ETIQUETACIÓN DEL *CORPUS DIMEx100*

INTRODUCCIÓN

El *Corpus* DIMEx100 es un corpus oral trascrito en alófonos y palabras, con la finalidad de construir modelos acústicos y diccionarios de pronunciación. Contiene 6000 frases (o archivos de audio) que fueron grabadas por 100 hablantes, cada hablante grabó 60 frases, 10 comunes y 50 individuales.

El *Corpus DIMEx100* está organizado en 100 carpetas una por hablante: s001, s002, s003,..., s100. Cada una de las carpetas tiene varias subcarpetas: audio, audio editado, etiquetas default, etiquetas fonemas, fonemas, sílabas, texto. A su vez dentro de cada una de ellas contiene sus respectivas carpetas de comunes y de individuales. (Pineda, Luis. et all. (2004) DIMEx100: a new phonetic and speech corpus for mexican spanish. En memorias del congreso Iberamia 2004, Tonanzintla, Puebla, Mex.)

El *Corpus DIMEx100* se etiqueta en dos niveles de representación de la lengua: un nivel T54 que corresponde a la segmentación alofónica según el inventario del español de México (Cuétara, 2004) más sus respectivos acentos, y un nivel Tp que corresponde a la representación de palabras-ortograficas. Existen otros dos niveles de etiquetación T44 y T22, que se derivan del nivel T54. En el nivel T44 se consideran principios acústicos básicos y se realiza de manera semiautomática derivada de T54, además de manera manual se etiquetan codas silábicas (ver etiquetación T44). El nivel T22 se realiza de manera automática basada en T54 y sólo se consideran a los 22 alófonos del español de México.

De manera general, la etiquetación del audio en el nivel de alófonos consiste en reconocer y delimitar cada sonido de la frase hablada mediante la asisgnación de etiquetas que representan a los alófonos.

ETIQUETACIÓN DE T54

Una vez asignada la carpeta correspondiente por ejemplo s001, s002, ..., s066, etc.

Paso 1.- La herramienta de etiquetación: el SpeechView.

El SpeechView se encuentra dentro del conjunto de herramientas de análisis de habla del CSLU-OGI (Center for Spoken Language Understanding Oregon Graduate Institute of Science and Technology), la cual está instalada en cada computadora disponible para el proyecto DIME bajo el icono:

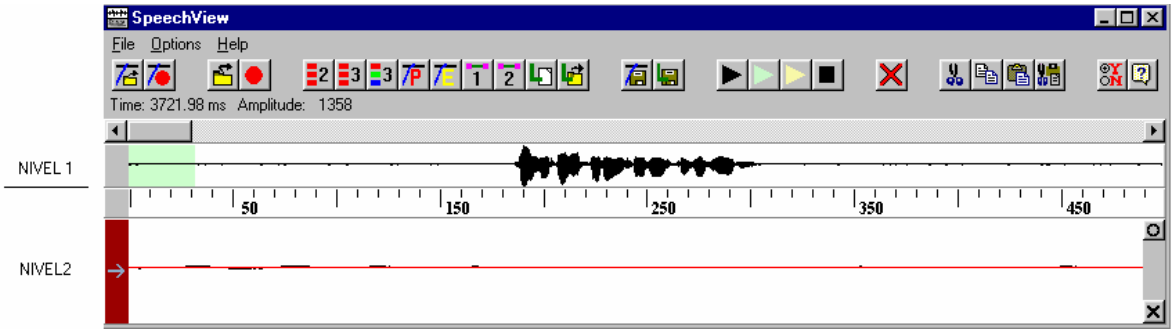
Speech Viewer.lnk

Al darle click, se abrirá la barra de trabajo del SpeechView .

Esta herramienta nos permite abrir el audio, editarlo y realizar la etiquetación del *CorpusDIMEx100*.

1.1 Abrir el audio con el icono siguiente  (el primero en la barra de izquierda a derecha) después ubicarlo en entorno de red/Servidor-dime2/Dimex100/(carpeta asignada)/audio /comunes o individuales (según carpeta que se esté trabajando, se trabajará con ambas carpetas)

1.2 Editar el audio. Como se ve en la imagen siguiente hay demasiado silencio al inicio y final de la frase-audio; por lo que tendremos que recortarlo

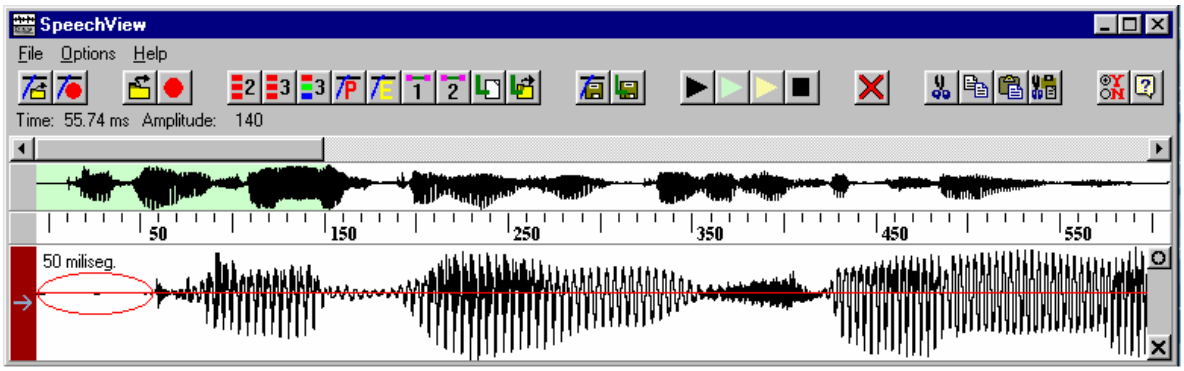


Antes del sonido inicial de la frase se dejará un silencio de 90 o 50 milisegundos según se muestra en la tabla.

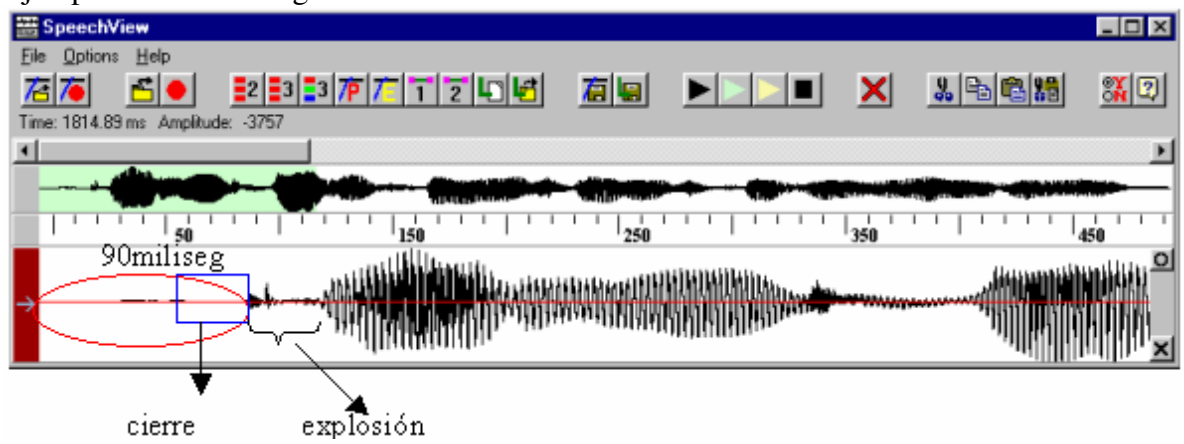
	90 milisegundos	50 milisegundos
Inicia en:	p, t, k, b, d, g, ch, ll,	en todos los demás casos
Finales	no aplica	Todos

Si la frase inicia con sonidos del tipo (p, t, k, b, d, g, ch, ll) se dejara un silencio de 90 milisegundos tiempo que incluye el silencio y la oclusión correspondiente a ese tipo de sonidos, es decir, los sonidos oclusivos presentan un cierre y una explosión.
Si la frase inicia con cualquier otro tipo de sonido el tiempo de silencio será de 50 milisegundos.
El silencio para final de frase en todos los casos es de 50 milisegundos.

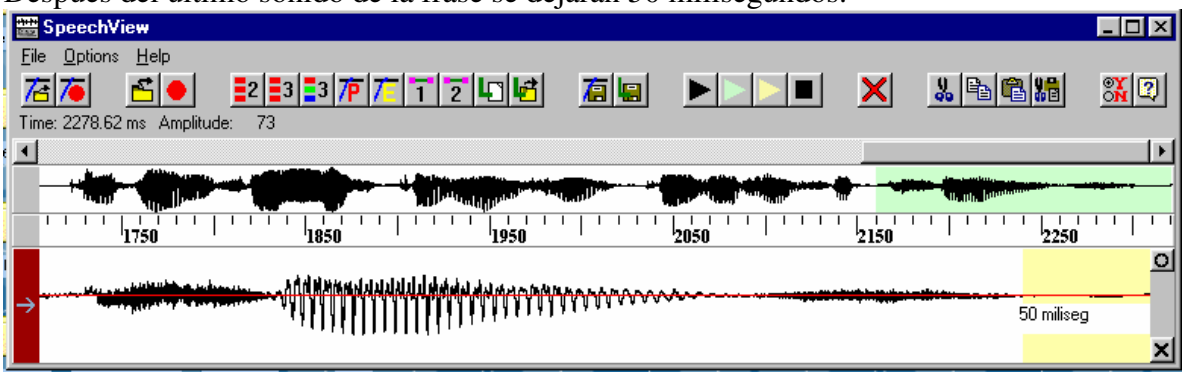
Ejemplo con 50 milisegundos al inicio de la frase.



Ejemplo con 90 milisegundos al inicio de la frase.



Después del último sonido de la frase se dejarán 50 milisegundos.



Para recortar el audio:
1) (nos colocamos en nivel 2 selección amarilla) nos desplazamos hasta el inicio de la frase y una vez identificado el primer sonido y después de haber considerado los 90 o 50 milisegundos se seleccionará la parte restante

2) para cortarlo hacer click en el icono de tijera 
3) para el final de la frase, una vez ubicado el final del último sonido (final de onda sonora) y adicionando los 50 milisegundos de silencio, se cortará lo restante.

1.3 Guardar el archivo de audio editado en la subcarpeta "audio_editado"

Este paso consiste en guardar el audio ya cortado dentro de la subcarpeta correspondiente ("comunes" o "individuales") incluida en la subcarpeta "audio_editado". Para eso hay que dar clic en  y cuidar que el archivo se grave en la dirección apropiada.

Paso 2 Los alófonos del español de México
Los alófonos son los sonidos que forman a un sistema de lengua en su realización (es decir, en su pronunciación)

2.1 Los alófonos del español de México (Cuétara, 2004)
El objetivo de hacer la trascripción en alófonos es obtener la descripción del contenido fonético de las frases del *corpus*, para la creación de modelos acústicos y diccionarios de pronunciación.
Para los sonidos del español de México se propone el siguiente inventario:

Fonemas y alófonos del español de México para el etiquetado fonético

Fonemas	Alófonos	Contexto
Bilabial oclusiva sorda p		
/p/	p_c p	En todos los casos
Dental oclusiva sorda t		
/t/	t_c t	En todos los casos
Velar oclusiva sorda k		
/k/	k_c k_j	_{e, i, j}
/k/	k_c k	En todos los demás casos
Bilabial oclusiva sonora b		
/b/	b_c b	///_
/b/	b_c b	{m, n}_
/b/	V	En todos los demás casos
Dental oclusiva sonora d		
/d/	d_c d	///_
/d/	d_c d	{m, n}_
/d/	D	En todos los demás casos
Velar oclusiva sonora g		
/g/	g_c g	///_

/g/	g_c g	{m, n}_
/g/	G	En todos los demás casos

Palatal africada sorda tS

/tS/	tS_c tS	En todos los casos

Labiodental fricativa f

/f/	f	En todos los casos

Alveolar fricativa sorda s

/s/	z	V_V
/s/	z	_{b, d, g, Z, m, n, n~, l, r, r(}
/s/	s_[	_{t}
/s/	s	En todos los demás casos

Velar fricativa sorda x

/x/	x	En todos los casos

Palatal fricativa sonora Z

/Z/	dZ_c dZ	///_
/Z/	dZ_c dZ	{m, n}_
/Z/	Z	En todos los demás casos

Nasal bilabial m

/m/	m	En todos los casos

Nasal alveolar n

/n/	n_[	_{t, d}
/n/	N	_{k, g}
/n/	n	En todos los demás casos

Nasal palatal n~

/n~/	n~	En todos los casos

Lateral alveolar l

/l/	l	En todos los casos

Vibrante simple r(

/r(/	r(	En todos los casos

Vibrante múltiple r

/r/	r	En todos los casos

Vocal alta palatal i

/i/	j	_{a, e, o, u}
/i/	j	{a, e, o, u}_
/i/	i	En todos los demás casos

Vocal media palatal e

/e/	E	_{r}
/e/	E	{r}_
/e/	E	_{p, t, k, b, g, d, tS, f, x, Z, l, r(}$
/e/	e	En todos los demás casos

Vocal abierta a

/a/	a_2	_{u, x}
/a/	a_2	_{l}$
/a/	a_j	_{tS, n~, Z, j}
/a/	a	En todos los demás casos

Vocal media velar o

/o/	O	_{r}
/o/	O	{r}_
/o/	O	_{consonante}$
/o/	o	En todos los demás casos

Vocal alta velar u

/u/	w	_{a, e, o, i}
/u/	w	{a, e, o, i}_
/u/	u	En todos los demás casos

silencio: .sil **ruido: .bn**

La columna que se emplea para representar los sonidos de los audios editados es la que lleva el nombre de Alófonos.

La columna de la derecha: contexto; nos presenta dos situaciones, la primera es la situación más común, es decir, el sonido se representa *X* en cualquier situación que se presente; la segunda indica que en determinas situaciones la representación de ese sonido *X* cambia; por ejemplo:

El sonido k se representa k_c k (k_c por el cierre que tienen las consonantes oclusivas y k indica la explosión) en cualquier posición en que aparece; pero si este sonido k va antes de una e o una i entonces su representación es k_c k_j, ejemplo la palabra *casa* se transcribe k_c k a s a, la palabra *actuar* a k_c k t w a r(; pero si se tienen palabras como *querer* su representación adecuada es con k_c k_j e r(E r(, la palabra *quitar* se representa k_c k_j i t_c t a r(.

Significados de los contextos

{} antes de, por ejemplo n se trascribe n[antes de t y d. *antes* a n_[t_c t e s

///_ inicio absoluto de palabra; por ejemplo b se transcribe b_c b en inicio absoluto de palabra : *vaso* se transcriben b_c b a s o; *basura* b_c b a s u r(a

{}_ después de, por ejemplo después de n trascribase d_c d.

indagar i n_[d_c d a G ar

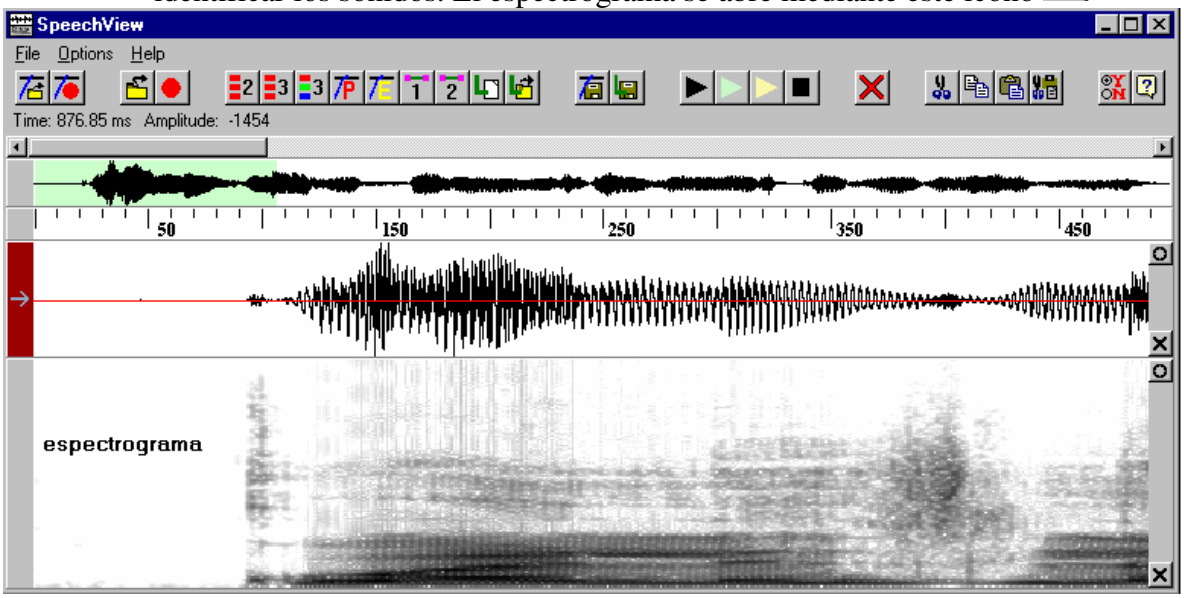
{}$ si a final de silaba hay..., por ejemplo *cuerpo* dividido en sílabas es cuer-po, entonces se transcribe E por que a final de silaba hay r(: k_c k w E r(p_c p o

2.2 Etiquetación T54.El SpeechView y los alófonos del español de México

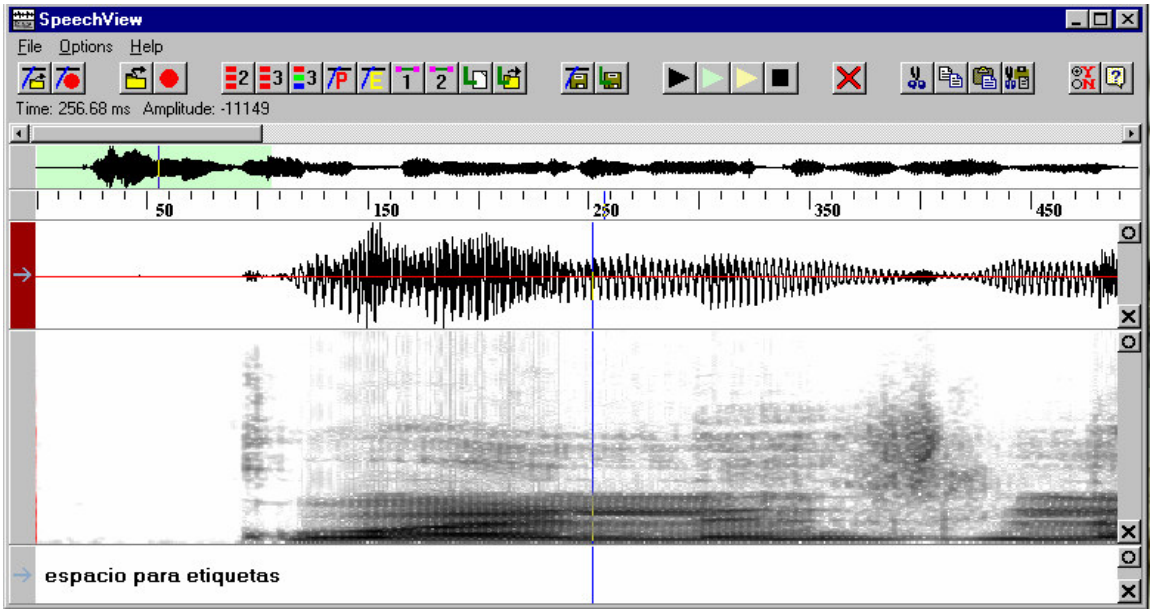
La etiquetación consiste en identificar y especificar cada uno de los sonidos del audio editado mediante la lista de alófonos.

Para ello se necesita:

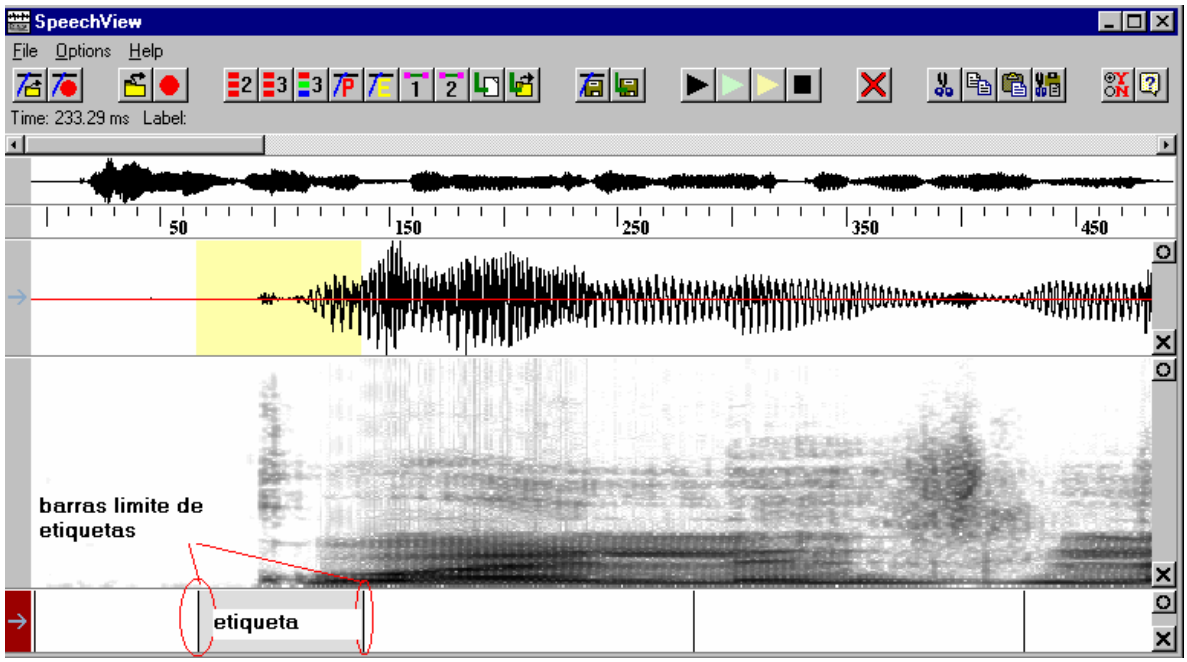
1) el audio editado con el que se está trabajando
2) se necesita abrir el espectrograma, representación gráfica de la onda sonora; ya que de manera visual (a demás de la percepción auditiva) les permitirá identificar los sonidos. El espectrograma se abre mediante este icono 



3) Abrir el espacio de etiquetas mediante el icono  se abrirá una barra blanca debajo del espectrograma,



4) en ese espacio para etiquetas mediante el botón inser o enter se marcaran lineas-temporales que formaran las etiquetas, las cuales podrán ser movidas al colocar el cursor del mouse sobre ellas



5) Su movilidad mediante el mouse les permitirá ajustar la etiqueta al sonido correspondiente. La reproducción del sonido se realiza mediante los iconos  si se desea escuchar el enunciado completo,  y  si se desea escuchar de manera parcial. Una vez identificado el sonido (mediante percepción auditiva y visual) hay que asignarle su representación alófonica de acuerdo con la tabla del paso 2.1 Fonemas y alófonos del español de México. (por supuesto que es a todo el enunciado, en la figura siguiente sólo se presenta de manera parcial)

Ejemplo de alineación- etiquetación:

La siguiente frase que se va a etiquetar dice:

¿cuál es la diferencia de este gobierno?

Su representación en alófonos correspondiente (dependerá de como lo haya dicho el hablante) si es:

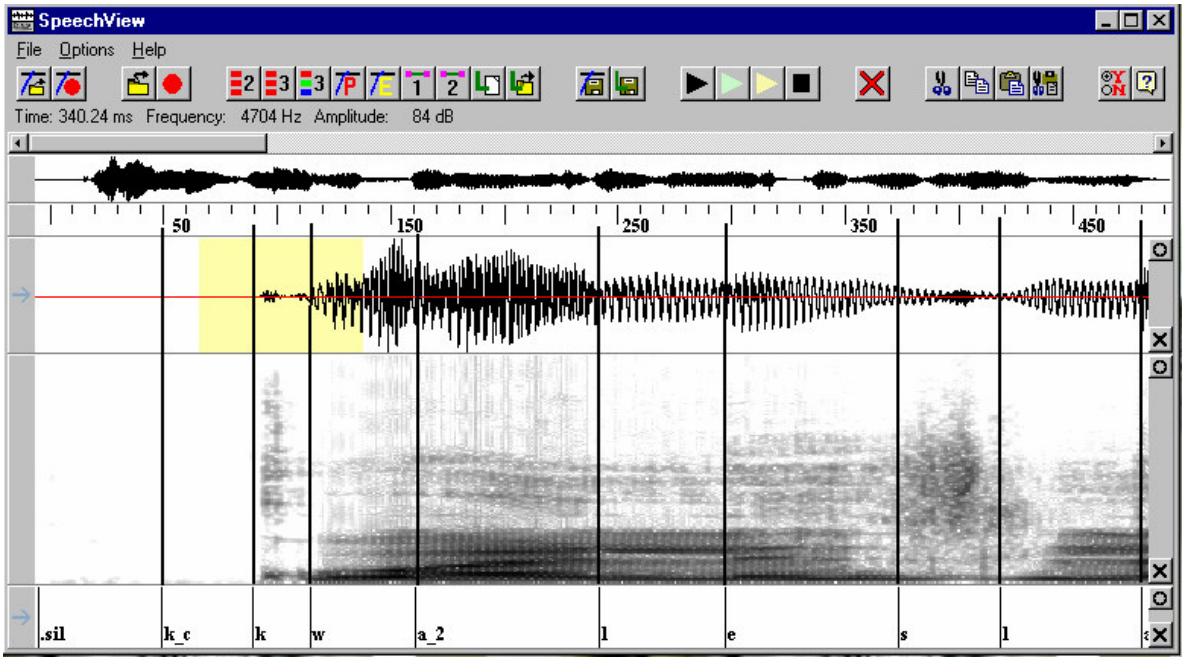
Pronunciando con una sola *e* entre *de este gobierno*

k_c k w a_2 l e s l a D i f e r(n s j a D e s_[t_c t e G o V j E r(n o

Pronunciando con las dos *e* entre *de este gobierno*

k_c k w a_2 l e s l a D i f e r(n s j a D e e s_[t_c t e G o V j E r(n o

Su alineación será de la siguiente manera:



Con respecto a está figura se puede observar en el espectrograma que los limites de etiquetas corresponde aproximadamente a cambios en él y en la representación gráfica de la onda sonora.