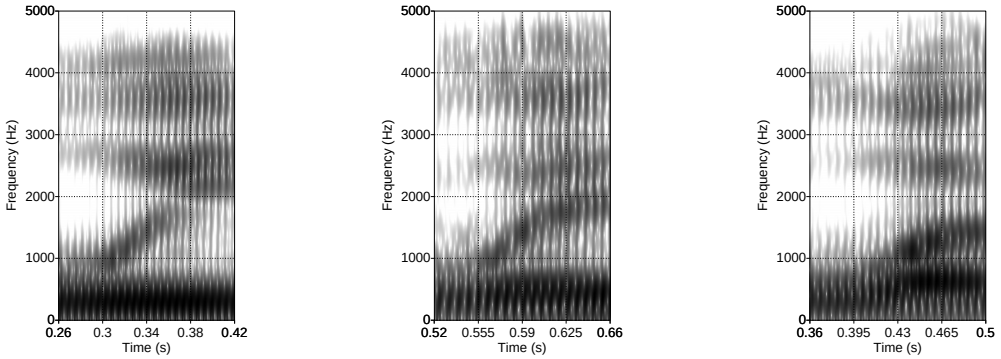
Figura 7.3: Transiciones [eyi], [ye], [yo], [ya]

La tabla 7.3 muestra los valores de los formantes para las aproximantes y las vocales de la misma manera que la tabla 7.2. Los espectrogramas de [yi] y [yo] fueron tomados de otro hablante masculino. Los valores de los formantes de la aproximante son muy similares en todos los casos. Los valores para las vocales también están muy cercanos a los valores discutidos en el capítulo anterior.

	aproximante [y]		vocal	
	F1	F2	F1	F2
[i]	282	2320	298	2354
[e]	293	2120	484	1900
[o]	303	2100	471	1261
[a]	305	2128	678	1596

Tabla 7.3: Valores de formantes de [y] y vocales

Los espectrogramas de la figura 7.4 que muestran las transiciones de la aproximante /w/ a tres vocales son también un reflejo del movimiento de formantes vocal a aproximante (ver la figura 7.2). Son extractos de las realizaciones [owi] (difícil), [kawehka'] (cerca) y [yowak] (noche). En la transición de [w] a [i], F1 permanece relativamente constante ya que ambos sonidos poseen un bajo F1, pero F2 toma un movimiento hacia arriba para alcanzar el valor de la [i], aproximadamente 2200 Hz. En los otros casos F1 también permanece relativamente constante y F2 se eleva gradualmente. De igual manera se observa que los formantes F2 y F3 de [w] tienen menor energía que aquellos de las vocales. En la transición de [w] a [a] F1 y F2 están muy cercanos.



i. [wi] ii. [we] iii. [wa]

Figura 7.4: Transiciones [wi], [we], [wa]

Resumiendo, las transiciones nos muestran aproximantes independientes al ser combinadas con vocales, no fueron hallados indicios de diptongos. Las transiciones entre vocales y aproximantes no parecen afectar la estructura formántica de cada uno de estos sonidos. No parece tampoco haber diferencias si las transiciones se presentan en una misma sílaba o en su frontera con otras. La duración de las transiciones entre vocales y aproximantes es relativamente consistente a pesar del movimiento requerido para articular sus sonidos. Cuando los articuladores se tiene que mover una mayor distancia, por ejemplo entre /i/-/w/, los formantes se mueven más rápido que cuando los sonidos tienen una estructura formántica similar entre sí, por ejemplo entre /i/-/y/.

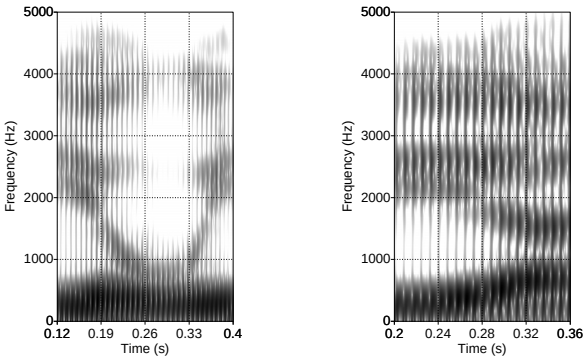
7.2 Transiciones vocal-vocal

Cuando las vocales están adyacentes dentro de una sola palabra, estas transiciones generalmente son continuas y suaves. Frecuentemente, cuando hay transiciones vocal a vocal entre palabras, a través de sus fronteras, estas transiciones son también continuas; pero esto no se presenta siempre. Se examinarán estas transiciones en la siguiente sección.

7.2.1 Transiciones continuas

La figura 7.5 muestra las transiciones de /i/ a las vocales baja central y media posterior; provienen de las realizaciones [kiowit] (lluvia) y [tiamiki] (tienes sed). Cabe señalar que la transición /ie/ no se halló en la bibliografía consultada, pero sí se encontró frecuentemente /ye/, esto nos indica un posible rasgo característico del náhuatl de San Miguel Tzinacapan.

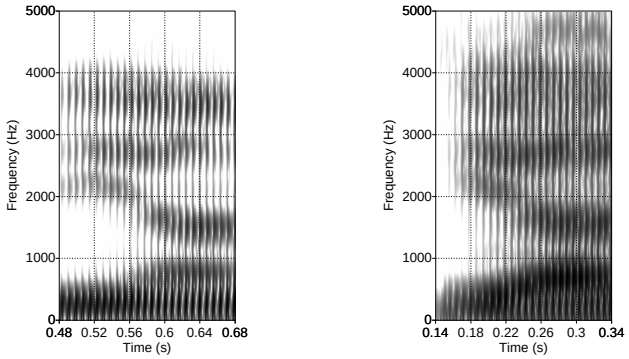
Las transiciones de esta figura son muy parecidas a aquellas entre la aproximante /y/ y las vocales (vea la figura 7.3). No es sorpresiva esta similitud ya que ambos sonidos son muy similares en su movimiento formántico. Como se ha dicho, una diferencia entre aproximantes y vocales es la baja energía de aquellas. Otra diferencia es la duración del estado estable de ambos sonidos, siendo más largo para las vocales. La transición de [i] a [a] muestra tales diferencias. El espectrograma de la derecha muestra el segmento [iowi]; se observan varios detalles: La transición de la aproximante con las vocales vecinas es distinguible ya que su F2 es de baja energía (el segmento entre 260 ms y 330 ms); además la transición [i]-[o] (entre 120 ms y 260 ms) es más lenta que la transición [w]-[i] (entre 330 ms y 400 ms).



i. [iowi] ii. [ia]

Figura 7.5: Transiciones [io], [ia]

Con el fin de contrastar las diferencias entre las aproximantes y las vocales, la figura 7.6 muestra el segmento [ia] de la realización [niamiki] (tengo sed) y el segmento [ya] de la realización [yalwa] (ayer). La vocal [i] y la aproximante [y] son mostrados en toda su realización. Se observa que la vocal mantiene un estado estable (80 ms) a diferencia de la aproximante, cuyo F2 desde el inicio no mantiene una posición estable y desciende hacia [a].

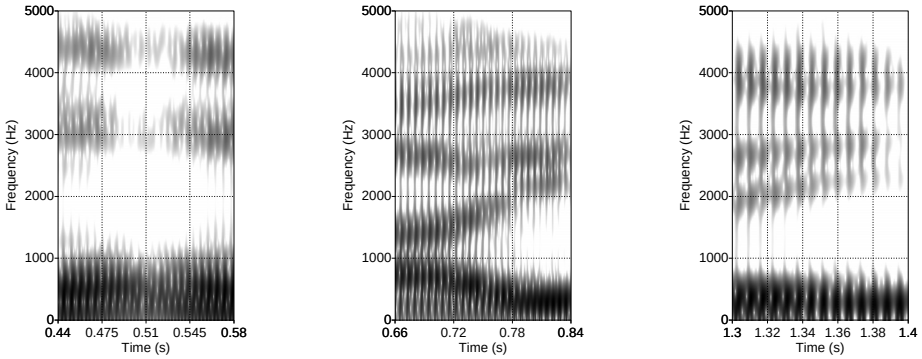


i. [ia] ii. [ya]

Figura 7.6: Comparación de transiciones [ia], [ya]

Para brindar más ejemplos de transiciones, la figura 7.7 muestra las transiciones desde [o] a otro fono [o] en el caso de la frontera entre las palabras [amo owi] (fácil), la transición de [a] a [i] de la frontera entre sílabas de la palabra [ait] (quizás), por último se presenta la transición de [e] a [i] de la palabra [tei]. En esta última transición existe un mínimo movimiento de los formantes, esto es debido a que hay un mínimo movimiento requerido a los articuladores, lo cual resulta en un cambio pequeño de la configuración del tracto vocal. La transición de [a] a [i] es como un reflejo de aquella mostrada en la figura 7.6. El espectrograma de la transición de [o] a [o] nos revela que la amplitud de la forma de onda es menor durante la transición, de esta manera el hablante marca que hay dos palabras y no una sola con una vocal de muy larga duración; debido a que la configuración del tracto vocal no cambia la estructura formántica es la misma.

Podemos concluir que la duración de las transiciones vocal a vocal permanece relativamente constante. En la producción de una vocal a otra, los articuladores no tardan demasiado tiempo en realizar sus movimientos, moviéndose más rápido cuando el cambio es mayor.



i. [oo] ii. [ai] iii. [ei]

Figura 7.7: Transiciones [oo], [ai], [ei]

7.2.2 Transiciones discontinuas: oclusivas glotales

En ocasiones los hablantes marcan la presencia de vocales adyacentes entre palabras apretando las cuerdas vocales para reducir el flujo de aire a través de ellas y así disminuir la vocalización. A este fenómeno se le llama oclusiva glotal [Oli93], un ejemplo es la interjección ah-ah.

Recuerde que la oclusiva glotal, indicada como / ' /, no es un fonema en el náhuatl de Tzinacapan y por lo general se utiliza para señalar el fin de una palabra.

La figura 7.8 muestra la transición entre los fonos [o] de [amo owi] (fácil). La constricción de la glotis se presenta en el intervalo (455 ms - 550 ms), el cual se señala con flechas en la forma de onda. Observe que en la oclusión hay un cambio abrupto en la periodicidad de la señal y su amplitud disminuye. La oclusiva glotal es identificada por la amplitud reducida y el evidente cambio de periodicidad en la forma de onda y el espectrograma (al que se le ha aplicado preénfasis). Debido a que las oclusivas glotales se forman en la glotis, la posición por el movimiento de los articuladores no es afectada. Por lo tanto no hay una estructura formántica o movimiento de formantes en particular asociados con este fenómeno. “Las oclusivas glotales no afectan cualquier movimiento de formantes esperado si la oclusiva glotal no estuviera presente. El uso y lugar de las oclusivas glotales es bastante impredecible e idiosincrásico para pronunciaciones particulares y hablantes individuales. Es imposible generalizar un patrón particular para su existencia” [Oli93].

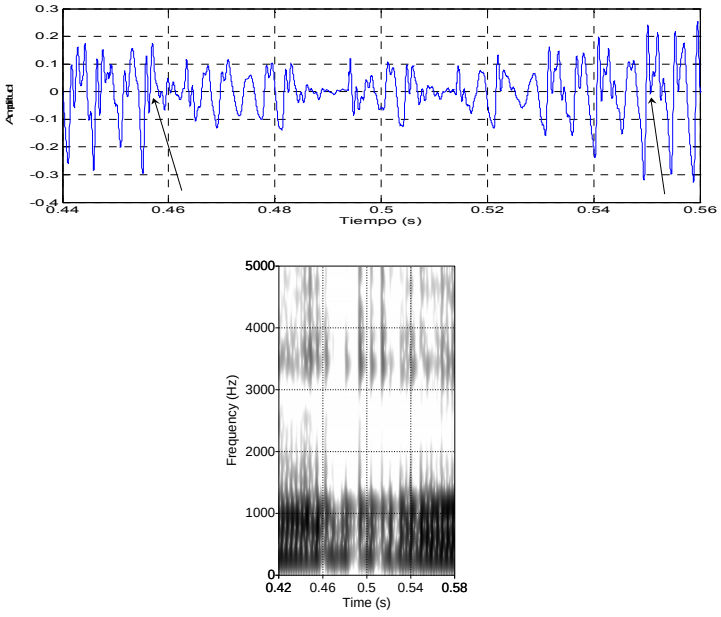


Figura 7.8: Transición /oo/

8. TRANSICIONES OBSTRUYENTES Y VOCALES

Las consonantes oclusivas, fricativas y africadas pertenecen a la clase de fonemas obstruyentes. En las siguientes secciones se analizarán las transiciones entre estos sonidos y las vocales para examinar cómo interactúan dinámicamente. En particular queremos ver si el loci de estos fonemas son influenciados por el contexto, y de ser así, en qué grado y medida.

8.1 Oclusivas y vocales

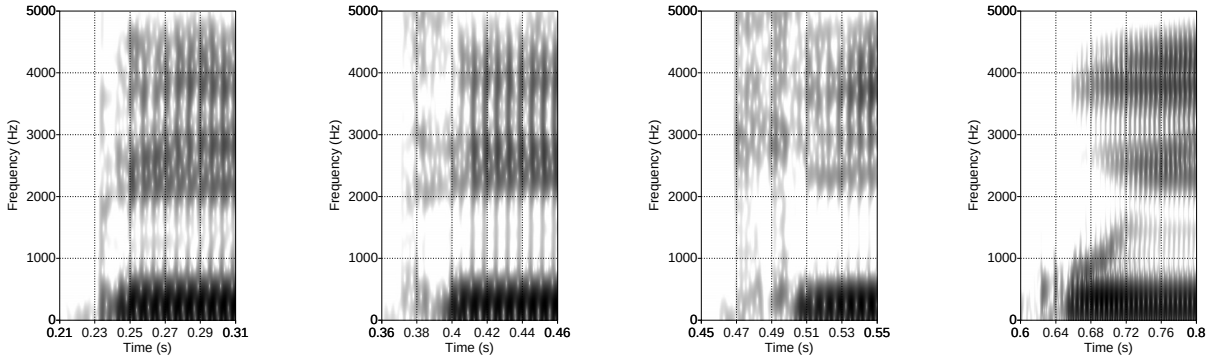
En esta sección se examinará cómo las oclusivas y las vocales interactúan dinámicamente. En particular queremos ver si los loci de estos fonemas están influenciados por contexto, y en ese caso, en qué grado. El análisis mediante espectrogramas de estos sonidos en transición debe proporcionar evidencia de efectos coarticulatorios.

8.1.1 Transiciones oclusiva-vocal

Aunque los sonidos oclusivos sean esencialmente ruido, por efectos de la coarticulación es posible hallar indicios de formantes en la transición a un sonido sonoro. No es fácil ver la trayectoria de los formantes en las transiciones de las oclusivas sordas a las vocales. En ocasiones solamente es posible inferir las transiciones por el movimiento de los formantes de las vocales en la frontera con las oclusivas.

Las figuras 8.1 a 8.4 muestran los espectrogramas de las transiciones de las oclusivas a las vocales. Es posible distinguir el sonido oclusivo del vocálico. Las figuras 8.1 y 8.2 muestran las oclusivas moviéndose a las vocales anteriores alta y baja.

En la figura 8.1 se presentan segmentos de las realizaciones [pili'] (niño), [tit] (fuego), [kili] (le dijo) y [mak^{w}il] (cinco). F2 muestra un movimiento hacia arriba para [p]-[i] y [t]-[i], un mayor movimiento se aprecia en la bilabial. En el caso [k]-[i] no se aprecia algún movimiento de F2 en la frontera de la vocal. Por último, en la transición [k^{w}]-[i], se observa que el F2 de la parte aproximante (similar a /w/) de la oclusiva trata de alcanzar el F2 de la vocal. Respecto a F1 no se aprecia un movimiento en los cuatro casos.



i. [pi] ii. [ti] iii. [ki] iv. [k^{w}i]

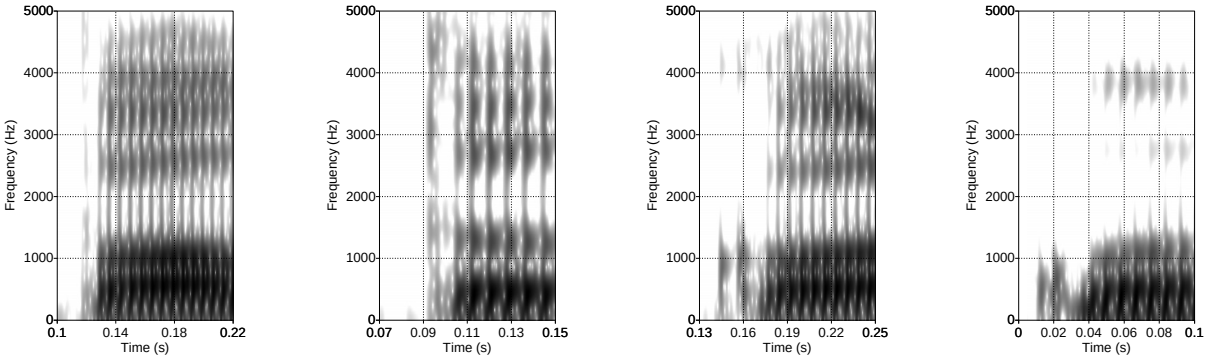
Figura 8.1: Transiciones [pi], [ti], [ki], [k^{w}i]

Las figuras mostradas en 8.2 corresponden a segmentos de [čikte] (pájaro) y [keye] (por qué). Difícilmente se aprecia algún movimiento de F1 y F2 durante la transición [t/]-[e]. En el caso de la transición velar se aprecia un **velar pinch**, este es un fenómeno donde F2 y F3 están tan cerca entre sí que casi se combinan en un solo formante, es un fenómeno común en las oclusivas velares. Aparentemente hay un muy ligero movimiento hacia arriba de F1 y hacia abajo de F2 durante la transición velar. No se contaba en el corpus con realizaciones que incluyeran las transiciones /pe/, /k^{w}e/.

i. [te] ii. [ke]

Figura 8.2: Transiciones [te], [ke]

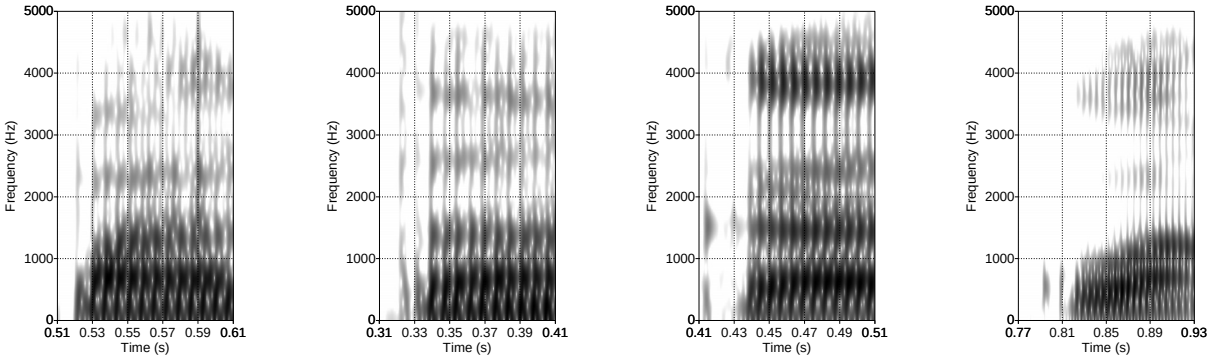
Las transiciones de las oclusivas hacia la vocal posterior /o/ se muestran en la figura 8.3, fueron tomadas de las realizaciones [pokti] (humo), [totonik] (caliente), [kolot] (alacrán) y [k^{w}owit] (árbol), esta última realización fue tomada de otro hablante masculino. Durante la transición de [p] a [o] no parece haber movimiento en los formantes. Respecto a la transición de [t] a [o] podemos observar un ligero descenso de F2 y F3. Sobre la transición de [k] a [o] no ocurre el velar pinch, de hecho F2 permanece constante. Finalmente, en la transición [k^{w}]-[o] es posible notar que la transición de F2 desde la parte aproximante de [k^{w}] hacia el F2 de [o] es más sutil; sin embargo es posible identificar esta transición con respecto a [ko] debido a que el F2 de [k^{w}] muestra un ascenso en su transición a [o].



i. [po] ii. [to] iii. [ko] iv. [k^{w}o]

Figura 8.3: Transiciones [po], [to], [ko], [k^{w}o]

Las transiciones de las oclusivas a la vocal central se muestran en la figura 8.4, pertenecen a las realizaciones [tapalol] (comida), [ta:l] (tierra), [takat] (hombre) y [titak^{w}a] (tu comes). Se observa que F2 tiene un movimiento hacia arriba en la transición [p]-[a]. Lo mismo ocurre en la transición [t]-[a], donde el movimiento es muy ligero. Es en la transición [k]-[a] donde F2 muestra un movimiento constante. En la transición [k^{w}]-[a] se observa el movimiento ascendente del F2 de [k^{w}], es posible distinguir que dicho formante posee menor energía que el F2 de la vocal siguiente cuando ésta se encuentra en un estado estable. En los tres casos F1 muestra un movimiento hacia arriba durante la transición.



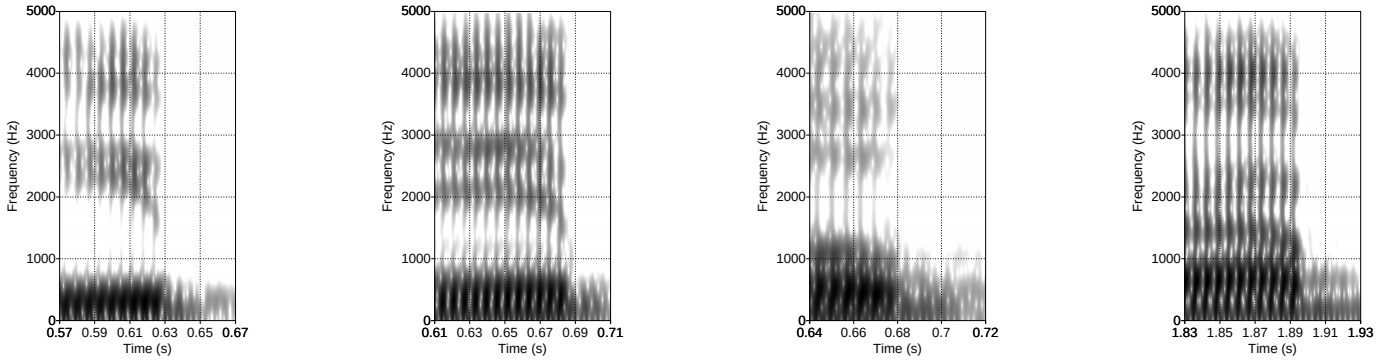
i. /pa/ ii. /ta/ iii. /ka/ iv. /k^{w}a/

Figura 8.4: Transiciones /pa/, /ta/, /ka/, /k^{w}a/.

8.1.2 Transiciones vocal-oclusiva

Las figuras 8.5, 8.6 y 8.7 muestran la transición de las vocales a las oclusivas sordas. Es notable que el movimiento de los formantes durante la transición sea muy similar a las transiciones oclusiva-vocal. Recordemos que debido a que la zona de cierre de un sonido oclusivo prácticamente no muestra energía, lo cual reduce información para realizar observaciones, el movimiento de los formantes es sugerido por el movimiento de éstos en la frontera con las vocales. También hay información en la zona de baja frecuencia durante el cierre.

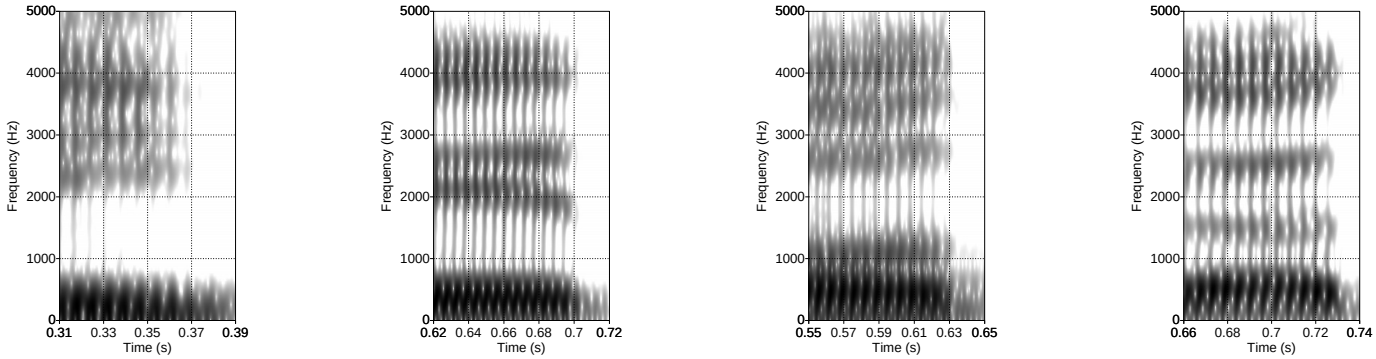
La figura 8.5 muestra la transición desde las vocales hacia /p/, se consideraron las realizaciones [yawipta'] (antier), [nepa'] (allá), [šolopi] (mentiroso) y [ɕinakapan] (Tzinacapan). En la transición desde [i], F1 permanece constante mientras que F2 muestra un rápido movimiento descendente antes de la región de cierre. Lo mismo ocurre en la transición desde [e]. La transición entre [o] y [p] muestra un movimiento descendente mínimo para F1 y F2; lo mismo se aprecia para la transición [a]-[p].



i. [ip] ii. [ep] iii. [op] iv. [ap]

Figura 8.5: Transiciones [ip], [ep], [op], [ap]

La figura 8.6 muestra a las cuatro vocales moviéndose a la oclusiva alveolar sorda /t/, las realizaciones correspondientes son [titakwa'] (tú comes), [miket] (muerto), [kolot] (alacrán) e [ilwikatik] (azul). En general, durante la transición, los formantes mantienen un estado estable excepto en [e]-[t], donde F2 presenta un marcado movimiento hacia abajo. Es posible apreciar en las transiciones [i]-[t] y [a]-[t] que sus F1 presentan un leve movimiento hacia abajo.

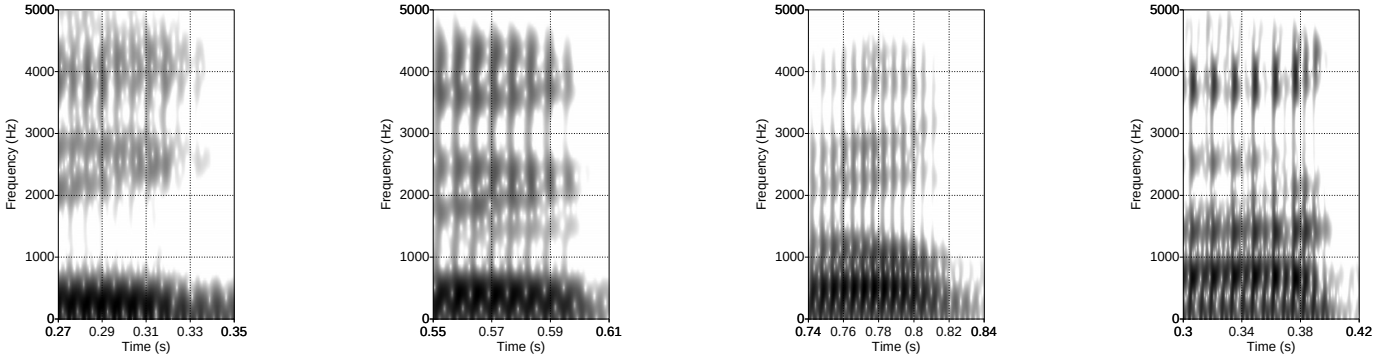


i. [it] ii. [et] iii. [ot] iv. [at]

Figura 8.6: Transiciones [it], [et], [ot], [at]

Las transiciones de las vocales a la velar oclusiva sorda /k/ se muestran en la figura 8.7, los ejemplos vienen de las realizaciones [tikneki'] (tú quieres), [iɕonteko] (su cabeza), [šošoktik] (verde) y [takat] (hombre). Observe que el movimiento de F1 es hacia abajo en las vocales [i] y [a]. Se aprecia un estado estable en las restantes vocales. El velar pinch es apreciable en las

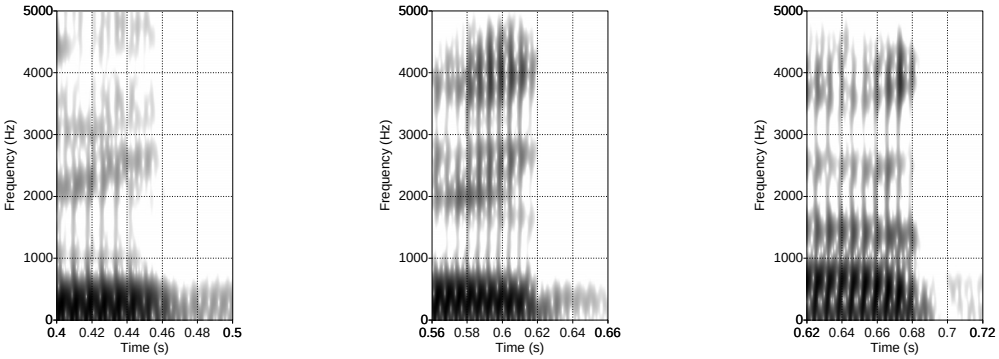
transiciones de las vocales [i]-[e]. Aunque por definición no se trata de un velar pinch (porque involucra a F2 y F3), es de llamar la atención que se unan F1 y F2 en la frontera de la vocal [o]. En [a] hay un descenso de F3 aproximándose bastante a F2 aunque sin llegar a formar un velar pinch. Note que la altura del velar pinch es proporcional a la posterioridad de las vocales, más posterior la vocal, a más baja frecuencia se ubica el velar pich.



i. /ik/ ii. /ek/ iii. /ok/ iv. /ak/

Figura 8.7: Transiciones de /ik/, /ek/, /ok/, /ak/

Como último bloque, en la figura 8.8 se muestran las transiciones de las vocales a la velar oclusiva labial sorda /k^{W}/, estos ejemplos vienen de las realizaciones [nehnik^{W}a'] (yo como), [nisek^{W}i] (tengo frío), y [titak^{W}a'] (tu comes). El hablante de la primera palabra es distinto al de los otros dos ejemplos de dicha figura, ambos son masculinos. Observe que los espectrogramas son similares a los de la figura 8.7. El velar pinch es apreciable en la transición desde [i], mientras que es menos aparente en la transición desde [e]. En [a] hay un ligero descenso de F1 y F2 al final de la misma, de hecho sus formantes conservan un estado ciertamente estable, sin rasgos de existir un velar pinch. En las otras dos vocales, el F1 es estable. No se contaba en el corpus con la transición /ok^{W}/.



i. [ik^{W}] ii. [ek^{W}] iv. [ak^{W}]

Figura 8.8: Transiciones [ik^{W}], [ek^{W}], [ak^{W}]

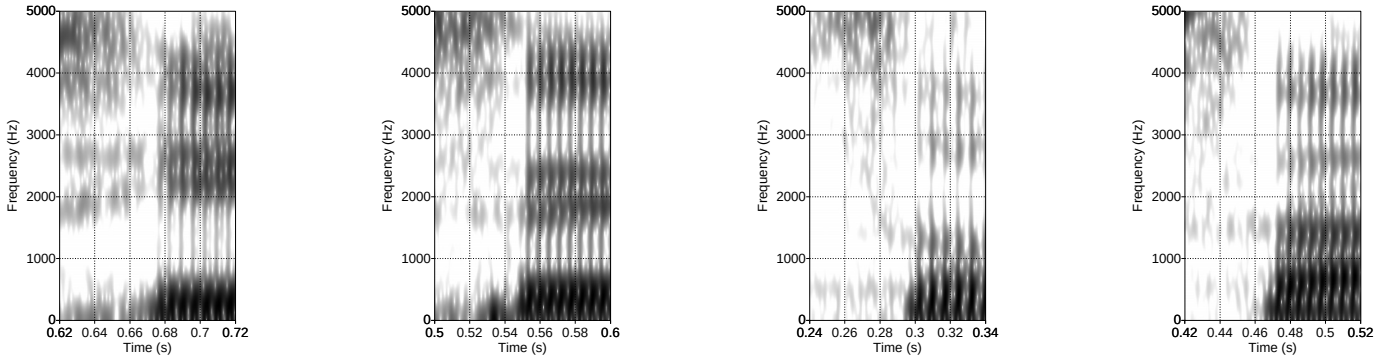
8.2 Fricativas y vocales

Los sonidos oclusivos son únicos porque el sonido ocurre en el punto de constricción del tracto vocal más que en la glotis, la turbulencia del sonido es debida a que la constricción es estrecha. Debido a que la fuente de sonido está en la constricción, el tracto vocal detrás del punto de constricción (la cavidad posterior) tiene menos influencia sobre la acústica del sonido que a parte del tracto que está adelante de la constricción. La influencia de la cavidad posterior en el espectro del sonido depende del tamaño de la apertura de la constricción. Ya que el ruido es filtrado primordialmente por el área al frente de la constricción (la cavidad anterior), el espectro del ruido está principalmente controlado por cualquier cámara resonante existente en la cavidad anterior. Las fricativas alveolares y alveopalatales, producidas atrás de la cavidad oral, tienen cavidades frontales que son considerablemente más pequeñas que sus cavidades posteriores. En consecuencia estos sonidos tienen patrones resonantes inusuales y no exhiben formantes.

Los lugares de articulación para algunas fricativas están relacionados con los de las oclusivas. En el náhuatl, la alveolar /s/ se produce en la misma área que la oclusiva alveolar; la alveopalatal /š/ se articula en el área entre las oclusivas alveolar y velar. Ya que conocemos los valores de los loci de las oclusivas y hemos visto cómo éstas interactúan con algunas vocales representativas, debemos poder anticipar cómo es que las fricativas producidas en la misma área del tracto vocal interactuarán con tales vocales. Las similitudes entre las transiciones de las oclusivas y las fricativas son debidas a su vez a la similitud en los lugares de articulación; las diferencias en la manera como estos fonemas interactúan son el resultado de las diferentes maneras de producción sonora.

8.2.1 Alveolares

Las fricativas alveolares se producen constriñendo el tracto vocal con la lengua en los alveolos, lo suficiente para producir fricción pero sin impedir por completo el flujo de aire. La figura 8.9 muestra la fricativa alveolar sorda /s/ en su transición a las vocales. Las realizaciones correspondientes son [siwat] (mujer), [sesek] (frío), [tasohkamatik] (gracias) y [aksa'] (alguien).



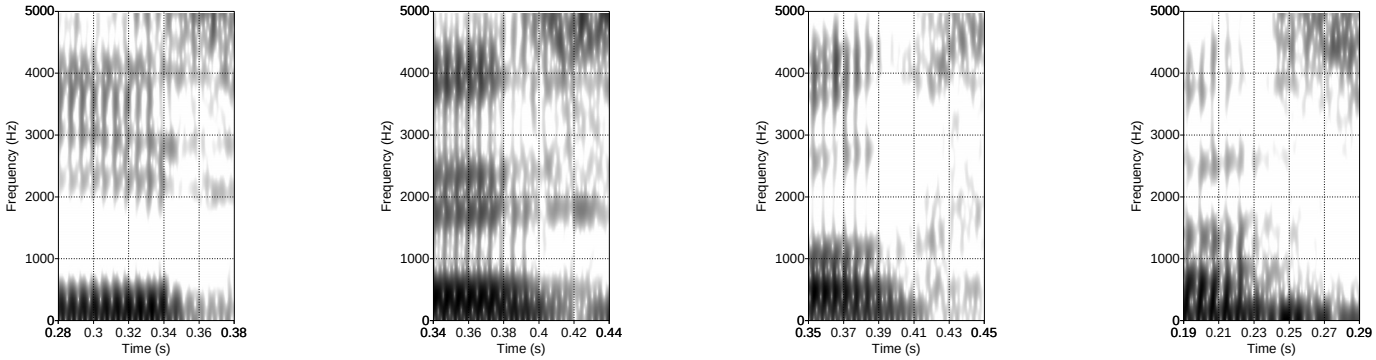
i. [si] ii. [se] iii. [so] iv. [sa]

Figura 8.9: Transiciones [si], [se], [so], [sa]

Los espectrogramas nos revelan que hay una coarticulación anticipada durante la producción de las fricativas alveolares. Por eso podemos apreciar una estructura formántica en los fonos fricativos [s], de hecho observamos un mayor movimiento de formantes durante la fricativa que en la vocal. Se observa que al inicio de las vocales sus formantes no presentan algún movimiento. Podemos notar que, descontando la energía a baja frecuencia, la ubicación del borde inferior de la región de la fricativa depende de F2 de la vocal. Antes de la vocal [i] hay mucha anticipación en la coarticulación. Por el contrario, antes de las vocales [e] y [a] no hay necesidad de coarticulación anticipada para F2 porque el valor de este formante para el locus alveolar es aproximadamente el mismo que el valor target para dichas vocales, indicando que la lengua está en una posición adecuada para producir el mismo valor F2. En la transición [s]-[o] los formantes muestran una coarticulación anticipada indicando que el cuerpo de la lengua puede prepararse para la vocal media posterior aún con la constricción frontal alta.

Es posible apreciar un ligero movimiento hacia arriba de F1 y hacia abajo de F2 de la vocal [o] en su frontera con la fricativa. Para articular la fricativa [s] la lengua mantiene una posición baja, también conserva esa posición cuando se articula la vocal [e], es por eso que el comportamiento formántico de la fricativa nos muestra que los formantes tienen un comportamiento estable. La transición hacia [o] muestra a F2 descendiendo.

La figura 8.10 muestra las transiciones en dirección contraria, desde las vocales hacia las fricativas, provenientes de las realizaciones [istak] (blanco), [sesek] (frío), [kostik] (amarillo) y [tasohkamatik] (gracias). Durante la transición, F1 muestra un ligero descenso para la vocal [i]. Lo mismo ocurre con la vocal [o]. En el caso de la vocal [e] su F1 muestra un movimiento estable. F2 permanece constante en todas las vocales salvo para la transición [o]-[s], donde se observa un movimiento hacia arriba.



i. [is] ii. [es] iii. [os] iv. [as]

Figura 8.10: Transiciones [is], [es], [os], [as]

Aunque la lengua se utiliza para generar la constricción en las fricativas alveolares, hemos apreciado evidencia de una coarticulación anticipada. Para observar la coarticulación durante la producción de las fricativas, daremos un ejemplo de en un contexto intervocálico. La figura 8.11 muestra el segmento de voz [osi], de la realización [nosiwaw]. La primera vocal tiene un F2 a una

frecuencia más baja que su correspondiente de la segunda vocal; observamos que, a la altura de cada F2, en la región de la fricativa existe un movimiento hacia arriba. Esto es evidencia de movimiento articulatorio durante la fricativa. Este movimiento es lento; [Oli93] agrega que es también "más limitado ya que la lengua no se puede mover tan libremente durante una constricción alveolar".

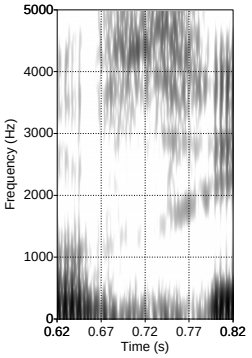
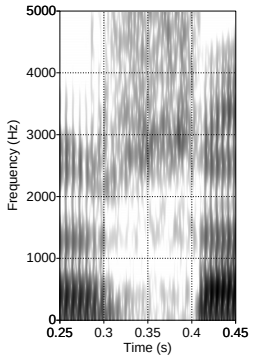


Figura 8.11: Segmento [osi]

8.2.2 Alveopalatales

La fricativa alveopalatal /š/ se articula al tocar la lámina de la lengua la zona posalveolar y el paladar, el ápice de la lengua se levanta y forma una constricción. [Oli93] indica que cuando la constricción para producir un sonido lingüístico se presenta en el velo, F2 y F3 a menudo se mezclan para formar un velar pinch. [Oli93] señala además que "los sonidos articulados cerca del velo pueden presentar un velar pinch para algunas vocales. Sin embargo, si no hay un velar pinch, F2 debe ser debe ser aún de frecuencia alta para los sonidos alveolapalatales que para las otras fricativas. Además las transiciones hacia adentro y afuera de las vocales pueden ser más lentas para estas fricativas".

La figura 8.12 muestra las transiciones [o]-[š] y [š]-[o] en un solo espectrograma, éste corresponde al segmento [ošo] de la realización [šošoktik] (verde). Durante la región de la fricativa el espectrograma muestra similitudes con el de la alveolar /s/. Se observa durante la fricativa una cierta estructura formántica debida por el efecto de coarticulación. La región de mayor energía en la fricativa se produce arriba del F3 de las vocales que la rodean. Nos centraremos ahora en el segmento [šo], F2 al inicio de la vocal tiene un valor de 1375 Hz mientras que en el caso de la transición [so] mostrada anteriormente su F2 tiene un valor de 1313 Hz. Se confirma lo observado por [Oli93], aunque la diferencia es pequeña.



/ošo/

Figura 8.12: Transiciones [šo] y [oš] del segmento [ošo]

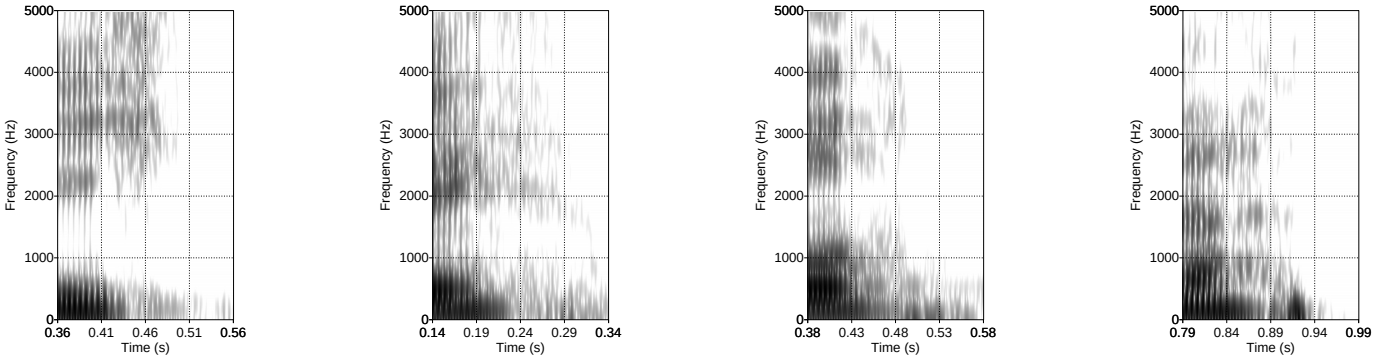
En la figura 8.13 presentamos los espectrogramas de los segmentos [siw] y [ši] de las realizaciones [siwat] (mujer) y [šinečpalewi'] (ayúdeme), fueron pronunciadas por una voz femenina, se trata de la misma mujer. Observemos en ambos espectrogramas el inicio de la vocal /i/; el valor de cada F2 es de 2593 Hz y de 2758 Hz al coarticularse con la alveolar y con la alveopalatal respectivamente. Nuevamente F2 para el caso alveopalatal se localiza a una frecuencia mayor que el F2 del caso alveolar.

i. [siw] ii. [ši]

Figura 8.13: Transiciones [s] y [š] a la vocal [i]

8.2.3 Glotal

La figura 8.14 muestra la transición de la fricativa glótica /h/ hacia las vocales. Las realizaciones correspondientes son [ihkon] (así), [teh] (tú), [ohti] (camino) y [amo nitahtowa koyotahtol] (no hablo español), el extracto [ah] presentado proviene de la palabra [nitahtowa].



i. [ih] ii. [eh] iii. [oh] iv. [ah]

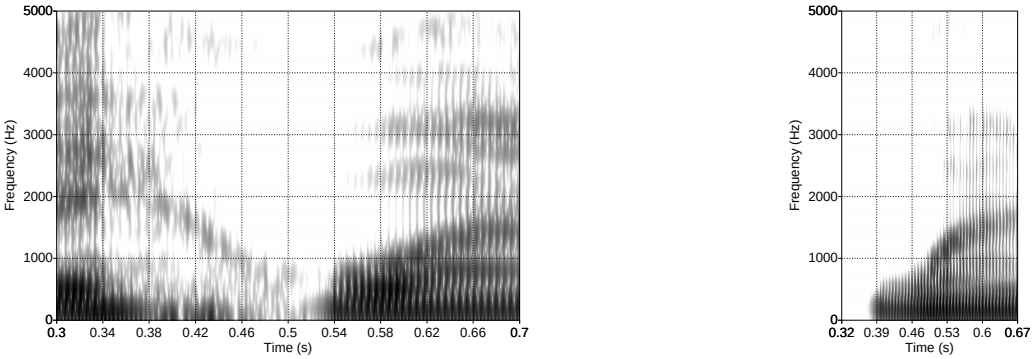
Figura 8.14: Transiciones [ih], [eh], [oh], [ah]

Los espectrogramas nos permiten identificar fácilmente a la vocal o a la fricativa que la acompaña. Además de un comportamiento ruidoso debido a la fricción del sonido, es posible apreciar en los espectrogramas una estructura cuasi-formántica en cada fono [h].

En los anteriores espectrogramas se podría llegar a la conclusión que la estructura cuasi-formántica de la aspiración está en función del sonido sonoro que la precede. Sin embargo no es del todo cierto, también la naturaleza del sonido siguiente (y en consecuencia, de la sílaba siguiente) influye en la fricativa glotal. En la figura 8.15 se presentan los espectrogramas de los segmentos de dos realizaciones: [tehwan] (nosotros) y [wan] (y). Gracias al movimiento de la estructura cuasi-formántica de la fricativa glotal en el segmento [ehwa] es posible apreciar que durante su realización los órganos articulatorios están libres para articular el sonido aproximante [w]. Sin embargo surge la pregunta, ¿la región de debilitamiento del fono [w] es parte de la fricativa glotal o de la aproximante?

Por otra parte, en el espectrograma del segmento [wa] (figura derecha) es posible apreciar la aproximante [w] ubicada en inicio de palabra, note que su F2 parte de la posición más baja. Este espectrograma es muy parecido al de la izquierda en la región siguiente al de la fricativa e incrementa el cuestionamiento si la región de debilitamiento mencionada efectivamente forma parte de la aproximante.

Se presenta este caso porque a veces no es posible determinar si una región sonora pertenece al sonido precedente o siguiente. Baste decir que la coarticulación interviene de una manera muy importante en el segmento [ehwa]; gracias a ella, durante la producción de la fricativa glotal, el tracto vocal articula la aproximante [w]; note que conforme transcurre la fricativa su presencia ruidosa se va debilitando. Gracias al fenómeno silábico, hay un momento donde [w] en posición onset contiene todos sus rasgos (es el instante donde el fono [w] contiene más energía). Al final, tenemos la combinación de fenómenos coarticulatorios y silábicos.



i. [ehwa] ii. [wa]

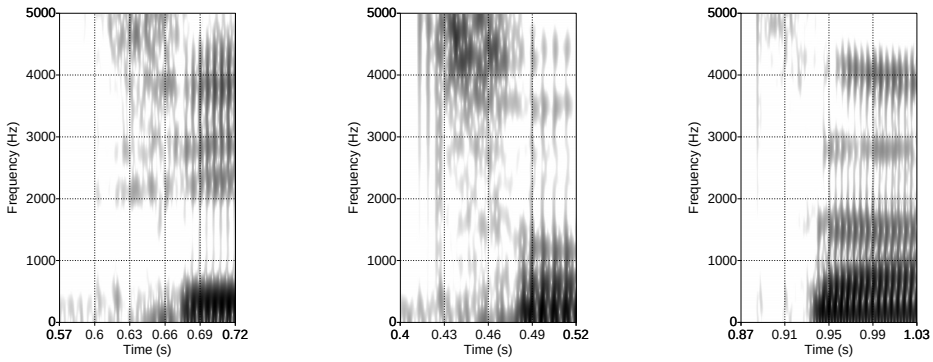
Figura 8.15: Segmentos de [tehwan] (nosotros) y [wan] (y)

8.3 Africadas y vocales

En este último apartado se revisarán las transiciones entre los dos sonidos africados del náhuatl con las vocales. No se contaba con todas las transiciones entre africadas y vocales en el corpus, por lo que se presentan las mayores transiciones posibles.

8.3.1. Alveolar

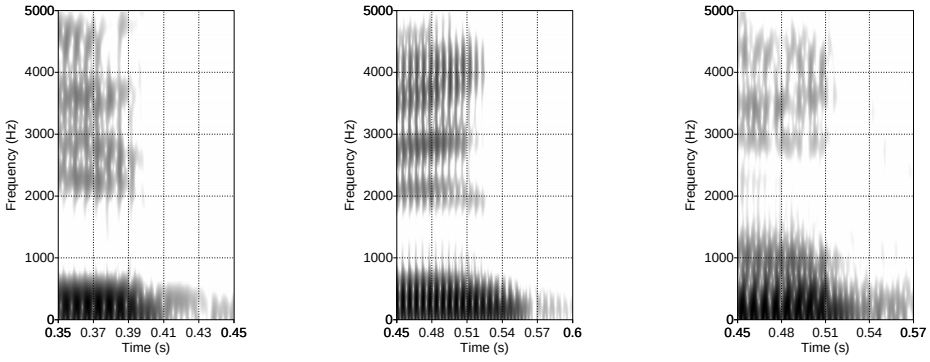
En la figura 8.16 se muestran las transiciones entre la africada alveolar /¢/ y las vocales /i/, /o/, /a/. Las realizaciones de las que se toman estos segmentos son: [mi¢ili] (te dijo), [i¢onteko] (su cabeza) y [keniwtimono¢a'] (¿cuál es tu nombre?). A pesar de la característica región fricativa de esta africada, los espectrogramas nos muestran movimientos formánticos en las transiciones muy similares a las halladas en [ti], [to], [ta], presentadas en las figuras 8.1, 8.3 y 8.4. Lo anterior se puede explicar debido a /¢/ y /t/ tienen el mismo punto de articulación. De hecho, la región fricativa de la africada nos facilita observar las transiciones, observe por ejemplo el espectrograma de [¢i]. Mientras que en la transición de [¢i] F2 muestra un movimiento hacia arriba, en [¢o] F2 presenta un movimiento hacia abajo; esto está en correspondencia con las transiciones [ti], [to] ya presentadas. Respecto a la transición [¢a], y al contrario de la transición [ta], F2 muestra un ligero movimiento hacia abajo, aunque el F1 de ambas transiciones presentan un claro movimiento hacia arriba.



i. [¢i] ii. [¢o] iii. [¢a]

Figura 8.16: Transiciones [¢i], [¢o], [¢a]

En la figura 8.17 se aprecian las transiciones de tres vocales a /¢/, fueron tomadas de [i¢onteko] (su cabeza), [ime¢] (su pierna), [no¢onteko] (mi cabeza). Se observa que F2 desciende en la transición, mientras que F1 se mantiene es estado estable.



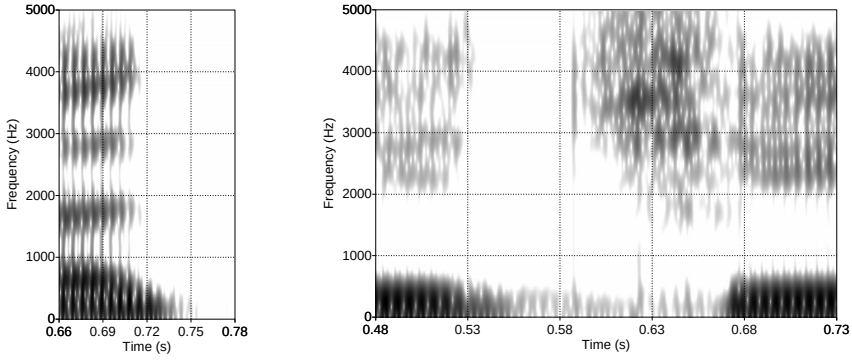
i. [i¢] ii. [e¢] iii. [o¢]

Figura 8.17: Transiciones [i¢], [e¢], [o¢], [a¢]

8.3.2. Alveopalatal

Por último se analizan algunas transiciones entre vocales y la africada alveopalatal. La figura 8.18 muestra dos espectrogramas, el segmento [ač] fue extraído de la realización [kanači'] ("cuánto", para formular preguntas); el segmento [iči] proviene de la realización [čičiltik] (rojo).

En la transición [ač] se observa que F2 y F3 se mueven hacia arriba mientras F1 toma un movimiento descendente. En el caso de las transiciones [ič] y [či] agrupadas en un solo espectrograma podemos observar que sus formantes F1, F2 y F3 mantienen un estado estable.



i. [ač] ii. [iči]

Figura 8.18: Transiciones [ač], [iči]

9. TRANSICIONES CONSONANTES SONORANTES Y VOCALES

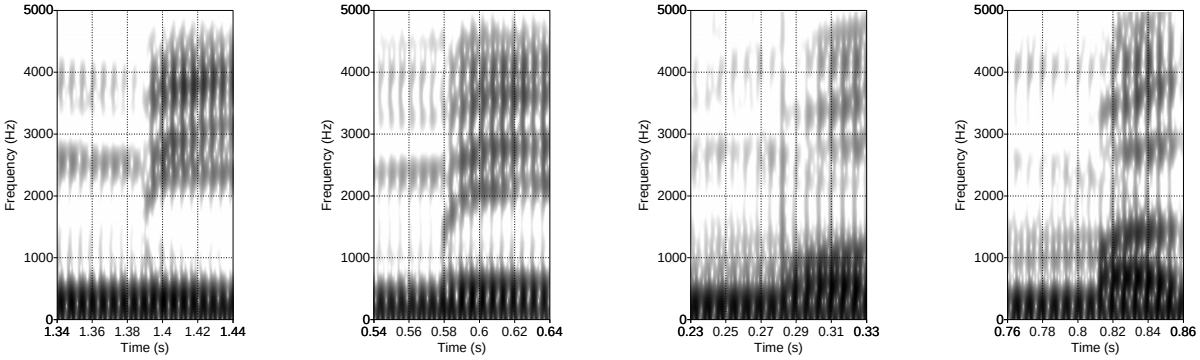
La clase de fonemas sonorantes consiste de vocales, aproximantes, nasales y líquidas. Estos tres últimos sonidos tienen características vocálicas y sonoras, por lo que entran en esta clase. Estos fonemas son diferentes de los sonidos obstruyentes (oclusivos, fricativos y africados) ya que la corriente de aire no está tan obstruída en el tracto vocal. Las consonantes nasales y líquidas forman un grupo separado dentro de la clase de sonorantes ya que, a diferencia de las aproximantes /y/ y /w/, no tienen una contraparte vocálica. Las nasales y líquidas están hechas con más constricción que los sonidos vocálicos y aproximantes. Las nasales se producen cerrando la cavidad oral y abriendo la cavidad nasal, sin haber constricción en ésta. Las líquidas se producen con alguna constricción, pero usualmente no lo suficiente para producir ruido turbulento. Analizaremos las transiciones entre vocales y sonidos nasales y líquidos.

9.1. Nasales y vocales

Los dos fonemas nasales del náhuatl se producen con cierres en los mismos lugares del tracto vocal que los sonidos oclusivos. La posición de los articuladores para la nasal bilabial /m/ es la misma que en /p/; la alveolar /n/ tiene aproximadamente la misma configuración que /t/, pues ésta tiene un punto de articulación apicodental mientras que el de /n/ es apicoalveolar, es decir, se articulan en puntos muy cercanos. Los fonemas nasales son sonoros. El hecho de que el flujo de aire fluya a través de la cavidad nasal, ya que el velo está abierto, provoca que no haya un estallido cuando el cierre oral es liberado. La apertura del velo marca una diferencia básica entre sonidos nasales y oclusivos sonoros, aunque recordemos que éstos no existen en el náhuatl.

9.1.1 Transiciones nasal a vocal

Las figuras 9.1 y 9.2. muestran las transiciones desde /m/ y /n/ hacia las 4 vocales del náhuatl. Las realizaciones de la figura 9.1 corresponden a [komit] (olla), [ome'] (dos), [timota] (despedida) y [tasohkamatik] (gracias). Las realizaciones de la figura 9.2 corresponden a [totonik] (caliente), [ninemi] (vivo), [nopili'] (mi hijo) y [kanači'] (cuánto). Puede observarse que la nasal tiene un F1 muy bajo, esto ya había sido comentado en el capítulo 6. La región nasal se ve muy similar en todas las gráficas, lo cual haría difícil discriminar [m] de [n]. Sin embargo, es posible diferenciar estos sonidos por el efecto de los sonidos vecinos.

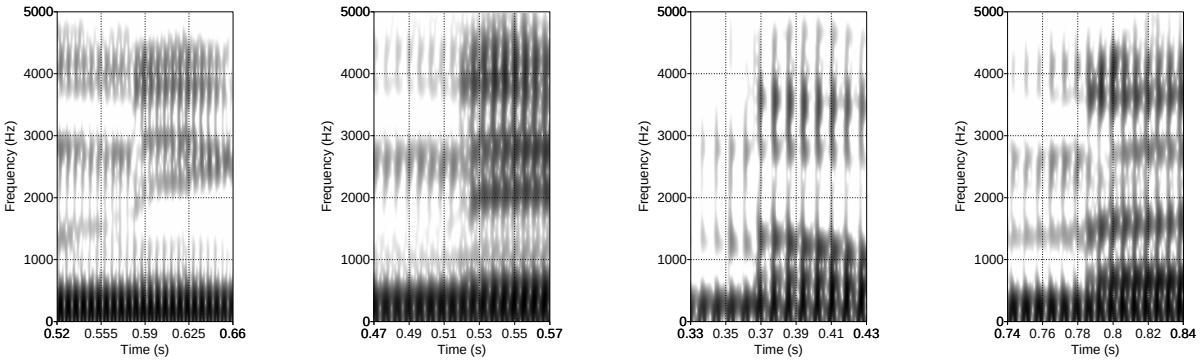


i. [mi] ii. [me] iii. [mo] iv. [ma]

Figura 9.1: Transiciones [mi], [me], [mo], [ma]

Los espectrogramas muestran que la frontera entre los sonidos nasal y vocal es distintiva. En algunos casos la frontera puede ser reconocida por una discontinuidad abrupta en la región de F1 al comienzo de la vocal. Esta discontinuidad es más aparente en aquellas vocales con un alto F1 (vea la transición [m]-[a]). Otra indicación de la de la frontera es el repentino oscurecimiento de los formantes superiores al comienzo de la región de la vocal. Esto lo explica [Oli93], "Cuando el tracto oral se abre, el balance de energía cambia y se nota en los espectros de las vocales una mayor energía en regiones de alta frecuencia que en las nasales".

Nótese además en los espectrogramas una resonancia debajo de F1 durante la región vocálica, observe los casos de vocales con un F1 alto. Esta resonancia es una continuación del formante nasal. El formante nasal también está presente en las vocales con bajo F1, sólo que se combina con éste; observe que F1 luce más ancho que un formante oral sencillo. [Oli93] lo explica, "la presencia del formante nasal indica que a pesar de que la cavidad oral se ha abierto para articular la vocal, el velo permanece abierto, permitiendo que el aire escape a través de los tractos oral y nasal".



i. [ni] ii. [ne] iii. [no] iv. [na]

Figura 9.2: Transiciones [ni], [ne], [no], [na]

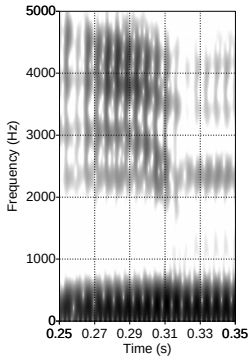
La frecuencia al principio de las vocales así como el movimiento de F2 durante la región vocálica es el factor para identificar las nasales. La tabla 9.1 muestra los valores iniciales de F2 en la región vocálica inmediatamente después de las nasales. Para cada vocal, las entradas para los fonos [m] son siempre las más bajas; para los fonos [n] los valores son consistentemente más altos. Los renglones de la tabla están establecidos para que el valor de F2 de las vocales aparezca en orden descendente. Sin embargo el valor de [o] en ambas columnas resulta más bajo que el de [a]. Cabe mencionar que los datos en la tabla corresponden al mismo hablante, excepto en la transición [no], donde se presentan los datos de un segundo hablante entre paréntesis. La tabla no muestra una consistencia en los valores de F2 que se hubiera esperado al ordenar las vocales de menor a mayor posterioridad.

	[m]	[n]
[i]	2035	2138
[e]	1633	1938
[o]	958 (767)	(1381)
[a]	1127	1479

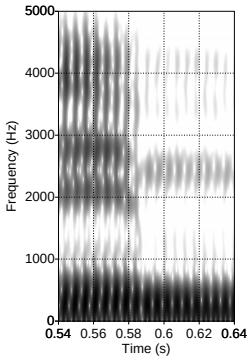
Tabla 9.1: Valores iniciales de F2 de vocales precedidas por nasales (Hz)

9.1.2 Transiciones vocal a nasal

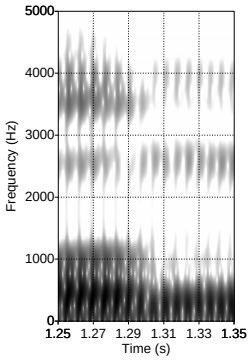
Las figuras 9.3 y 9.4 muestran las transiciones de las vocales a las nasales bilabial /m/ y alveolar /n/, respectivamente. Las realizaciones de la figura 9.3 corresponden a [imeɬ] (su pierna), [ninemi] (vivo), [komit] (olla) y [tasohkamatik] (gracias). Las realizaciones de la figura 9.4 corresponden a [yekinɬi] (ahora), [tonalɬi] (sol), [kanači'] (cuánto). En la región de la consonante se observa una prominente línea oscura en la región de baja frecuencia. Sin embargo también es observable energía en la nasal a mayor frecuencia, la multiplicidad de resonancias nos revela una estructura formántica. En la región F1 de la vocal, los espectrogramas muestran al formante nasal como un doble formante para las vocales con un alto o mediano F1, y un ensanchamiento de F1 para lasa vocales con bajo F1. La presencia del formante nasal en la región vocálica indica que el velo se ha abierto en anticipación de la nasal venidera, de aquí que la vocal esté nasalizada. La nasalización es la característica principal en la transición vocal-nasal.



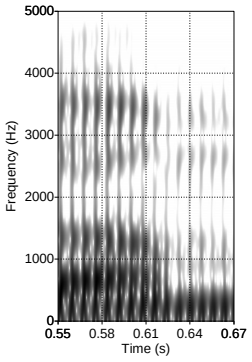
i. [im]



ii. [em]

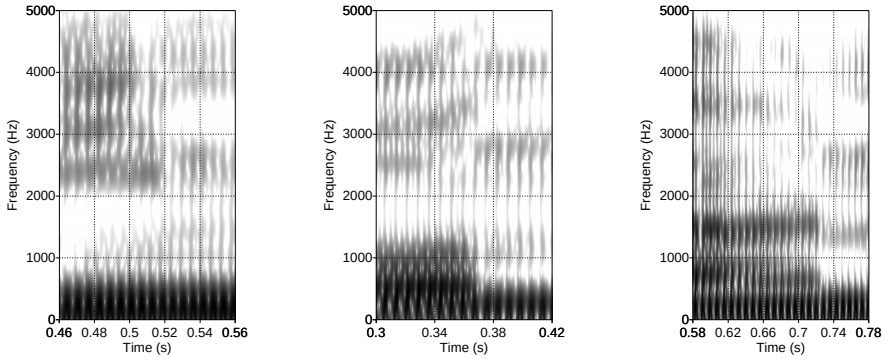


ii. [om]



iv. [am]

Figura 9.3: Transiciones [im], [em], [om], [am]



i. [in] ii. [on] iii. [an]

Figura 9.4: Transiciones [in], [on], [an]

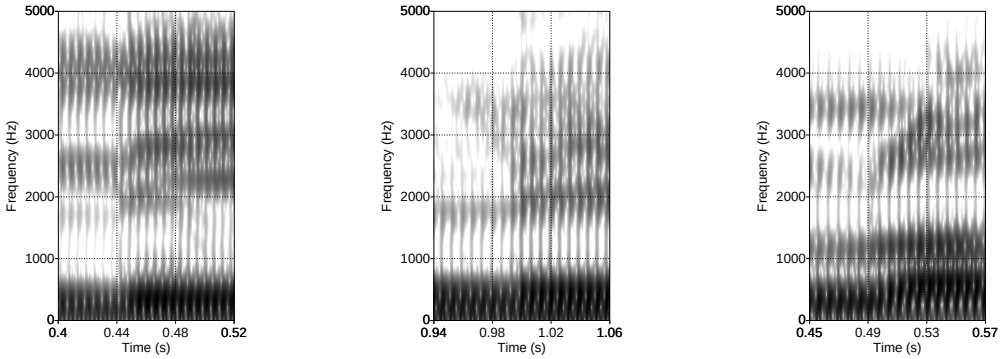
Las resonancias observadas en la región nasal de los espectrogramas son difíciles de interpretar debido a que en estos sonidos intervienen dos cavidades resonantes (o dos tractos paralelos, diferentes y parcialmente separados) en los que viaja el sonido. Esto produce una estructura resonante múltiple y complicada de interpretar.

9.2 Lateral y vocales

La lateral /l/ se produce con una constricción de la punta de la lengua, o ápice, en los alveolos. La constricción para /l/ no para el flujo de aire, como ocurre con las oclusivas y nasales; tampoco la constricción es tan severa como para causar un ruido fricativo ya que la lengua está ubicada en una posición que permite el paso de aire al exterior. El menor grado de constricción es reflejado en las transiciones hacia dentro o hacia afuera del sonido /l/. En lugar de un valor medio de F2, típico de los sonidos alveolares, /l/ tiene un target bajo, similar a la bilabial /w/.

9.2.1 Transiciones lateral-vocal

La figura 9.5 muestra las transiciones de /l/ a tres vocales del náhuatl. Corresponden a las realizaciones [pili'] (niño), [šinečpalewi'] (ayúdeme) y [šolopi] (mentiroso). En esta ocasión se presentan espectrogramas de tres hablantes masculinos distintos. Estos ejemplos corresponden a fonemas dentro de la misma sílaba. Debido al patrón más usual CV del náhuatl, las transiciones /l/-vocal, donde cada fonema pertenezca a diferentes sílabas, se podrían hallar en fronteras de palabras; sin embargo recuérdese que el corpus sólo consta de palabras aisladas. El valor de F1 de /l/ es casi el mismo en cada hablante: 300 Hz, 376 Hz y 370 Hz en [le], [li] y [lo] respectivamente. A diferencia de este mismo tipo de transiciones presentadas por [Oli93] en su estudio fonético del inglés estadounidense donde el F2 de /l/ tiene un valor por debajo de 1000 Hz, en náhuatl los valores de F2 varían por abajo y por arriba de dicho valor. Se puede observar que la lateral muestra estabilidad en los valores de sus formantes F1 y F2. También se nota que en la transición [l]-[i] existe una discontinuidad entre los formantes de ambos sonidos, lo cual no es apreciable en las transiciones hacia [e] y [o]. Las regiones de [l] para cada vocal son diferentes en los valores de F2 y F3, podemos concluir que las vocales vecinas ejercen influencia en la producción de /l/. Una discontinuidad o un cambio en el nivel de energía en la región arriba de F2 (hay mayor energía al comienzo de la vocal) nos permite distinguir la unión entre /l/ y vocal.

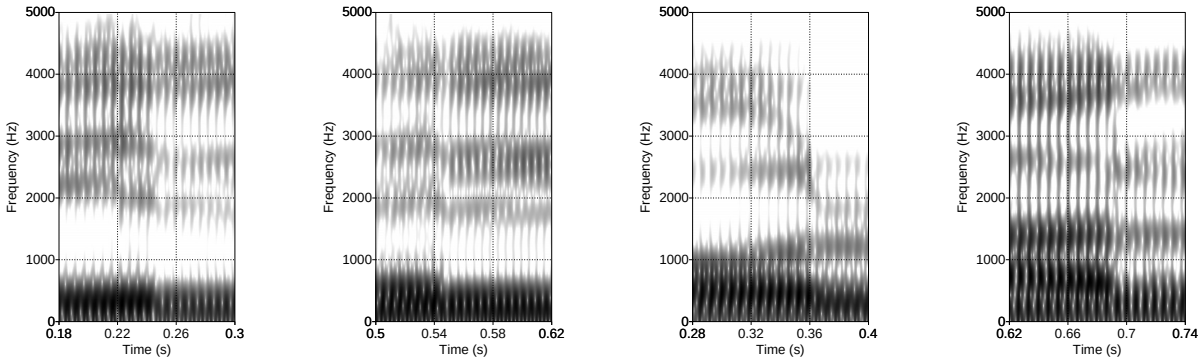


i. [li] ii. [le] iii. [lo]

Figura 9.5: Transiciones [li], [le], [lo]

9.2.2 Transiciones vocal-lateral

Los espectrogramas de la figura 9.6 corresponden a las cuatro vocales en transición a la consonante lateral. La transición [a]-[l] se desprende de la realización [ta:l] (tierra), por lo que ambos fonemas pertenecen a la misma sílaba. El resto de las transiciones mostradas corresponden a fonemas que pertenecen a las fronteras de distintas sílabas, son extractos de las realizaciones [pili'] (niño), [kaneli'] (falso) y [kolot] (alacrán). Observe que los espectros son casi imágenes espejo de las transiciones mostradas en la sección anterior. Nuevamente se observan dos importantes características en [l]: un bajo valor de su F1 y los diferentes valores que toma su F2, dependiendo de la vocal que le sigue, lo cual es evidencia de coarticulación. También se observa estabilidad de dichos formantes durante la producción de la lateral. Es de notar que [l] en ocasiones presenta una mayor energía en su F3 que en su F2, por ejemplo vea la transición [e]-[l].



i. [il] ii. [el] iii. [ol] iv. [al]

Figura 9.6: Transiciones [il], [el], [ol], [al]

No se aprecia claramente si hay existe discontinuidad en las transiciones pero si es notorio que en ellas hay un movimiento de los formantes. El movimiento más drástico se presenta en los formantes superiores a F2 en la transición [o]-[l]. De no ser por los cambios de energía entre vocal y lateral, no sería sencillo distinguir un sonido de otro al observar sus espectrogramas.

10. INTERACCIONES CONSONÁNTICAS

Como se comentó anteriormente, el patrón silábico del náhuatl es (C)V(C), siendo CV la combinación más común, seguida de CVC, VC y V. No existen combinaciones de consonantes en una sílaba con el patrón CCV o VCC. Por lo tanto la combinación de consonantes se presenta entre sílabas o palabras; bajo esta consideración cabe mencionar que nunca se combinan tres consonantes (CCC).

Los sonidos consonánticos son divididos en dos grandes grupos: obstruyentes (oclusivas, fricativas y africadas) y sonorantes (líquidas, nasales y aproximantes).

Se estudiarán las interacciones consonánticas observando las transiciones obstruyente-obstruyente, obstruyente-sonorante y sonorante-obstruyente. Al final se analizarán las interacciones sonorante-sonorante.

10.1 Interacciones obstruyente-obstruyente

Se observarán las transiciones e interacciones entre:
Oclusiva-oclusiva.
Oclusiva-fricativa.
Fricativa-oclusiva.

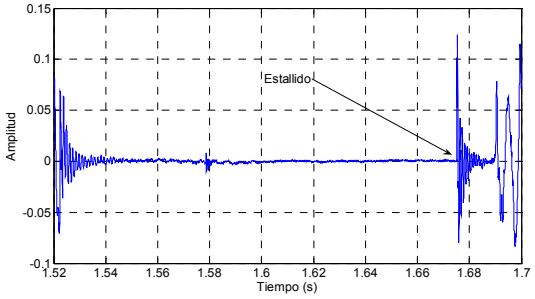
Las africadas están formadas por grupos compuestos de una oclusiva y una fricativa. Ya se hizo una presentación de las africadas en la sección de fonética estática, sin embargo estos sonidos serán estudiados a mayor detalle en la sección correspondiente a la interacción oclusiva-fricativa. El corpus no cuenta con interacciones fricativa-fricativa, por lo que no fueron abordadas.

10.1.1 Interacciones oclusiva-oclusiva

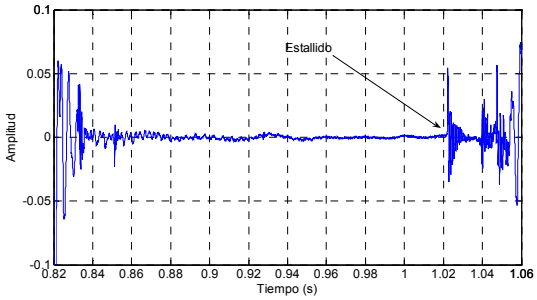
En el habla natural no siempre es posible esperar una secuencia continua de oclusivas que sea cada una de ellas plenamente distinguibles.

Se observará ahora la interacción entre dos oclusivas velares /k/. En la figura superior 10.1 se muestra la forma de onda de una sola [k] en posición intervocálica, está extraída de la realización [amo nitahtowa koyotahtol] (no hablo español). En la figura inferior 10.1 se presenta la forma de onda de una doble [k] intervocálica, tomada de la realización [kani yetok komit] (dónde está la olla). Se observa claramente que el intervalo de cierre de la doble [k] (aproximadamente de 180 ms) es 40% más largo que el cierre de la [k] sencilla

(aproximadamente 130 ms). El intervalo de cierre prolongado nos da evidencia de la interacción oclusiva-oclusiva.



i. [k] sencilla



ii. [k] doble

Figura 10.1: [k] contra [k-k]

Evidentemente en el espectrograma de la interacción [k]-[k] sólo apreciaremos una sola región de cierre y una sola explosión, vea la figura 10.2, donde para una mejor apreciación se presenta el espectrograma con preénfasis. Por lo anterior, resulta útil examinar la forma de onda. Es posible concluir que los hablantes diferencian entre la [k] sola y la doble [k] alargando el intervalo de cierre para esta última.

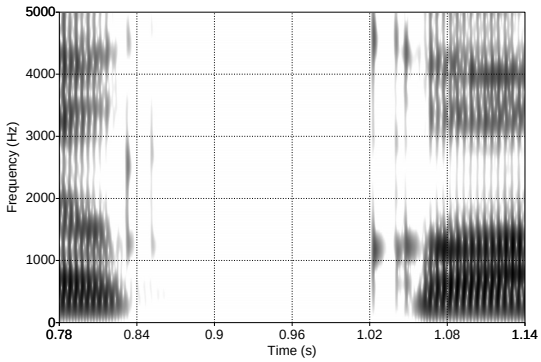


Figura 10.2: Espectrograma de la unión [k]-[k] de la forma de onda de la figura 9.1

Ahora examinaremos la interacción entre dos oclusivas diferentes. La figura 10.3 muestra la forma de onda y espectrograma del segmento [kt] de la realización [pokti'] (humo). El inicio del estallido de [k] está señalada por una flecha en la forma de onda, después se presenta la región de cierre de [t] y al final ocurre su estallido. El espectrograma muestra los estallidos de

ambas oclusivas. Se observa una posible coarticulación en el segmento [okti], el segundo formante de [o] desciende ligeramente al final de la vocal, pero en el estallido de [k] se observa una estructura formántica con movimiento hacia arriba, esto también se aprecia en el estallido de [t] para así alcanzar al F2 de [i]. Para una mejor apreciación se presenta el espectrograma con preénfasis.

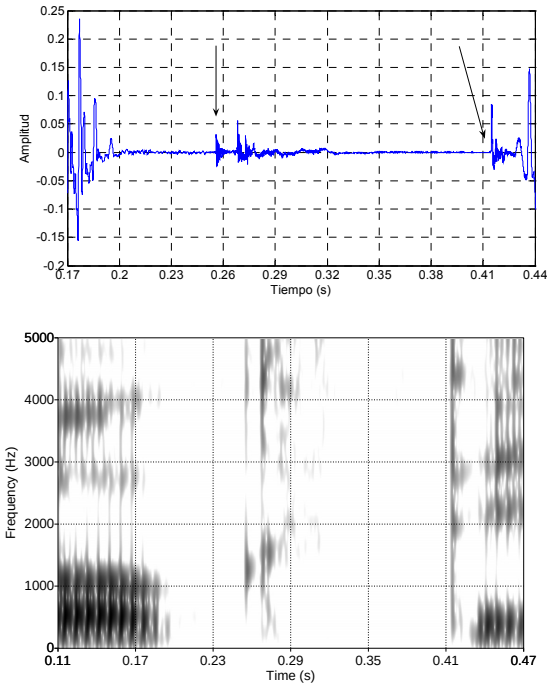


Figura 10.3: Unión [kt]

En la figura 10.4 se presenta la forma de onda y espectrograma del segmento [pt] de la realización [wipta'] (pasado mañana). La forma de onda muestra el estallido de [p], comenzando en la flecha, seguida por una región de cierre para [t]. El cierre está seguido de un estallido. El espectrograma muestra los estallidos de [p] y [t]. Como en las figuras anteriores, se presenta el espectrograma con preénfasis.

Estos ejemplos muestran que la articulación de oclusivas sordas puede variar, siendo más factible que no se libere la primera oclusiva entre fronteras de palabras que dentro de una palabra.

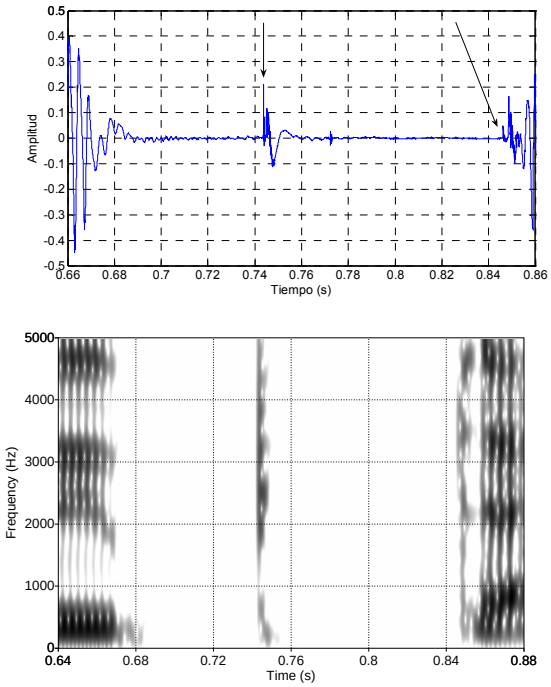


Figura 10.4. Unión [pt]

10.1.2 Interacciones oclusiva-fricativa

Además de los sonidos africados, contamos en nuestro corpus con la transición de [k] a [s] tomada de la realización [aksa'] (alguien). Note que se trata de una transición a través de sílabas VC.CV. En la figura 10.5 se muestran la forma de onda y espectrograma de dicha transición. La región de cierre y el estallido es visible en la forma de onda, éste último está señalado por una flecha. La duración del cierre no es breve (aproximadamente 87 ms). También es claramente apreciable el segmento de la fricativa ya que la señal empieza a adquirir gradualmente una mayor amplitud. En el espectrograma también se observa la región de cierre y el estallido, cuyo inicio se aprecia mediante una línea vertical de energía que abarca un gran número de frecuencias poco después de los 56 ms, aunque el estallido posee mayor energía por debajo de los 2.5 kHz.

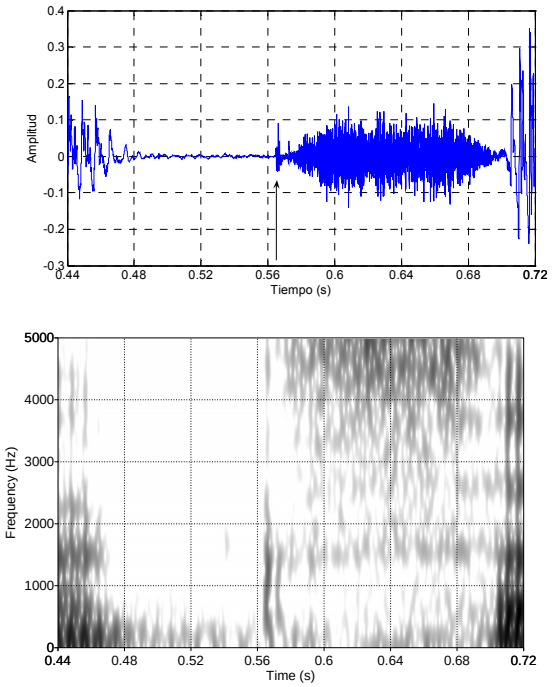


Figura 10.5: Unión [ks] de la palabra [aksa']

10.1.3 Interacciones fricativa-oclusiva

La figura 10.6 muestra la forma de onda y espectrograma de la combinación fricativa-oclusiva [st], correspondiente a la realización [esti'] (sangre). Note que cada obstruyente pertenece a diferentes sílabas de acuerdo al patrón silábico CV.

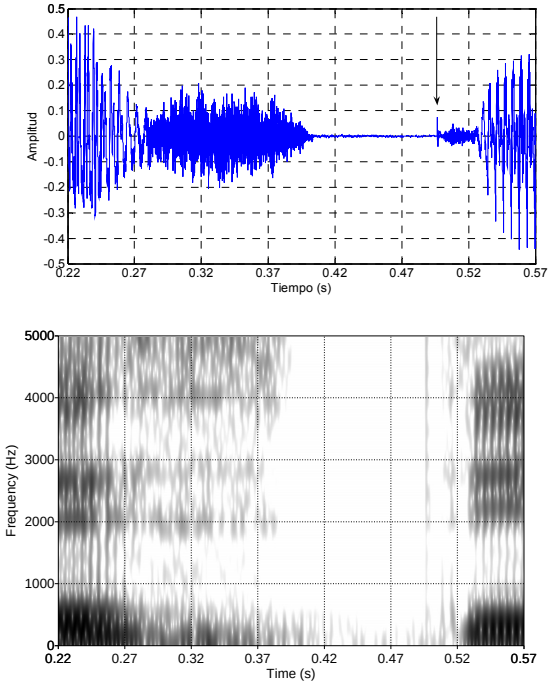


Figura 10.6: Unión [st] de [esti]

La forma de onda y el espectrograma muestran parte de la vocal [e], la fricativa, la oclusiva y parte de la vocal [i]. Observamos que la región de cierre de la oclusiva está bien definida y contiene muy poca energía. El estallido es señalado en la forma de onda con una flecha. No se hallaron casos, como en [Oli93] en su estudio de la fonética acústica del idioma inglés, donde el silencio sea muy breve o haya una alguna vibración en el mismo.

La figura 10.7 muestra la forma de onda y espectrograma del mismo tipo de interacción fricativa-oclusiva [s]-[t]. En esta ocasión la interacción es a través de fronteras de palabras [tayis tepi¢in] (de la oración [ti tayis tepitsin a:t], ¿quieres tomar poquita agua?). Observamos que el efecto es el mismo: la zona de cierre contiene muy poca energía y el estallido está bien definido. En consecuencia se percibe auditivamente con claridad la articulación de los dos sonidos.

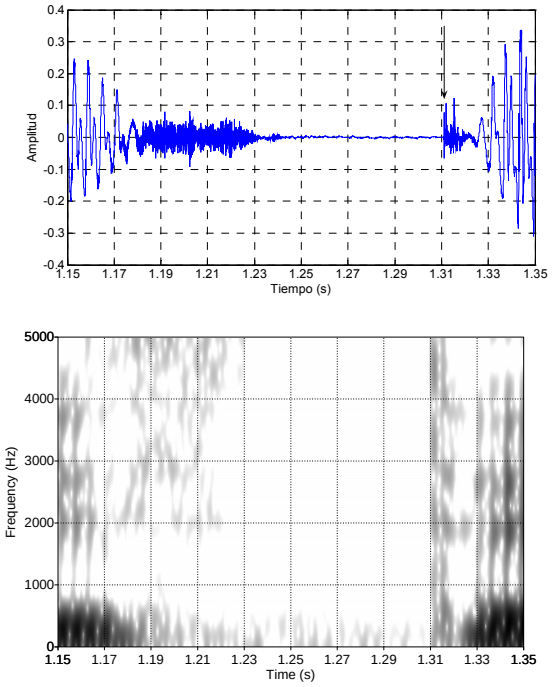


Figura 10.7: Secuencia [s]-[t] a través de fronteras de palabras

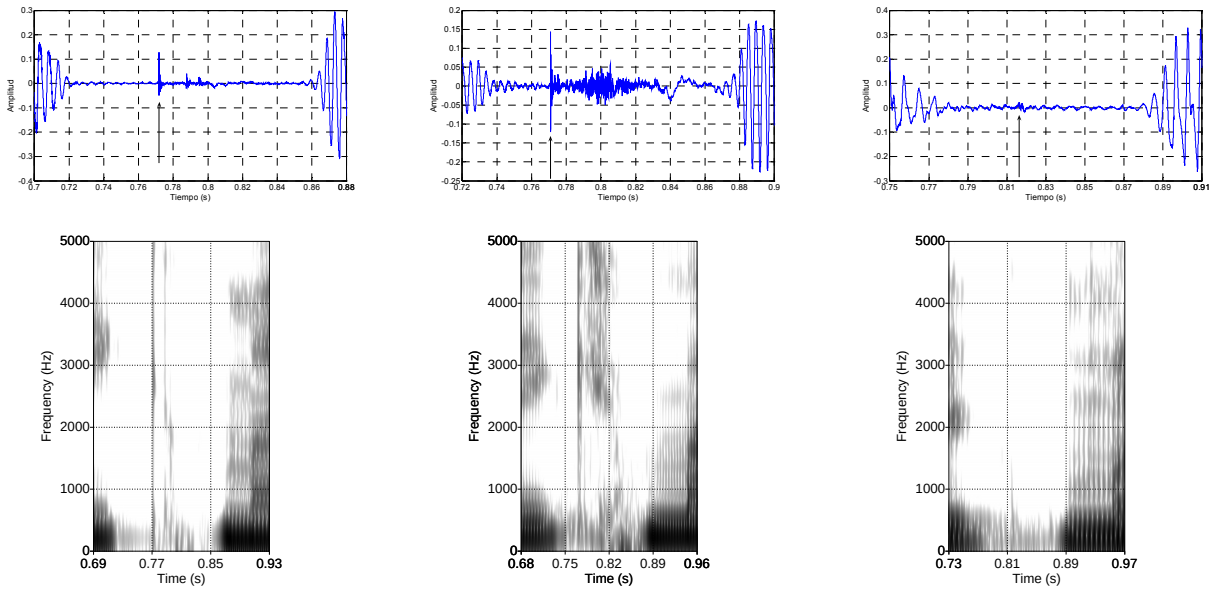
También se inspeccionaron a través de formas de onda y espectrogramas las transiciones /¢t/ de la realización [me¢ti'] (luna), las características han sido las mismas que en los casos anteriores.

10.2 Interacciones obstruyente-sonorante

Esta sección examinará las interacciones de oclusivas y fricativas con las consonantes sonorantes (nasales, líquidas y aproximantes) cuando las obstruyentes preceden a las sonorantes.

10.2.1 Interacción oclusiva-nasal

La interacción que presentamos en la figura 10.8 se extrajo de la oración [amo nikmati'] (no lo sé) por tres distintos hablantes. Se observa que la interacción [km] es a través de fronteras de sílabas en [nikmati]. Las distintas realizaciones nos muestran resultados distintos. La pronunciación canónica se muestra en las figuras de la izquierda, antes del estallido (señalado en la forma de onda con una flecha) hallamos la región de cierre, el estallido dura unos 35 ms para volver a un silencio que precede a la sonorante [m]; podemos observar también en el espectrograma la explosión de la oclusiva a los 77 ms. En las figuras del centro se vuelve a señalar con una flecha en la forma de onda el estallido de la oclusiva, se observa una mayor variación de la señal en la región de cierre, de hecho hay cierta periodicidad, lo que implica que la oclusión no es completa y las cuerdas aún vibran un poco; también llama la atención que la oclusiva está aspirada debido a la gran variación de la señal después del estallido; en conclusión esta realización se parece más a una oclusiva aspirada sonora. Por último se observó también que los hablantes tendían a debilitar el estallido, observe la columna de la derecha; en esta realización el estallido es muy débil y ha sido señalado por una flecha en la forma de onda, note que la duración total de la señal desde el inicio de la región de cierre hasta el inicio de /m/ es muy semejante al de los otros casos.



i. Mujer ii. Mujer iii. Hombre

Figura 10.8: Secuencias [km] intersilábicas pronunciadas por tres hablantes distintos

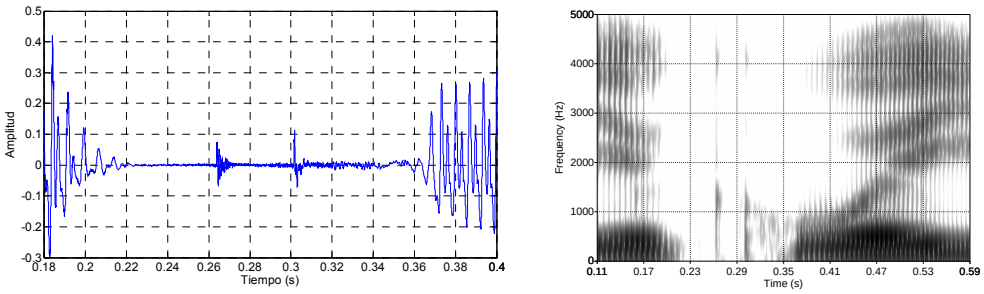
Hemos visto en la figura anterior diferentes realizaciones donde las secuencias son completamente articuladas o bien la explosión de la oclusiva es muy débil, siendo el caso más

común. Sin embargo la duración total de la oclusiva es muy semejante. Hubo variaciones donde se aspira el sonido oclusivo.

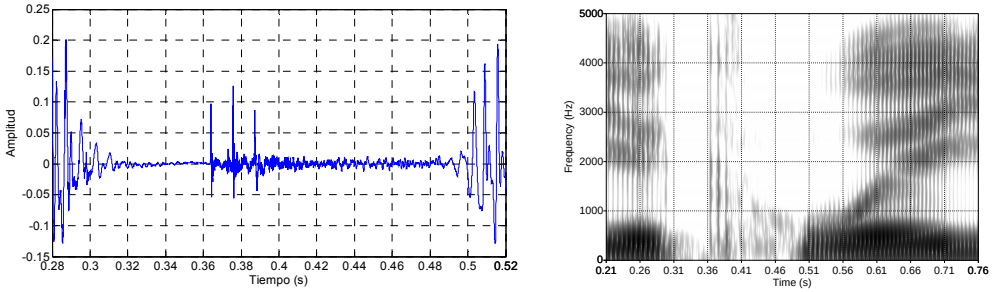
10.2.2. Interacción oclusiva-aproximante

En la figura 10.9 se muestran dos realizaciones de la oración [yek weyi ne kowit] (está muy grande ese árbol), pronunciadas por el mismo hablante. Se presentan las formas de onda y espectrogramas de cada realización. Se muestra a la aproximante [w] en combinación con la oclusiva velar [k]. Por lo tanto se trata de una transición entre fronteras de palabras.

Las formas de onda muestran el fin de la vocal [e], la oclusiva completa [k] y el inicio de la aproximante. Los espectrogramas toman la mayor parte de [e], la oclusiva completa, la vocal siguiente completa hasta llegar a la aproximante [y].



i. Primera realización



ii. Segunda realización

Figura 10.9: Dos realizaciones de un mismo hablante de la secuencia [k]-[w] en fronteras de palabras

Se aprecia que la oclusiva velar [k] está aspirada. También observe que la transición desde [k] a [w] ocurre por completo durante la aspiración. Cuando comienza la aproximante su F2 empieza su ascenso hacia la vocal adyacente.

En la figura 10.10 se muestra la interacción de la aproximante palatal [y] con la oclusiva alveolar [t] a través de la oración [ne takat yek nemi] (ese hombre es un vago). La interacción es a través de fronteras de palabras. La forma de onda muestra el estallido de la oclusiva, después de esto observe que hay un cierto comportamiento fricativo; el F2 de la aproximante [y] está descendiendo al comienzo de la sonoridad de dicho fonema. Se presenta también el

espectrograma aplicando preénfasis y con rango frecuencial de 0 Hz a 7 kHz para observar con mayor claridad la existencia de la sección fricativa en esta interacción.

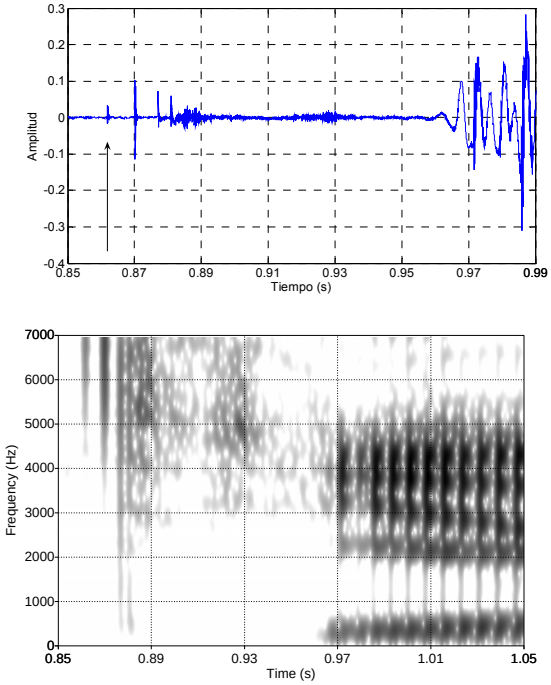


Figura 10.10: Secuencia [t]-[y] en fronteras de palabras

10.3 Interacciones sonorante-obstruyente

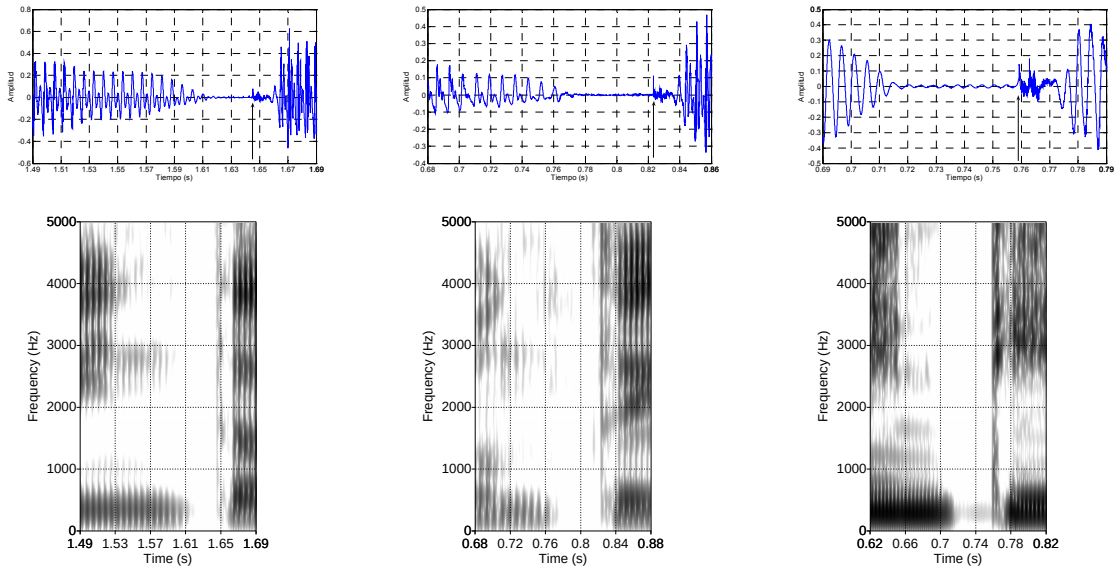
10.3.1 Interacciones nasal-obstruyente

La figura 10.11 muestra la secuencias [n]-[t] en dos modalidades: a través de fronteras de palabras por medio de la oración [ka:nači nipatiw in tasal] (¿Cuánto cuesta esta prenda?) y a través de sílabas por medio de las realizaciones [no¢onteko] (mi cabeza) y [kanintinemi'] (¿dónde vives?).

Las dos primeras oraciones son pronunciadas por el mismo hablante masculino, la última oración es pronunciada por una mujer. En las formas de onda se señala con una flecha el estallido de la oclusiva. Tanto las formas de onda como los espectrogramas, con preénfasis, muestran desde una parte de la vocal anterior a [n] hasta una parte de la vocal que sigue de [t]. Sin embargo hay una excepción, la forma de onda correspondiente a [kanintinemi'] se presenta con un mayor acercamiento de la región de cierre.

En las formas de onda se observa cómo [n] decae hasta combinarse con la región de cierre de la oclusiva. Ante esto, [Oli93] menciona que "ni las formas de onda, ni los espectrogramas muestran una clara frontera entre la nasal y el cierre de la oclusiva. La falta de una distinción clara entre la nasal y el cierre es el resultado de la manera como estos sonidos son producidos". Agrega también que "cuando una oclusiva sigue de una nasal y ambas tienen en mismo punto de articulación, el cierre ocurre al principio de la nasal. La única diferencia entre una secuencia nasal-oclusiva y una oclusiva sola está en la posición del puerto del velo. En el segundo caso el velo está cerrado para la duración entera del cierre de una oclusiva, mientras que está brevemente abierto al principio del cierre en una secuencia nasal-oclusiva. Así, mientras se articula la nasal, el velo se abre y el aire escapa a través del tracto nasal. Durante la producción de una oclusiva oral, el flujo de aire se termina. Sin embargo, es difícil interpretar las formas de onda o los espectrogramas para establecer el punto de cierre velar."

En la transición [nt] de la palabra [kanintinemi'] observamos que en la región de cierre la señal muestra, a diferencia de los otros dos ejemplos, un comportamiento periódico.



Extracto de [ka:nachi nipatiw in tasal] | Extracto de [no¢onteko] | Extracto de [kanintinemi']

Figura 10.11: Secuencias [n]-[t] en fronteras de palabras y sílabas

La figura 10.12 muestra un ejemplo de una nasal seguida de una oclusiva sorda, se trata de la transición /nk/ proveniente de [inkali'] (esta casa). El espectrograma tiene preénfasis. Sabemos que esta nasal y oclusiva se articulan en diferentes lugares, esto colabora para poder diferenciar más claramente la frontera entre ambos sonidos; observe que la frontera entre [n] y la región de cierre de [k] es claramente visible en el espectrograma. Esta frontera está indicada por los formantes nasales F1 y F2, donde ambos se extinguen al mismo tiempo.

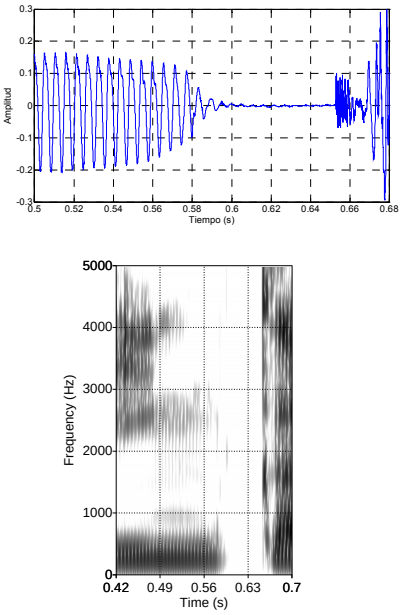


Figura 10.12: Interacción [nk]

El intervalo de cierre de una oclusiva después de nasal varía mucho, puede ser muy corto o inclusive casos donde no se diferencie de la nasal. Se presenta un ejemplo de esto en la interacción nasal-africada [n¢] de la figura 10.13, ésta presenta un extracto de la palabra [yekin¢in] (ahora). El segmento incluye una porción de la vocal que precede al sonido nasal para que se aprecie en el espectrograma (el cual tiene preénfasis) la distinción entre vocal y nasal. La forma de onda muestra la ausencia del intervalo de cierre, pero es posible apreciar la liberación de la oclusión.

De acuerdo a Olive, resulta un fenómeno común que las interacciones nasal-africada a final de sílaba carezcan de un intervalo de cierre e incluso del estallido. Lo anterior resulta comprensible debido a que en el idioma inglés esos sonidos están en posición coda, en esa situación el escucha hace uso de su conocimiento contextual o de léxico para identificar la secuencia nasal-africada. Olive agrega que este fenómeno también se presenta en la secuencia /nč/, la cual podría confundirse con /nš/.

Como sabemos, la interacción /n¢/ en el náhuatl sólo puede darse a través de sílabas, por lo que es de esperar que la africada /¢/ cuente con la mayor parte de sus rasgos al estar ubicada en posición onset. Esto tiene sentido ya que las realización que revisé por lo general contaban con el intervalo de cierre, y en los casos donde no se diferenciaba de la nasal sí se podía observar el estallido. Es decir, las africadas contaban con la mayoría de sus rasgos característicos.

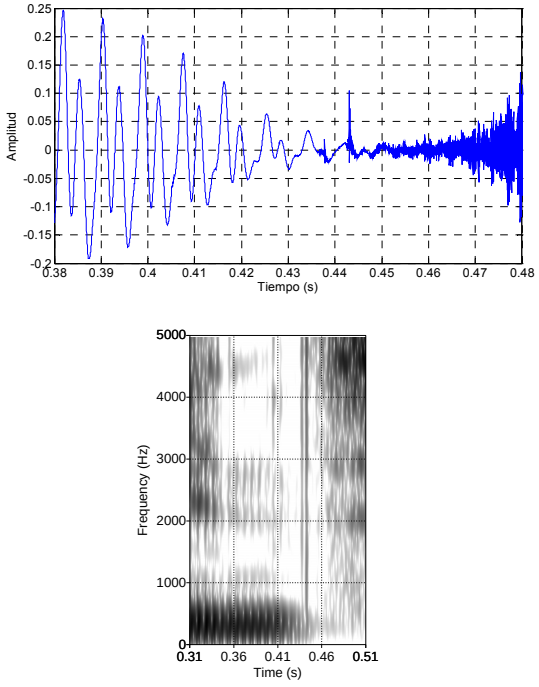


Figura 10.13: Interacción [n¢]

10.3.2 Interacciones líquida-obstruyente

Tanto en la interacción /lt/ como en /l¢/ se aprecia, en diferentes grados, una sección ruidosa al final de la lateral, ejemplos de este fenómeno con dicha africada se presentaron en la sección de fonética estática en fin de sílaba.

Por otro lado, en la figura 10.14 se revisa lo que ocurre en cuanto al movimiento de los formantes de la lateral en su interacción con un sonido obstruyente, el espectrograma tiene preénfasis. El segmento proviene de la realización [taltik] (café). El espectrograma muestra parte de los sonidos vecinos que interactúan con [lt], los fonos [a], [i]. Observe que los formantes del sonido lateral tratan de alcanzar los de la vocal [i] a pesar de que entre ambos sonidos se encuentre la oclusión. Debido a que el fono [k] sigue de la [i], los formantes F2 y F3 de esta vocal tienden a combinarse.

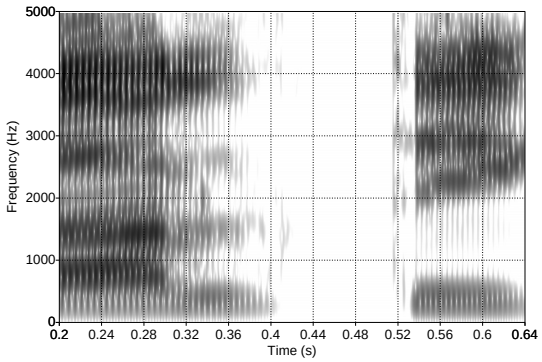
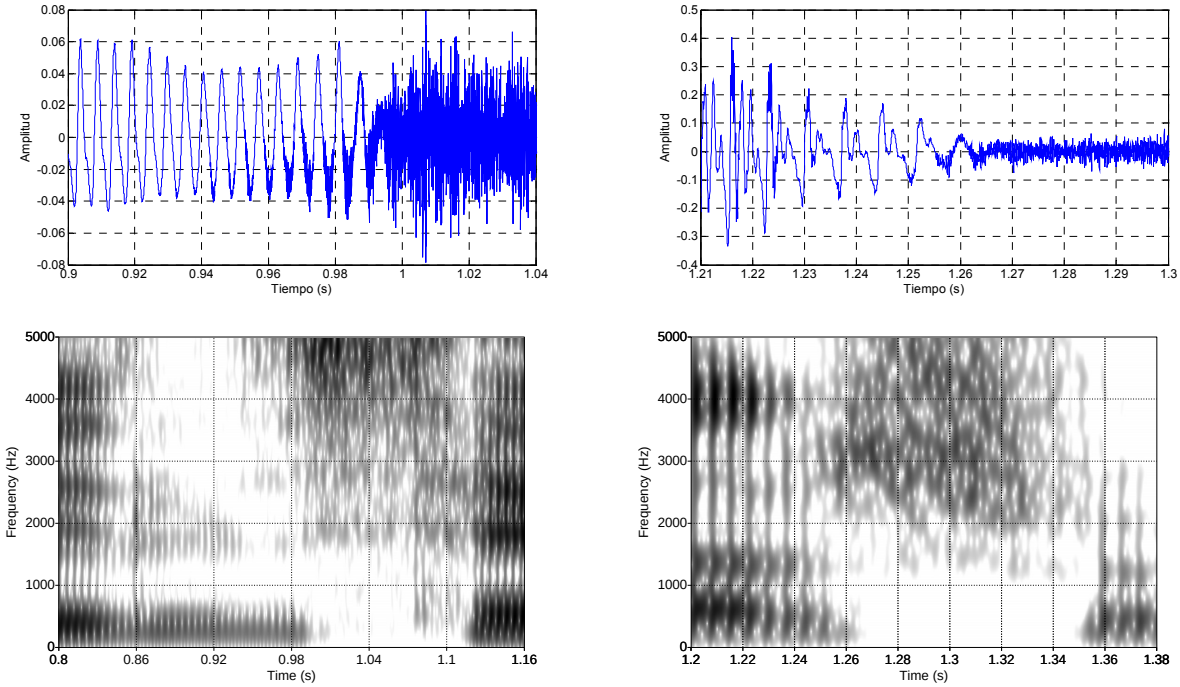


Figura 10.14: Interacción [lt]

La figura 10.15 muestra las interacciones [l]-[s] y [l]-[š] de las realizaciones [nokičpil se ta¢iw] (mi hijo es un flojo) y [ne siwat itasal šošoktik] (esa mujer tiene la ropa verde). Se ve que las interacciones son a través de palabras. Estas realizaciones fueron pronunciadas por distintos hablantes masculinos. Los espectrogramas cuentan con preénfasis e incluyen tanto a las consonantes en inspección como las vocales que las rodean; esto ha sido así ya que se observa una mayor interacción entre la lateral con la vocal después de fricativa que con la misma fricativa. Razón de esto es el movimiento de los formantes F1 y F2 de cada [l] para alcanzar a los de la vocales [e], [o] después de [s] y [š] respectivamente. Observe que la duración de cada fono lateral es muy diferente y que gradualmente la forma de onda, que muestra un segmento de lateral y fricativa, se va "contaminando" con los rasgos de la fricativa, sobre todo en la interacción [l]-[s].



i. [l]-[s] ii. [l]-[š]

Figura 10.15: Interacciones lateral - fricativa

10.4 Interacciones sonorante-sonorante

10.4.1 Interacción lateral-aproximante

En la figura 10.16 se muestra el espectrograma, con preénfasis, de la interacción de la lateral [l] seguida de la aproximante [w]. La realización es [yalwa'] y se muestra un segmento de la vocal [a], la lateral completa y un segmento de [w]. Como ocurre en los casos lateral-obstruyente ya vistos, la transición es suave. En su transcurso, F2 muestra un descenso suave. F1 presenta un movimiento muy ligero hacia arriba, pero es casi imperceptible.

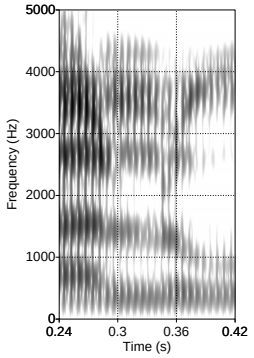


Figura 10.16: Interacción [lw]

11. VARIACIONES

En secciones previas se han presentado los correlatos acústicos del náhuatl de San Miguel Tzinacapan; también se ha ilustrado como estos correlatos acústicos cambian en el tiempo en la combinación de fonemas. Por lo general se presentaron ejemplos con segmentos de sonido bien formados cuyas propiedades acústicas cuentan con la mayoría de los rasgos distintivos del fonema en estudio; para estos ejemplos también se tomó en cuenta que dicho segmento de sonido fuera buen representante de las realizaciones de todos los hablantes. De esta manera se procuró presentar la tendencia de los fenómenos relacionados con los fonemas en su interacción con otros. Cabe mencionar que hubo algunas excepciones y se presentaron algunas variaciones o casos excepcionales que resultaban convenientes para incluir en este estudio y así clarificar algunos fenómenos o advertir de la presencia de otros.

Si las señales de voz fueran siempre bien comportadas, estables, con atributos acústicos fácilmente identificables, las confusiones en la comunicación rara vez ocurrirían; más aún, los problemas en las tecnologías del habla como reconocimiento y síntesis de voz serían fácilmente identificables y resueltos. Sin embargo la señal de voz natural es extremadamente variable como ha podido apreciar a lo largo de este estudio. Aún si las variaciones a causa de la edad, características anatómicas y otros factores pudieran ser eliminados, habría aún variación en la señal de voz debido a que incluso el mismo hablante nunca producirá la misma combinación de fonemas de manera idéntica en cada ocasión.

Uno de los sonidos extraordinariamente variable en el náhuatl es la fricativa glotal /h/ debido a que depende enormemente del contexto en el que se produce. Es decir, su realización es muy flexible y sus propiedades cambian cuando ocurre en diferentes ambientes.

En esta sección se presentarán variaciones que fueron encontradas durante la realización del estudio fonético. Como se ha dicho algunas variaciones importantes ya fueron reportadas, por lo que se mencionarán algunas otras.

11.1 Variaciones vocálicas

Algunos hablantes solían aspirar la vocal en inicio de palabra. La figura 11.1 presenta la forma de onda y espectrograma del segmento [es] de la realización [esti'] (sangre). La aspiración es observable tanto en la forma de onda como en el espectrograma (al que se le ha aplicado preénfasis), note que su intensidad es muy baja y tiene muy poca energía.

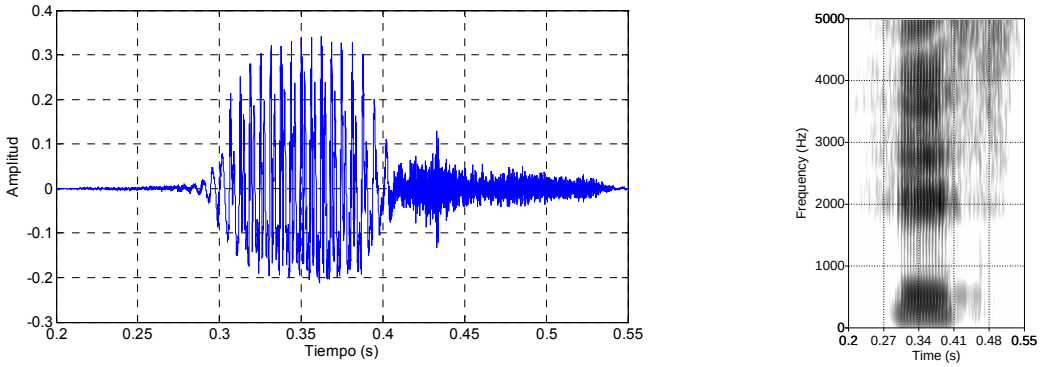


Figura 11.1: Segmento [es] de la palabra [esti'] (sangre)

Al respecto de las vocales aspiradas es conveniente hacer un comentario más, comúnmente se presenta este fenómeno en las vocales ubicadas a final de palabra. No resultaba sencillo determinar la presencia del fonema /h/ a final de palabra o de una vocal aspirada. La recomendación obtenida a través de expertos en lingüística es, en caso de duda, al momento de la grabación transcribir la información como si se tratara de una vocal aspirada y posteriormente de manera más fina, con la ayuda de software de análisis acústico por ejemplo, analizar la señal de voz para determinar la clasificación del sonido lingüístico. Como el náhuatl es una lengua muy estudiada es posible presentar la transcripción fonética del cuestionario con bastante exactitud (por ejemplo, todos los pronombres personales usan la oclusiva glotal /h/, la cual mediante un estudio fonético acústico es sencilla de identificar).

Por último vale pena hacer otro comentario, el espacio formántico de las vocales de una lengua permite el traslape de las fronteras de cada una de ellas (en consecuencia, por ejemplo, una /i/ puede casi escucharse como /e/ debido a que los formantes están ubicados en alturas cercanas). Mientras más vocales tenga una lengua, el espacio estará más ocupado por ellas y por lo tanto habrá más restricciones para que una vocal pueda variar. Debido a que el náhuatl tiene 4 vocales, éstas pueden cubrir mayor espacio y por ende las vocales están más libres para variar. Una manifestación de lo anterior es que ante la ausencia de la vocal /u/ en la lengua náhuatl, la vocal /o/ cuenta con mayor espacio en el espacio formántico para variar; en consecuencia los sonidos /o/ y /u/ son la misma vocal. El efecto en un hablante náhuatl al escuchar el sonido /u/ en lugar de /o/ es que se trató de una pronunciación no muy correcta de este último sonido; sin embargo la palabra que incluya el sonido /u/ es fonológicamente correcta [Sem1].

11.2 Variaciones nasales

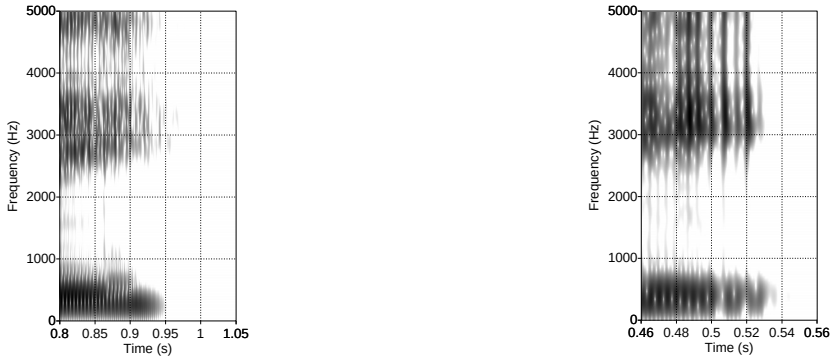
Durante las sesiones de grabación algunos hablantes pronunciaban determinadas palabras del vocabulario con /n/ al final de las mismas, mientras que otros hablantes no lo hacían así. La presencia o ausencia de la nasal /n/ a final de palabra fue registrada en [tonal¢in] / [tonal¢i] (sol), [nikan] / [nika] (aquí), [yekin¢in] / [yekin¢i] (ahora), [ih kon] / [ih ko] (así). Como se ha dicho, algunos hablantes daban una variación, otros la otra, pero ninguno las alternaba. Ante este hecho se decidió explorar si existía alguna característica de nasalidad en la última vocal en aquellas realizaciones, evidentemente, con aparente ausencia de nasal.

La figura 11.2 muestra los espectrogramas de los segmentos [kan] y [ka] extraídos de las realizaciones [nikan] y [nika] pronunciadas por dos hablantes distintos. Se aplicó preénfasis a los espectrogramas. Es completamente notorio que la realización que carece de /n/ no cuenta con rasgos nasales en la vocal, es decir, la presencia de un formante nasal por debajo de F1. El espectrograma de [kan] muestra todas las características de la nasal: múltiples resonancias manifestadas por una estructura formántica en [n] y presencia del formante nasal en la vocal adyacente por efectos de coarticulación.

i. [kan] ii. [ka]

Figura 11.2: Extractos de las realizaciones [nikan] y [nika]

Para hacer una última inspección se presentan en la figura 11.3 los espectrogramas de la vocal final de las realizaciones [tonal¢i] (sol) y [pokti'] (humo) pronunciadas por la misma hablante. La realización [pokti'] indudablemente no tiene nasal al final y resulta útil para hacer una comparación con la otra realización, observe que es posible sospechar la presencia de un comportamiento nasal en [tonal¢i] debido a que aparentemente hay un formante nasal en la vocal, es más notorio al final de ella; mientras que en [pokti'] no se observa este fenómeno. Esto nos da una pista para determinar si hay nasalización a final de palabra (vocal nasal) a pesar de la ausencia del cierre nasal para producir /m/ o /n/.

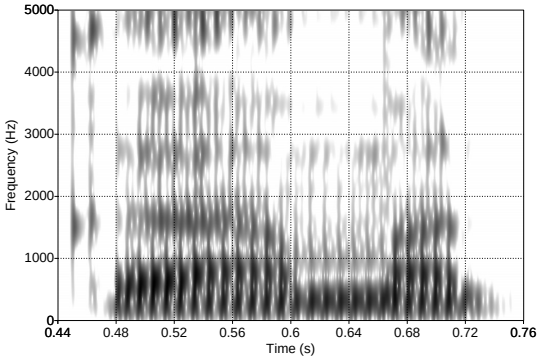


i. [i] ii. [i]

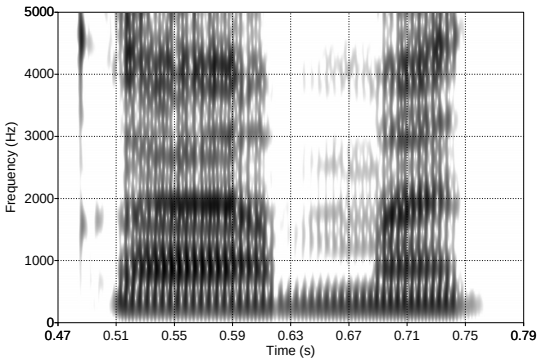
Figura 11.3: Extractos de las realizaciones [tonal¢i] y [pokti']

11.3 Factores adicionales de variación acústica

Vale la pena mostrar una diferencia predecible en el habla de un hombre y una mujer y que se refleja en un espectrograma, se trata de la frecuencia fundamental. Vea la figura 11.4, se muestran los espectrogramas con preénfasis de los segmentos [kama] de [tasohkamatik] (gracias) dichos por un hombre y una mujer. Observe que las líneas verticales de estos espectrogramas son diferentes entre los hablantes, estando estas líneas más cercanas entre sí en la mujer que en el hombre, incluso se distinguen menos en ella. Esta diferencia en el espectro es debida a que la frecuencia fundamental es más alta en la mujer. Note además que estos espectrogramas también difieren en la altura de los formantes de las vocales.



hablante masculino



hablante femenino

Figura 11.4: Comparación de hablantes diferentes

12. ENTRENAMIENTO Y RECONOCIMIENTO DE VOZ

A continuación se describe el sistema de reconocimiento de voz, de palabras aisladas y dependiente del locutor utilizando MSBC (Multisection Bookcode). Este sistema se programó en una computadora personal (PC). El Laboratorio de Procesamiento de Voz de la Facultad de Ingeniería de la UNAM cuenta con una larga trayectoria en el área de reconocimiento de voz, por lo que el autor reconoce la contribución en esta etapa del trabajo de tesis de la experiencia adquirida por el laboratorio a través de las siguientes publicaciones y que se recomienda al lector interesado consultar: [Her95] en lo que respecta a detección de inicio y fin de palabra, [Her00] y [Her01-01] en lo que respecta a reconocimiento por VQ, [Her01-02] en lo referente a preprocesamiento, y [Nie03] sobre temas varios de reconocimiento de voz.

El sistema de entrenamiento se divide en dos etapas: entrenamiento y reconocimiento. La etapa de entrenamiento obtiene los patrones de referencia de un conjunto de palabras (repeticiones) de entrenamiento. La etapa de reconocimiento emplea el conjunto de patrones de referencia y a través de medidas de distorsión entre dicho conjunto y la palabra de entrada en esta etapa, determina la palabra reconocida.

Las figuras 12.1 y 12.2 ilustran los diagramas de bloques que muestran, de forma general, los procesos que se realizaron en el entrenamiento y reconocimiento de palabras aisladas, respectivamente.

Como se observa en los diagramas, existe un módulo común en ambas etapas: el preprocesamiento. Por esta razón, el preprocesamiento se manejará como un módulo independiente, para efecto del análisis.

Como se muestra en la figura 12.1, durante la etapa de entrenamiento se capturan diferentes señales de voz, a la que denominaremos repeticiones de palabras. A cada repetición se le aplica preprocesamiento para obtener una palabra delimitada. Posteriormente, se agrupan todas las palabras recortadas y con ellas se obtienen sus centroides correspondientes. Estos centroides se guardan en un archivo para posteriormente cargarlos en memoria al ejecutar el reconocedor.

En la etapa de reconocimiento, como se observa en la figura 12.2, se captura la palabra que se desea reconocer. A esta palabra, también se le aplica preprocesamiento, para recortarla. Utilizando los centroides, calculados durante el entrenamiento y almacenados en la memoria, se realizan las comparaciones necesarias, para efectuar el reconocimiento. El éxito del evento dependerá de ciertos parámetros estadísticos obtenidos mediante el análisis de una población de resultados.

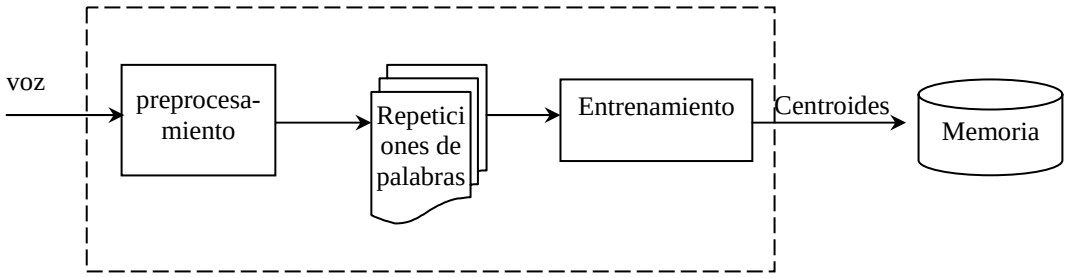


Figura 12.1: Etapa de Entrenamiento

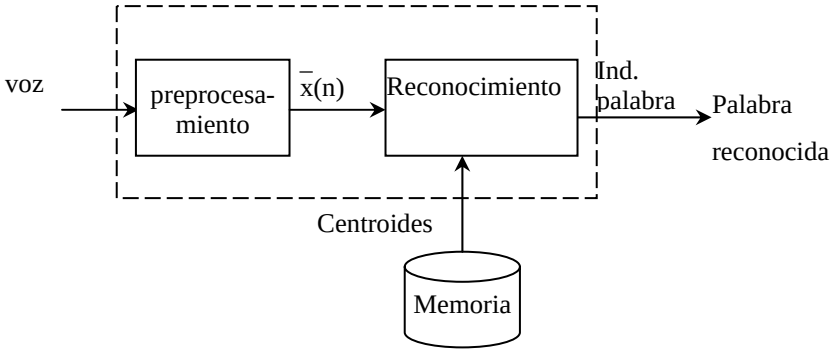


Figura 12.2: Etapa de Reconocimiento

12.1 Preprocesamiento

El preprocesamiento consta de los siguientes métodos:

1. Filtrado paso bajas y submuestreo. Esto fue llevado a cabo con el fin de analizar el rango frecuencial más importante de la señal de voz (0-5kHz) (e incluso eliminar ruidos de altas frecuencias) así como procesar un menor número de muestras. Se empleó un filtro Chebyshev tipo II de orden 24 con frecuencia de banda de rechazo (stopband) de 5,512 Hz. Tras el filtrado, la señal de voz fue submuestreada por un factor de 4 (conservando cada cuarta muestra empezando desde la primera de la señal de entrada). De esta manera la frecuencia de muestreo final se modificó de 44,100 Hz a 11,025 Hz. Cabe mencionar que el filtro paso bajas además evita el efecto de traslape espectral.

2. Preénfasis. Con el fin de realzar las frecuencias altas presentes en la voz, donde a=0.95.

3. Determinación de los límites de la señal de voz. Se empleó el algoritmo de detección de inicio y fin de palabra.

4. División de la señal en tramas de 256 muestras (23.22 ms), a cada trama se le aplica una ventana de tipo Hamming, para suavizar el espectro. El traslape entre ventanas es de un 15%, por ende el corrimiento es de 218 muestras (19.77 ms).

La figura 12.3 muestra el diagrama de bloques para este módulo.

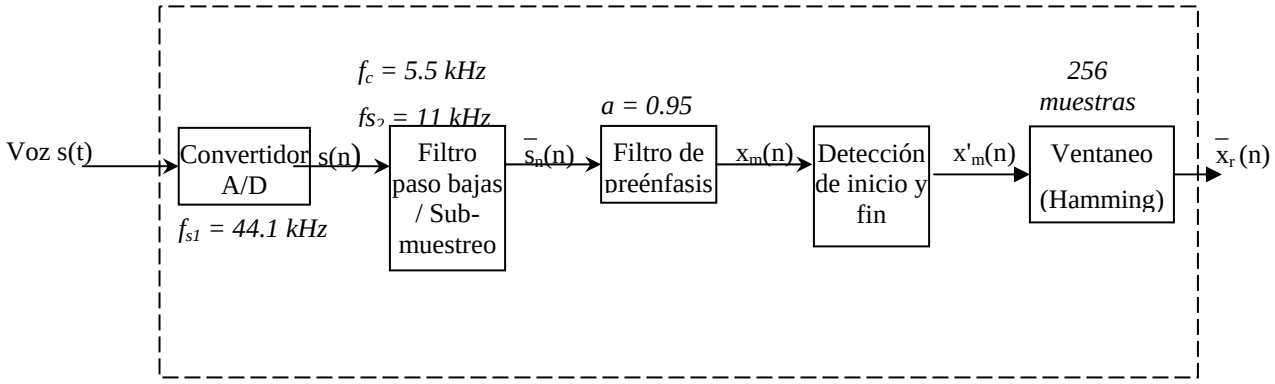


Figura 12.3 Diagrama de bloques para el módulo de preprocesamiento

Como ya se mencionó, el preprocesamiento es un módulo común en el entrenamiento y en el reconocimiento. Sin embargo, existe una diferencia sutil: durante el entrenamiento, cada palabra recortada se almacena en un archivo independiente que posteriormente se utilizará para determinar los centroides representativos. Mientras que en el reconocimiento la palabra recortada se utiliza directamente para hacer las comparaciones con los centroides.

12.2 Entrenamiento

Para el entrenamiento, se deben capturar diversas palabras recortadas, por medio del módulo de preprocesamiento, que representan las diferentes repeticiones. A partir de estas repeticiones, se obtiene un conjunto de centroides que se guardan en un archivo y cargarlos en memoria al ejecutar el reconocedor.

Los archivos de palabras, previamente almacenados, son utilizados para aplicarles varios procesos a cada uno. Primero se determinan sus vectores de autocorrelación para obtener los coeficientes de predicción lineal. Enseguida, estos coeficientes son segmentados linealmente en 4 partes. Se sigue el mismo procedimiento para todas las repeticiones de una misma palabra. Posteriormente, una vez que se tienen segmentadas todas las repeticiones, se agrupan en conjuntos del mismo segmento. Para cada conjunto se obtienen centroides, que representarán al segmento. Finalmente, estos centroides son guardados en un archivo para su posterior empleo. La figura 12.4 esquematiza el proceso completo del entrenamiento. Los centroides/vectores de referencia se almacenaron en archivos independientes.

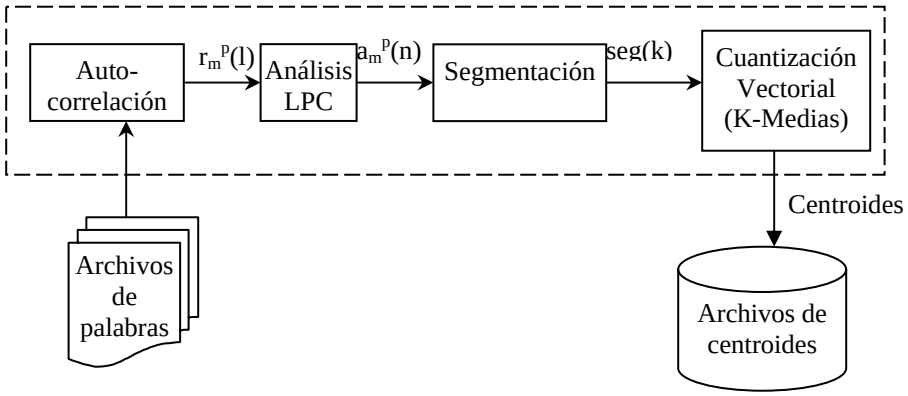


Figura 12.4 Módulo de Entrenamiento

El entrenamiento, es la etapa que nos permite obtener los patrones de comparación, para las señales de voz. Consta de los siguientes módulos: autocorrelación, análisis LPC, segmentación y cuantización vectorial.

12.3 Reconocimiento

El reconocimiento, recibe las tramas de una señal de voz, proveniente del preprocesamiento, para efectuar comparaciones con los centroides y obtener un indicador a la palabra reconocida.

Cada trama recibida es utilizada para calcular su autocorrelación y sus coeficientes de predicción lineal. Este proceso se repite hasta que el preprocesamiento detecta el fin de la palabra. Una vez detectado el final, se segmentan los vectores LPC y los vectores de autocorrelación correspondientes, de forma lineal, nuevamente 4 partes, y con segmentos iguales al patrón de comparación. Con cada segmento se realiza una comparación, utilizando la distancia de Itakura, con los centroides que correspondan al mismo segmento.

El resultado de esta comparación, aunado con ciertos parámetros estadísticos, determina el éxito o fracaso del reconocimiento. La figura 12.5 muestra el diagrama de bloques del reconocimiento.

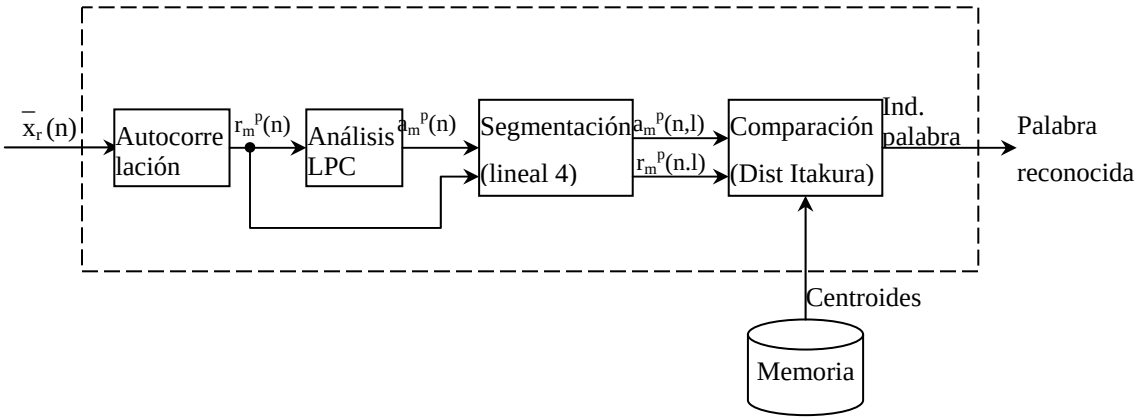


Figura 12.5: Módulo de Reconocimiento

El reconocimiento permite identificar una señal de voz similar a ciertos patrones definidos previamente. Consta de los siguientes módulos: autocorrelación, análisis LPC, segmentación y comparación.

12.4 Resultados

Para efectuar el reconocimiento de una palabra en particular, se emplearon 8 repeticiones de ella por cada uno de los 11 hablantes. Es decir que en total se tenían 88 pronunciaciones de esa palabra y que se emplearon en el reconocedor.

Adicionalmente se probó la confiabilidad del sistema reconociendo las mismas repeticiones que se emplearon para el entrenamiento; es decir, 12 repeticiones pronunciadas por cada uno de los 11 hablantes, en total 132 pronunciaciones. Reconocer las mismas realizaciones del entrenamiento es conocido como límite superior (upperbound).

Como se explicó, el reconocimiento se lleva a cabo calculando las distancias de la palabra a reconocer –extrayendo su respectiva información: valores de autocorrelación, coeficientes LPC- con los centroides del diccionario que fueron previamente calculados por el método de K-medias. Por lo tanto, cada pronunciación que ingresó al sistema fue comparada con el diccionario de 61 palabras.

Cabe mencionar que el diccionario incluye palabras fonéticamente semejantes, inclusive algunos pares mínimos.

En las siguientes páginas se presentan las matrices de confusión del sistema de reconocimiento tanto del límite superior como de las nuevas realizaciones. Se ha preferido emplear la traducción al español para una consulta más rápida. Las matrices se interpretan de la siguiente manera: los renglones indican las pronunciaciones que entraron al sistema de reconocimiento, las columnas indican las palabras reconocidas. Por ejemplo, de 88 repeticiones de la palabra [ámo'], 87 fueron reconocidas correctamente, mientras que 1 fue reconocida equivocadamente como [á:t] (agua).

MATRÍZ DE CONFUSIÓN LÍMITE SUPERIOR – UPPERBOUND (PARTE 1/2)

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
		agua	alguien	allá	blanco	caliente	casa	cierto	cinco	comida	con	cuánto	cuatro	despedida	día	difícil	dónde	dos	él / ella	él quiere	falso	frío	fuego	humo	joven	lejos	mañana	mentiroso	mi cabeza	mi esposa	mi hijo
1	agua	132																													
2	alguien		132																												
3	allá			131										1																	
4	blanco				131						1																				
5	caliente					132																									
6	casa						132																								
7	cierto							131																							
8	cinco								132																						
9	comida									132																					
10	con				1						131																				
11	cuánto											131																			
12	cuatro												132																		
13	despedida													132																	
14	día														132																
15	difícil															132															
16	dónde							1									131														
17	dos																	132													
18	él / ella																		131												
19	él quiere																			123											
20	falso																1				131										
21	frío																					132									
22	fuego																						131								
23	humo																							132							
24	joven																								132						
25	lejos																									130					
26	mañana																										132				
27	mentiroso																											132			
28	mi cabeza																												121		
29	mi esposa																													132	
30	mi hijo																														132
31	mi mamá																														
32	mi papá																														
33	mosquito																														
34	muerto																														
35	negro																														
36	niño																														
37	no (amo)																														
38	no (kana)																														
39	pájaro																														
40	pasado mañana																														
41	por qué																														
42	qué														1																
43	quizás																														
44	rojo																														
45	saludo																														
46	sangre																														
47	su cabeza																														
48	su pierna																														
49	tengo hambre																														
50	tengo sed																														
51	tienes sed																														
52	tierra								1																						
53	todavía no																														
54	tres																														
55	tú																														
56	tu cabeza																												11		
57	tu quieres																			6											
58	uno																														
59	yo																														
60	yo como																														
61	yo quiero																			2											

MATRÍZ DE CONFUSIÓN LÍMITE SUPERIOR – UPPERBOUND (PARTE 2/2)

31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61		
mi mamá	mi papá	mosquito	muerto	negro	niño	no (amo)	no (kana)	pájaro	pasado mañana	por qué	qué	quizás	rojo	saludo	sangre	su cabeza	su pierna	tengo hambre	tengo sed	tienes sed	tierra	todavía no	tres	tú	tu cabeza	tu quieres	uno	yo	yo como	yo quiero		
																															agua	1
																															alguien	2
																															allá	3
																															blanco	4
																															caliente	5
																															casa	6
																														1	cierto	7
																															cinco	8
																															comida	9
																															con	10
																			1												cuánto	11
																															cuatro	12
																															despedida	13
																															día	14
																															difícil	15
																															dónde	16
																															dos	17
																		1													él / ella	18
																										7				2	él quiere	19
																															falso	20
																															frío	21
				1																											fuego	22
																															humo	23
																															joven	24
									2																						lejos	25
																															mañana	26
																															mentiroso	27
																									11						mi cabeza	28
																															mi esposa	29
																															mi hijo	30
132																															mi mamá	31
8	124																														mi papá	32
		131																							1						mosquito	33
			132																												muerto	34
				132																											negro	35
					131																			1							niño	36
						132																									no (amo)	37
							132																								no (kana)	38
								132																							pájaro	39
									132																						pasado mañana	40
										132																					por qué	41
											131																				qué	42
												132																			quizás	43
													132																		rojo	44
														132																	saludo	45
															132																sangre	46
																132															su cabeza	47
																	132														su pierna	48
																		132													tengo hambre	49
																			132												tengo sed	50
																			5	127											tienes sed	51
																					131										tierra	52
																						132									todavía no	53
										1													131								tres	54
																								132							tú	55
																1									120						tu cabeza	56
																										118				8	tu quieres	57
																											132				uno	58
																												132			yo	59
																													132		yo como	60
																										5				125	yo quiero	61

MATRÍZ DE CONFUSIÓN RECONOCEDOR (PARTE 1/2)

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
		agua	alguien	allá	blanco	caliente	casa	cierto	cinco	comida	con	cuánto	cuatro	despedida	día	difícil	dónde	dos	él / ella	él quiere	falso	frío	fuego	humo	joven	lejos	mañana	mentiroso	mi cabeza	mi esposa	mi hijo
1	agua	88																													
2	alguien		88																												
3	allá			86										1																	
4	blanco				88																										
5	caliente					88																									
6	casa						86														1										
7	cierto							85																							
8	cinco								87																						1
9	comida									88																					
10	con	1									85																				
11	cuánto											88																			
12	cuatro												88																		
13	despedida										1			87																	
14	día														87																
15	difícil															88															
16	dónde																84				4										
17	dos																	88													
18	él / ella																		83												
19	él quiere																			77											
20	falso							1									1				86										
21	frío																					88									
22	fuego																						88								
23	humo																							88							
24	joven																								88						
25	lejos																									86	1				
26	mañana																										88				
27	mentiroso																											88			
28	mi cabeza																												67		
29	mi esposa																													88	
30	mi hijo																														88
31	mi mamá																														
32	mi papá																														
33	mosquito																														
34	muerto																														
35	negro																														
36	niño																														
37	no (amo)	1																													
38	no (kana)																														
39	pájaro																														
40	pasado mañana																														
41	por qué																														
42	qué																												1		
43	quizás											1																			
44	rojo																														
45	saludo																														
46	sangre																														
47	su cabeza																												2		
48	su pierna																														
49	tengo hambre																														
50	tengo sed																														
51	tienes sed																														
52	tierra	1																													
53	todavía no																2														
54	tres																														
55	tú																		1												
56	tu cabeza																												11		
57	tu quieres																			9											
58	uno																														
59	yo																														
60	yo como																														
61	yo quiero																			5											

MATRÍZ DE CONFUSIÓN RECONOCEDOR (PARTE 2/2)

31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61		
mi mamá	mi papá	mosquito	muerto	negro	niño	no (amo)	no (kana)	pájaro	pasado mañana	por qué	qué	quizás	rojo	saludo	sangre	su cabeza	su pierna	tengo hambre	tengo sed	tienes sed	tierra	todavía no	tres	tú	tu cabeza	tu quieres	uno	yo	yo como	yo quiero		
																				1			1					1		1	agua	1
																															alguien	2
																		1													allá	3
																															blanco	4
																															caliente	5
																				1											casa	6
																							1					1		1	cierto	7
																															cinco	8
																															comida	9
									1																				1		con	10
																															cuánto	11
																															cuatro	12
																															despedida	13
						1																									día	14
																															difícil	15
																															dónde	16
																															dos	17
			1																1									1		2	él / ella	18
																										6				5	él quiere	19
																															falso	20
																															frío	21
																															fuego	22
																															humo	23
																															joven	24
									1																						lejos	25
																															mañana	26
																															mentiroso	27
																									21						mi cabeza	28
																															mi esposa	29
																															mi hijo	30
88																															mi mamá	31
	88																														mi papá	32
		88																													mosquito	33
			88																												muerto	34
				88																											negro	35
					85															1			2								niño	36
						87																									no (amo)	37
							88																								no (kana)	38
								88																							pájaro	39
									88																						pasado mañana	40
										88																					por qué	41
											87																				qué	42
												87																			quizás	43
													88																		rojo	44
														88																	saludo	45
															88																sangre	46
																86															su cabeza	47
																	88														su pierna	48
																		88													tengo hambre	49
																			83	5											tengo sed	50
																			5	83											tienes sed	51
																					87										tierra	52
																						86									todavía no	53
					1					4										1			82								tres	54
																								87							tú	55
																									77						tu cabeza	56
																										69				10	tu quieres	57
																											88				uno	58
																								2				86			yo	59
																													88		yo como	60
																										10				73	yo quiero	61

Los porcentajes de reconocimiento del sistema considerando todas sus respectivas pronunciaciones son:

Límite superior: 98.98% (8052 pronunciaciones; de ellas, 7970 bien reconocidas,).
Reconocedor: 97.52% (5368 pronunciaciones; de ellas, 5235 bien reconocidas).

Donde "reconocedor" se refiere al reconocimiento de nuevas pronunciaciones no empleadas en el entrenamiento.

Los resultados de reconocimiento son bastante satisfactorios. Vale la pena ahondar en los resultados para indagar cuáles son las palabras con el mayor número de errores de reconocimiento y analizar la razón de dichos errores.

Análisis del límite superior

Las palabras con 5 o más repeticiones mal reconocidas se presentan en la tabla 13.1.

Palabra (traducción)	Palabra(s) con la(s) que hubo confusión
[tiknéki'] (tu quieres)	[niknéki'], [kinéki']
[mo¢ontéko] (tu cabeza)	[no¢ontéko], [i¢ontéko] (su cabeza)
[no¢ontéko] (mi cabeza)	[mo¢ontéko]
[kinéki'] (el quiere)	[tiknéki'], [niknéki']
[nopá] [nópa] (mi papá)	[nomá] [nóma] (mi mamá)
[niknéki'] (yo quiero)	[tiknéki'], [kinéki']
[tiamíki'] (tienes sed)	[niamíki'] (tengo sed)

Tabla 13.1: Palabras con mayores problemas de reconocimiento (límite superior)

Es claro que el principal problema en el límite superior se debió a las palabras con pares mínimos y gran semejanza fonética.

Análisis del reconocedor

Se sigue el mismo procedimiento que en el anterior análisis, ahora con nuevas pronunciaciones, para el caso de palabras con 4 o más repeticiones mal reconocidas. Los datos se presentan en la tabla 13.2.

Palabra (traducción)	Palabra(s) con la(s) que hubo confusión
[no¢ontéko] (mi cabeza)	[mo¢ontéko]
[tiknéki'] (tu quieres)	[niknéki'], [kinéki']
[niknéki'] (yo quiero)	[tiknéki'], [kinéki']
[mo¢ontéko] (tu cabeza)	[no¢ontéko]
[kinéki'] (el quiere)	[tiknéki'], [niknéki']
[éyi'] (tres)	[keyé] (por qué), [tiamíki'] (tienes sed), [píli'] (niño)
[tiamíki'] (tienes sed)	[niamíki']
[niamíki'] (tengo sed)	[tiamíki']
[yéh] (él, ella)	[niknéki'], [néh] (yo), [niamíki'], [miket]
[káni'] (dónde)	[kanéli'] (falso)

Tabla 13.2: Palabras con mayores problemas de reconocimiento (reconocedor)

Nuevamente se observa que los mayores problemas de reconocimiento se presentan con las palabras fonéticamente semejantes. Además se nota en la matriz de confusión que las pronunciaciones mal reconocidas están más dispersas en ella.

13. CONCLUSIONES

En este trabajo se ha presentado la fonética-acústica del náhuatl de San Miguel Tzinacapan, también se desarrolló un reconocedor de comandos de voz para el dialecto, de dicha lengua, hablado en dicha localidad. Para realizar esta tarea se tuvo que adquirir la materia prima realizando las grabaciones directamente en este poblado ubicado en la Sierra Norte del Estado de Puebla. Es muy valioso conocer el sistema fonológico de esta lengua para así formular un cuestionario que contenga todos los sonidos lingüísticos existentes. Sin embargo, para realizar un estudio fonético-acústico es importante tener claridad sobre las interacciones de interés de los sonidos lingüísticos y procurar que el cuestionario tome esto en cuenta también.

Esta labor pudiera ser complicada en el caso de lenguas que no han sido muy estudiadas; aún así, aunque se trate de una lengua estudiada, pueden existir dialectos de dicha lengua de los que tenga poca información. Afortunadamente se contó con diversas fuentes de consulta para este trabajo.

Por otra parte, cabe resaltar que la labor de preprocesamiento de las grabaciones, en especial la segmentación, toma un tiempo considerable. Por ejemplo, el algoritmo de inicio y fin de palabra presentado en secciones pasadas tuvo que modificarse para que detectara varios inicios y fines de palabras debido a que cada archivo de audio original contenía varias repeticiones. La meta era que cada repetición se guardara en un archivo diferente para ir creando el corpus o base de datos. Es muy conveniente realizar ajustes al algoritmo mencionado por las razones descritas en la sección 3.2.3. Con estos ajustes se llegó a un detector (y segmentador) más preciso de varios inicios y fines de palabras para ambientes relativamente controlados; pero no se llegó a un detector infalible, pues la variación del ruido de fondo en segmentos de silencio y la variación de las propias pronunciaciones ocasionalmente orillaban a falsas detecciones, aunque ciertamente muy pocas.

Dado lo anterior, para un vocabulario de 61 palabras, 11 hablantes y al menos 20 repeticiones, se emplearon 13420 pronunciaciones que fueron guardadas independientemente en el mismo número de archivos.

A continuación se presentan conclusiones más específicas sobre los dos temas principales de este trabajo.

13.1 Estudio fonético-acústico

Se comenzó describiendo las propiedades acústicas de cada fonema y que distinguen a un sonido de otro. Prestando atención a la estructura silábica del náhuatl, dichas propiedades fueron estudiadas de dos maneras, con los fonemas en posición onset y coda. En posición onset los rasgos distintivos de los fonemas están completamente caracterizados, mientras que en posición coda ocurren debilitamientos y pérdida de propiedades.

Después de la presentación de acústica estática, se mostró cómo la acústica de los fonemas se ve afectada por sus sonidos vecinos y cómo cambian debido a la coarticulación o movimiento continuo de los articuladores.

Precisamente una consecuencia de la coarticulación es que a menudo es imposible determinar dónde termina un sonido y empieza el siguiente. Hemos visto que algunos sonidos exhiben discontinuidades cuando se cierra completamente el tracto vocal, tanto en el momento de cierre como en su liberación (como por ejemplo los sonidos nasales). Resulta atractiva la idea de considerar esa discontinuidad como la unión entre los fonemas, sin embargo la discontinuidad no es mas que un indicativo de transición, siendo insuficiente para señalar la frontera de un fonema debido a que otras propiedades del mismo fonema se manifiestan en otros lugares de la señal. Por ejemplo, evidencias del lugar de articulación del cierre ocurren antes del momento del cierre, por lo que basarnos en el cierre no constituye una manera correcta de identificar la frontera de un fonema. Pensemos en un sonido nasal, donde el formante nasal se manifiesta ya desde la vocal que lo antecede, allí tenemos evidencia de un sonido nasal antes de la discontinuidad (y del cierre total del tracto vocal). Las fronteras entre sonidos sonoros y fricativos (y viceversa) tampoco son muy marcadas debido a que el comienzo y fin del ruido fricativo es gradual, conforme los articuladores formen una constricción más y más (o menos y menos) estrecha; por lo tanto es difícil determinar la frontera exacta entre ellos. Hemos visto también transiciones entre sonidos sonoros que tienen transiciones largas, ciertamente es posible muchas veces identificar a la señal en un instante del tiempo como un sonido u otro, pero no es posible determinar la frontera exacta entre dos sonoras.

Otra consecuencia de la coarticulación es la variación en la acústica de los fonemas. Los fonemas se ven influenciados por sus fonemas vecinos y sus rasgos son algo diferentes de un contexto a otro. Un escucha constantemente usa el contexto para determinar los diferentes sonidos y qué es lo que se dice.

También se ha mostrado que los valores absolutos de los formantes o configuraciones espectrales no determinan la identidad de los fonemas, sino es el movimiento (o falta de movimiento) de los formantes lo que determina sus identidades; de esa manera podemos identificar a la vocal /i/ de la aproximante /y/ aún si los formantes no alcanzan sus targets.

Por lo tanto, el habla no consiste de secuencias de sonidos discretos, tampoco hay fronteras exactas entre los correlatos acústicos de los fonemas. El habla consiste de sonidos continuos entretejidos y dinámicos.

Surgen varias recomendaciones para trabajos futuros, una de ellas es analizar el acento, averiguar si hay diferentes grados del mismo o quizás, a pesar de que se sabe que en el náhuatl la penúltima sílaba es la acentuada, el acento se distribuya en sonidos vecinos. Otro trabajo futuro podría ser el análisis de los elementos suprasegmentales del náhuatl (p.e. la entonación), para esto resultaría conveniente grabar oraciones procurando que haya un alto grado de espontaneidad en las pronunciaciones.

También podría realizarse el estudio fonético-acústico de las variaciones dialectales de la lengua náhuatl. Esto es un trabajo muy amplio, por lo que sería posible acotarlo concentrándose en fonemas que sean de mayor interés.

En cuanto a tecnologías de procesamiento de voz, la información obtenida del presente estudio fonético-acústico del náhuatl podría incorporarse al desarrollo de sistemas de conversión de texto a habla (síntesis de voz) así como la creación de diccionarios de unidades acústicas. En lo que respecta a sistemas de reconocimiento de voz, a través del presente estudio fonético-acústico se pueden determinar las unidades de reconocimiento si se planteara que el sistema sea sensible a los fonemas del náhuatl.

13.2 Reconocimiento de voz por VQ

Uno de los objetivos principales de este trabajo fue analizar el comportamiento del método LPC para el reconocimiento de palabras en lengua náhuatl. Un rasgo característico del náhuatl es la presencia del saltillo, que, aunque no sea fonema en el dialecto náhuatl de San Miguel Tzinacapan, es un elemento muy presente en el habla. Por lo tanto, las palabras segmentadas incluyeron el saltillo tanto en el entrenamiento como en el reconocimiento. Cabe mencionar que en el proceso de segmentación, el detector de inicio y fin de palabra detectaba el saltillo en la mayoría de los casos. Es de notar que el saltillo podía tener grados de intensidad muy diversos, muchas veces estaba muy bien marcado, pero en otras era muy sutil. Además de eso, el tiempo entre el final de la palabra y el saltillo también variaba.

Sin duda que, además de la variabilidad propia de la voz, es saltillo es un elemento particular que desempeñaría un papel en la tarea de reconocimiento.

Los porcentajes de reconocimiento del sistema han sido bastante satisfactorios: Límite superior, 98.98%; reconocedor, 97.52%.

El sistema de reconocimiento encuentra dificultades para reconocer palabras del diccionario fonéticamente semejantes, incluyendo algunos pares mínimos.

Esto no quiere decir de ninguna manera que el sistema de reconocimiento haga mal su papel, por el contrario, el reconocimiento lo lleva a cabo en un porcentaje esperado por esta técnica de cuantización vectorial empleando coeficientes LPC.

En conclusión, el desempeño del reconocimiento de palabras aisladas en náhuatl por medio de cuantización vectorial, manipulando coeficientes LPC, se lleva a cabo de manera muy satisfactoria y con tasas muy semejantes a estos mismos métodos de reconocimiento empleados en otros proyectos de reconocimiento de voz del español de México y que han sido llevados a cabo en el Laboratorio de Procesamiento de Voz.

Como una mejora se debe trabajar en el uso de algoritmos que permitan una mejor tasa de reconocimiento de palabras fonéticamente semejantes. Esta propuesta cobra mayor peso debido a que en la lengua náhuatl muchos sustantivos son poseídos, como las partes del cuerpo o parentescos, y la diferencia fonética es pequeña (p.e. [nome¢] (mi pierna), [mome¢] (tu pierna)); aún más, los verbos conjugados también son muy semejantes fonéticamente.

Por otro lado, si se considerara la aplicación de un sistema de reconocimiento de náhuatl por VQ deben establecerse umbrales de rechazo que permitan discriminar palabras que no pertenezcan al vocabulario del reconocedor. Para ello se recomienda consultar las referencias mencionadas al inicio del capítulo 12.

Como siguiente trabajo se podría abordar el reconocimiento de palabras continuas en náhuatl y posteriormente abordar proyectos más complejos como la traducción automática del

habla entre el español de México y el náhuatl; este proyecto está fuertemente considerado por el Laboratorio de Procesamiento de Voz de la Facultad de Ingeniería de la UNAM.

APÉNDICES

Apéndice 1. Cuestionario

Se presenta el cuestionario elaborado. Note que hay palabras que no fueron consideradas en el diccionario del reconocedor, esto es debido a varias razones: no se contaba con el suficiente número de repeticiones o de hablantes; también hubo variaciones pero no alcanzaban el número suficiente para incorporarlas al reconocedor. También observe que se grabaron algunas oraciones; sin embargo, como no fueron empleadas para el reconocedor, se grabaron con un número de hablantes y repeticiones limitados.

Como indica Launey [Lau92], "la ortografía náhuatl nunca ha sido fijada realmente", por esta razón se ha preferido únicamente emplear la transcripción fonética.

	Español	Náhuatl (Fonético)
1	lluvia	[kiówit]
2	camino	[óhti'] [óhti]
3	fuego	[tít]
4	agua	[á:t]
5	tierra	[tá:l]
6	comida	[tapálo:l]
7	casa	[káli] [káli']
8	hombre	[tákat] [tágat]
9	mujer	[síwat]
10	mi esposo	[notákaw]
11	mi esposa	[nosíwaw]
12	mi mamá	[nóma] [nomá]
13	mi papá	[nópa] [nopá]
14	mi hijo (para niño pequeño)	[nopíli']
15	niño	[píli']
16	no	[kána]
17	no	[ámo']
18	sí	[kéma]
19	frío	[sések]
20	caliente	[totónik]
21	joven	[okíčpil]
22	difícil	[ówi]
23	fácil	[ámo ówi]
24	humo	[pókti'] [pókti]
25	sol	[tonál¢in] [tonál¢i]
26	luna	[mé¢ti'] [mésti']
27	día	[tónal]
28	noche	[yówak]
29	alacrán	[kólot]

30	árbol	[kówit] [k^{w}ówit]
31	sangre	[ésti']
32	pájaro	[číkte]
33	mosquito	[móyot]
34	yo como	[niták^{w}a']
35	tu comes	[titák^{w}a']
36	yo bebo	[nitái'] [nitáyi']
37	tu bebes	[titái'] [titáyi']
38	tienes sed	[tiamíki']
39	quizás	[á:it]
40	alguien	[aksá']
41	ahora	[yekín¢in] [yekín¢i]
42	todavía no	[ká:ya']
43	con	[íka']
44	así	[ihkó] [ihkón]
45	dónde	[káni']
46	cuánto	[kanáči'] [kanáči]
47	qué	[tóni']
48	quién	[akóni']
49	por qué	[keyé]
50	saludo local	[niów]
51	despedida local	[timóta]
52	gracias	[tasohkamátik]
53	mi cabeza	[no¢ontéko]
54	tu cabeza	[mo¢ontéko]
55	su cabeza	[i¢ontéko]
56	mi brazo	[nomái] [nómai]
57	tu brazo	[momái] [mómai]
58	su brazo	[imái] [ímai]
59	mi pierna	[nóme¢]
60	tu pierna	[móme¢]
61	su pierna	[íme¢]
62	uno	[sé']
63	dos	[óme']
64	tres	[éyi']
65	cuatro	[náwi']
66	cinco	[mák^{w}il]
67	aquí	[níka] [níkan]
68	allá	[népa']
69	cerca	[kawéhka'] [ámo wéhka']
70	lejos	[wéhka']
71	ayer	[yálwa']
72	hoy	[áma] [áman]
73	mañana	[mósta']
74	antier	[yawípta']
75	pasado mañana	[wípta']

76	yo	[néh]
77	tú	[téh]
78	él/ella	[yéh]
79	nosotros	[téhwan] [téhwa]
80	ustedes	[naméhwan] [naméhwa]
81	ellos	[yéhwan] [yéhwa]
82	yo quiero	[niknéki']
83	tu quieres	[tiknéki']
84	él quiere	[kinéki']
85	¿dónde está la olla?	[káni yétok kómit]
86	tengo sed	[niamíki']
87	tengo frío	[nisék^{w}i']
88	muerto	[míket]
89	tengo hambre	[nimayána']
90	te dijo	[mi¢íli] [mi¢ílwi]
92	me dijo	[netčíli] [netčílwi]
94	le dijo	[kíli] [kílwi]
96	mentiroso	[šolópi]
97	cierto	[néli']
98	falso	[kanéli']
99	estoy enfermo	[nimokokówa']
100	negro	[tíltik]
101	rojo	[čičíltik]
102	azul	[ilwikátik]
103	blanco	[ístak]
104	amarillo	[kóstik]
105	verde	[xoxóktik]
106	café	[táltik]
107	¿dónde vives?	[kanintinémi'] [kanitinémi'] [kantinémi']
108	vivo en San Miguel Tzinacapan	[ninémi san migél ¢inakápan]
109	y	[wan]
110	buenos días	[takahti]
111	buenas tardes	[tiotáki]
112	ayúdeme	[xinečpaléwi']
113	no lo sé - no entiendo	[ámo nikmáti']
114	no hablo español	[ámo nitahtówa' koyotáhtol] [ámo nitahtóa koyotáhtol]
115	mi hijo (para niño mayor)	[nokónew]
116	estoy cansado	[nisiówtok]
117	quieres tomar poquita agua	[titáyis tepí¢in á:t]
118	está muy difícil este trabajo que estamos haciendo	[yék ówi in tékit téi tikčiwtoke]
119	está muy grande ese árbol	[yék wéyi né kówit
120	ese hombre es un vago	[né tákat yék némi]
121	¿Cuánto cuesta esta prenda?	[ka:náči nipátiw in tásal]
122	esa mujer tiene la ropa verde	[né síwa:t itásal šošóktik]

Apéndice 2. Relación de figuras con archivos de audio y hablantes

En el nombre del archivo de audio se puede identificar al hablante cuya pronunciación fue analizada, para ello observe el par de letras mencionadas en él y consulte la siguiente tabla:

Hablante	Correspondencia
Clara	cl
Cote	co
Ernesto	er
Esteban	es
Francisco	fr
José Heradio	jh
Linda	li
Miguel	mi
Neri	ne
Quelita	qu
Vicente	vi
Yolanda	yo

Los números corresponden al número de repetición del hablante. Cuando la información se extrajo de una oración, el nombre del archivo contiene las letras “or”.

Propiedades estáticas de los sonidos del habla
Figura 1a. buenas_tardes-er-01
Figura 1b. blanco-er-03
Figura 2i. casa-qu-13
Figura 3i. hombre-es-02
Figura 3ii. comida-vi-03
Figura 3iii. buenas_tardes-er-01
Figura 3iv. comida_takual-fr-05
Figura 4i. frio-er-03
Figura 4ii. verde-jh-03
Figura 5i. no_hablo_espanol-er-04
Figura 5ii. camino-vi-05
Figura 6i. nosotros-jm-04
Figura 6ii. estoy_cansado-ne-04
Figura 7i. rojo-er-04
Figura 7ii. su_cabeza-er-02
Figura 8i. hoy-jh-05
Figura 8ii. tengo_hambre-jh-05
Figura 8iii. tengo_hambre-jh-03
Figura 9i. mujer-vi-08
Figura 9ii. todavia_no-jh-05
Figura 9iii. niño-jh-05
Figura 9iv. comida-jh-04
Figura 10 (hablante masculino). fuego-jh-10

Figura 10 (hablante masculino). uno-jh-02
Figura 10 (hablante masculino). caliente-jh-06
Figura 10 (hablante masculino). agua-jh-03
Figura 10 (hablante femenina). fuego-li-02
Figura 10 (hablante femenina). uno-li-11
Figura 10 (hablante femenina). caliente-li-01
Figura 10 (hablante femenina). agua-li-08
Figura 11. comida-er-03
Figura 12 (superior). es-or-05e-01
Figura 12 (central). mi-or-05e-01
Figura 12 (inferior). agua-er-06
Figura 13i. antier-ne-03
Figura 13ii. antier-es-03
Figura 13iii. hombre-li-03
Figura 13iv. humo-qu-04
Figura 14. sangre-er-02
Figura 15i. joven-er-03
Figura 15ii. joven-jm-04
Figura 16i. luna-jm-03
Figura 16ii. luna-li-03
Figura 17i. donde_vives-jh-03
Figura 17ii. es-or-05e-01
Figura 18i. mi_esposo-jh-03
Figura 18ii. mi_esposa-er-07
Figura 19i. sol-es-03
Figura 19ii. sol-cl-04
Figura 19iii. sol-er-03
Figura 20i. mi_hijo-li-03
Figura 20ii. si-li-03
Figura 20iii. su_cabeza-er-03
Figura 21. camino-jm-03

Transiciones vocálicas
Figura 1i. por_que-jh-02
Figura 1ii. no_hablo_espagnol-er-03
Figura 1iii. todavia_no-jh-05
Figura 2i. mujer-jh-03
Figura 2ii. ayudeme-jm-02
Figura 2iii. no_hablo_espagnol-jm-03
Figura 2iv. cerca-jh-04
Figura 3i. tres-er-04
Figura 3ii. por_que-jh-02
Figura 3iii. no_hablo_espagnol-er-03
Figura 3iv. tengo_hambre-jh-03
Figura 4i. dificil-jh-02

Figura 4ii. cerca-jh-04
Figura 4iii. noche-jh-02
Figura 5i. lluvia-jh-05
Figura 5ii. tienes_sed-jh-02
Figura 6i. tengo_sed-jh-03
Figura 6ii. ayer-jh-02
Figura 7i. facil-jm-01
Figura 7ii. quizas-jh-01
Figura 7iii. es-or-07e-01
Figura 8. facil-qu-03

Transiciones obstruyentes y vocales
Figura 1i. niño-es-07
Figura 1ii. fuego-es-03
Figura 1iii. le_dijo-es-02
Figura 1iv. cinco-es-03
Figura 2i. pajaro-es-02
Figura 2ii. por_que-es-03
Figura 3i. humo-jh-03
Figura 3ii. caliente-jh-02
Figura 3iii. alacran-jh-02-2
Figura 3iv. arbol-er2
Figura 4i. comida-es-02
Figura 4ii. tierra-es-04
Figura 4iii. hombre-es-02
Figura 4iv. tu_comes-es-03
Figura 5i. antier-es-04
Figura 5ii. alla-es-05
Figura 5iii. mentiroso-es-03
Figura 5iv. vivo-es-03
Figura 6i. tu_comes-es-04
Figura 6ii. muerto-es-04
Figura 6iii. alacran-es-02
Figura 6iv. azul-es-02
Figura 7i. tu_quieres-es-02
Figura 7ii. su_cabeza-es-02
Figura 7iii. verde-es-05
Figura 7iv. hombre-es-05
Figura 8i. yo_como__nejnikua_-er-04
Figura 8ii. tengo_frio-es-04
Figura 8iii. tu_comes-es-03
Figura 9i. mujer-es-02
Figura 9ii. frio-es-05
Figura 9iii. gracias-jh-03
Figura 9iv. alguien-es-02

Figura 10i. blanco-es-03
Figura 10ii. frio-es-05
Figura 10iii. amarillo-es-03
Figura 10iv. gracias-es-05
Figura 11. mi_esposa-es-03
Figura 12. verde-jh-05
Figura 13i. mujer-qu-07
Figura 13ii. ayudeme-qu-03
Figura 14i. asi-jm-03
Figura 14ii. tu-jm
Figura 14iii. camino-jm-03
Figura 14iv. no_hablo_espagnol-jm-01
Figura 15i nosotros-jm-02
Figura 15ii. y-jm-04
Figura 16i. te_dijo-es-03
Figura 16ii. su_cabeza-es-05
Figura 16iii. cual_nombre-ne-03
Figura 17i. su_cabeza-es-03
Figura 17ii. su_pierna-es-04
Figura 17iii. mi_cabeza-es-05
Figura 18i. cuanto-es-02
Figura 18ii. rojo-es-01

Transiciones consonantes sonorantes y vocales
Figura 1i. donde_olla-es-03
Figura 1ii. dos-es-02
Figura 1iii. despedida-es-02
Figura 1iv. gracias-es-03
Figura 2i. caliente-es-05
Figura 2ii. vivo-es-03
Figura 2iii. mi_hijo-jh-03
Figura 2iv. cuanto-es-03
Figura 3i. su_pierna-es-06
Figura 3ii. vivo-es-03
Figura 3iii. donde_olla-es-03
Figura 3iv. gracias-jh-04
Figura 4i. ahora-es-05
Figura 4ii. sol-es-03
Figura 4iii. cuanto-es-03
Figura 5i. niño-es-06
Figura 5ii. ayudeme-jm-02
Figura 5iii. mentiroso-jh-03
Figura 6i. niño-es-04
Figura 6ii. falso-es-01
Figura 6iii. alacran-es-08

Figura 6iv. tierra-es-03

Interacciones consonánticas
Figura 1i. no_hablo_espagnol-qu-01
Figura 1ii. donde_olla-qu-02
Figura 2. donde_olla-qu-02
Figura 3. humo-er-05
Figura 4. pasado_magnana-qu-02
Figura 5. alguien-es-04
Figura 6. sangre-es-05
Figura 7. mi-or-05e-01
Figura 8i. no_lo_se-qu-03
Figura 8ii. no_lo_se-li-04
Figura 8iii. no_lo_se-jm-03
Figura 9i. es-or-13e-01
Figura 9ii. es-or-13e-02
Figura 10. es-or-26e-02
Figura 11i. es-or-14e-01
Figura 11ii. mi_cabeza-es-04
Figura 11iii. donde_vives-qu-02
Figura 12. esta_casa-ne-01
Figura 13. ahora-er-03
Figura 14. cafe-es-02
Figura 15i. mi-or-29e-02
Figura 15ii. es-or-23e-01
Figura 16. ayer-jh-03

Variaciones
Figura 1. sangre-jm-03
Figura 2i. aqui-jm-03
Figura 2ii. aqui-es-03
Figura 3i. sol-qu-03
Figura 3ii. humo-qu-03
Figura 4 (superior). gracias-er-04
Figura 4 (inferior). gracias-qu-04

REFERENCIAS

[Ace01] "Spoken Language Processing. A Guide to Theory, Algorithm, and System Development".
Huang, Xuedong; Acero, Alex; Hon, Hsiao-Wuen.
Prentice Hall PTR.
Primera edición, 980 pp.
2001, EUA.

[Are06] Apuntes de "Fonética y Fonología I".
Clase impartida por el Mtro. Francisco Arellanes.
Escuela Nacional de Antropología e Historia.
Otoño 2006.

[Bañ] "Diccionario Nahuatl-Español".
Bañuelos, Juan.
Escuela Telesecundaria Tetsijtsilin.
59 pp.
Sin año, México.

[Can88] "Nahuatl dialectology: survey and some suggestions".
Canger, Una.
International Journal of American Linguistics.
Vol. 54, No. 1.
Pages 28-72.
January 1988.

[Cas00] "El mundo del color en Cuetzalan: un estudio etnocientífico en una comunidad nahua".
Castillo Hernández, Mario Alberto.
Instituto Nacional de Antropología e Historia. Serie Etnología.
Primera edición, 137pp.
2000, México.

[Coh94] "Switchboard---the second year".
Cohen, J.; Gish, H.; Flanagan, J.
Technical Report /pub/caipworks2 en ftp.rutgers.edu
CAIP Summer Workshop in Speech Recognition: Frontiers in Speech Processing II.
Julio, 1994.

[Col92] "Workshop on spoken language understanding".
Cole, R.A.; Hirschman, L.

Technical Report CSE 92-014, Oregon Graduate Institute of Science & Technology
Portland, Oregon, EUA.
Septiembre, 1992.

[Cor90] "Los nahuas".
Cortés, Gabriela.
Universidad de México. Revista de la Universidad Nacional Autónoma de México.
Vol. XLV, No. 477.
Pág. 24–28.
Octubre, 1990.
México.

[Del87] "Discrete–Time Processing of Speech Signals"
Deller, John R. Jr; Proakis, John G.; Hansen, John H. L.,
Prentice Hall.
1987, EUA

[Dut97] "An Introduction to Text-to-Speech Synthesis".
Dutoit, Thierry.
Serie "Text, Speech and Language Technology", volumen 3.
Kluwer Academic Publishers.
Primera edición, 286 pp.
1997, Holanda.

[Fan95] "Alphabet recognition".
Handbook of Neural Computation.
Fanty, M; Barnard, E.; Cole, R. A.
Editorial desconocida.
Año 1995.

[Ger97] "Vector Quantization and Signal Compression".
Gersho, A.; Gray, R. M.
Kluwer Academic Publishers
Sexta Edición.
1997, EUA

[Her95] "Detección de Señales de Voz Utilizando coeficientes de Máxima Similitud".
Herrera, A.; Ramos, A; Yamasaki, K.
Memorias del XVII Congreso Internacional Académico de Ingeniería Electrónica.
Págs. 739-748.
1995.

[Her00] "A Multisection VQ Isolated Speech Recognition Method Using the KLT".
Herrera, A.; Gardida, A.
Proceedings of the International Conference on Communications, vol 2.

Págs. 1133-1136.
2000.

[Her01-01]"Reconocimiento automático de palabras aisladas usando VQ y la KLT".
Herrera, A.; Gardida, A.
Memorias del Simposio La Investigación en la Facultad de Ingeniería 1999.
Facultad de Ingeniería, UNAM.
Diskette 1.
2001.

[Her01-02]"Reconocimiento automático de palabras aisladas usando DTW".
Herrera, A.; Ibarra, R.
Memorias del Simposio La Investigación en la Facultad de Ingeniería 1999.
Facultad de Ingeniería, UNAM.
Diskette 1.
2001.

[Hor92] "Nahuatl práctico : Lecciones y ejercicios para el principiante".
Horcasitas, Fernando.
Instituto de Investigaciones Antropológicas, UNAM.
98 pp.
1992, México.

[Jak87] "La forma sonora de la lengua".
Jakobson, Roman; Waugh, Linda.
Fondo de Cultura Económica.
Primera edición, 286 pp.
1987, México.

[Ken79] "Generative Phonology. Description and Theory".
Kenstowicz, Michael; Kisseberth, Charles.
Academic Press.
Primera edición, 453 pp.
1979, USA.

[Lad05-1] "Phonetic data analysis. An introduction to fieldwork and instrumental techniques".
Ladefoged, Peter.
Blackwell Publishers.
Primera edición, tercera impresión, 224 pp.
2005, EUA.

[Lad05-2] "Vowels and Consonants".
Ladefoged, Peter.
Blackwell Publishing.
Segunda edición, 206 pp.

2005, Reino Unido.

[Lad97] “The Sounds of the World’s Languages”.
Ladefoged, Peter.
Blackwell Publishing.
Primera edición, 426 pp.
1997, Reino Unido.

[Lad75] “Three Areas of Experimental Phonetics”.
Ladefoged, Peter.
Oxford University Press.
Primera edición, cuarta impresión, 180 pp.
1975, Reino Unido.

[Las86] “Las áreas dialectales del náhuatl moderno”.
Lastra, Yolanda.
UNAM.
766 pp.
1986, México.

[Lau92] “Introducción a la lengua y a la literatura náhuatl”.
Launey, Michel.
Instituto de Investigaciones Antropológicas, UNAM.
1992, México.

[Med90] “La etnografía de México: Un cambiante y milenario mosaico de lenguas y culturas”.
Medina, Andrés.
Universidad de México. Revista de la Universidad Nacional Autónoma de México.
Vol. XLV, No. 477.
Pág. 10–18.
Octubre, 1990.
México.

[Nie03] “Mexican Spanish Speech Commands Recognition Using the TMS320C6711 DSP”.
Nieto, O.; López, V; Herrera, A.
Proceedings of GSPx-2003 International Signal Processing Conference.
CD.
2003.

[Oli93] “Acoustics of American English Speech. A Dynamic Approach”.
Olive, Joseph; Greenwood, Alice, Coleman, John.
Springer.
396 pp.
1993, EUA.

[Opp00] "Tratamiento de señales en tiempo discreto".
Oppenheim, Alan V.; Schafer, Ronald W.; Buck, John R.
Prentice Hall. Signal processing series.
Segunda edición, 873 pp.
2000, España.

[Osh00] "Speech Communications. Human and Machine".
O'Shaughnessy, Douglas.
IEEE Press.
Segunda edición, 547 pp.
2000, EUA.

[Pal94] "1993 benchmark tests for the ARPA spoken language program".
Pallett, D.; Fiscus, J.; Fisher, W.; Garofolo, J.; Lund, B.; Prysbocki, M.
Proceedings of the 1994 ARPA (*Advanced Research Projects Agency)* Human Language Technology Workshop.
Princeton, New Jersey, EUA.
Marzo, 1994.
Páginas 49-74.

[Qui92] "Curso de fonética y fonología españolas".
Quilis, Antonio; Fernández, Joseph.
Consejo Superior de Investigaciones Científicas.
Décimo cuarta edición, 223 pp.
1992, España.

[Qui99] "Tratado de fonología y fonética españolas".
Quilis, Antonio.
Gredos.
Segunda edición, 558 pp.
1999, España.

[Rab78] "Digital Processing of Speech Signals".
Rabiner, Lawrence; Schafer, Ronald W.
Prentice Hall.
Primera edición, 512 pp.
1978, EUA.

[Rab93] "Fundamentals of Speech Recognition".
Rabiner, Lawrence; Juang, Biing-Hwang.
Prentice Hall PTR.
Primera edición, 507 pp.
1993, EUA.

[Sem1] Apuntes de Lengua Náhuatl Oral.

Seminario impartido por el Mtro. Leopoldo Valiñas.
Instituto de Investigaciones Antropológicas, UNAM.
Otoño 2005.

[Tou84] "Vocabulario mexicano de Tzinacapan. Sierra Norte de Puebla".
Toumi, Sybille.
Chantiers Amerindia.
1984, Francia.

[Web] "Webster's New Encyclopedic Dictionary"
BD&L Press.
1995, EUA.

[Zue90] "The MIT SUMMIT speech recognition system: A progress report".
Zue, V.; Glass, J.; Phillips, M.; Seneff, S.
Proceedings of the Third DARPA (Defense Advanced Research Projects Agency) Speech and Natural Language Workshop.
Hidden Valley, Pennsylvania, EUA.
Junio, 1990.

Sitios de internet.

[Cdi] "¿Qué lengua hablas?".
Comisión Nacional para el Desarrollo de los Pueblos Indígenas.
México.
http://www.cdi.gob.mx/conadepi/index.php?option=news&task=viewarticle&sid=107

[Deu] "Apuntes elementales de fonética".
Universidad de Deusto, España.
http://paginaspersonales.deusto.es/airibar/Fonetica/Apuntes/02.html

[Dis] "El oído".
Discapnet.
http://usuarios.discapnet.es/ojo_oido/los_padres_del_ni%C3%B1o_sordo.htm

[Lli05] "Corpus Orales".
Llisterri, Joaquim.
Curso impartido en la UNAM, enero 2005.
http://homepage.mac.com/joaquim_llisterri/language_resources/UNAM_05/UNAM_05.html.

[Mag01] "Sistema de producción del sonido"
Magallanes, Carlos.
Revista electrónica: El Profesor de Ciencias.
Número 15, diciembre 2001

Facultad de Ciencias Físico Matemáticas y Naturales.
Universidad Nacional de San Luis, Argentina.
http://bib0.unsl.edu.ar/baea/prof-cs/numero15/index.html

[Ogi] "Survey of the State of the Art in Human Language Technology".
Center for Spoken Language Understanding. Oregon, USA.Cambridge University Press.
Año 1996.
http://cslu.cse.ogi.edu/HLTsurvey/

[Pra] Praat: doing phonetics by computer.
Boersma, Paul; Weenink David
http://www.praat.org/

[San06] "Otitis media secretora"
Sánchez, M. de los Angeles.
Revista Electrónica de PortalesMedicos.com
Mayo, 2006.
http://www.portalesmedicos.com/publicaciones/articles/163/1/Otitis-Media-Secretora.html

[Sil1] "Syllables and stress in Nahuatl".
Tuggy, David.
Summer Institute of Linguistics.
http://www.sil.org/Mexico/nahuatl/23i-SyllablesNah.htm

[Sil2] "What is a syllable?".
Summer Institute of Linguistics.
http://www.sil.org/LINGUISTICS/GlossaryOfLinguisticTerms/WhatIsASyllable.htm

[Sil3] "Fenómenos que ocurren en los extremos de la sílaba, palabra o enunciado".
Marlett, Stephen.
Summer Institute of Linguistics.
http://www.sil.org/capacitar/fonologia/cursos/M2004/Marlett2005-33.pdf