PRACTICA 6
(Puntuable)
Implementación de un sintetizador concatenativo de difonos

Consideraciones previas

Para la realización de la siguiente práctica se recomienda en primer lugar que se lea el siguiente documento hasta el final y se haga un esquema de lo que se pretende en dicha práctica:

Enunciado de la práctica

Sea *L* un lenguaje compuesto por seis fonos: la vocal [a], y cinco consonantes: [b], [f], [l], [m], [t], [s]. Estos fonos sólo se pueden agrupar para formar sílabas de los siguientes tipos:

- **V**: [a];
- **CV**: [ba], [fa], [la], [ma], [ta], [sa];
- **CCV**, sólo con [b], [f] o [t] seguido de la consonante líquida [l], es decir: [bla], [fla], o [tla];
- **VC, CVC** y **CCVC**, resultantes de agregar [s] final a los tres tipos anteriores: [as], [fas], [flas], etc.

Además, en el lenguaje L, de ahora en adelante L, hay dos restricciones fonotácticas:

1. El sonido [m] no puede ser inicial en una frase.
2. El sonido [m] no puede suceder a una [s].

Objetivo de la práctica

El trabajo práctico consiste en implementar un sintetizador concatenativo de difonos para L. El sistema debe tener en su inventario exactamente **una instancia de cada difono**; es decir, no debe realizar selección de unidades al sintetizar una frase nueva.

Se pide implementar el sintetizador propiamente dicho; o sea, el back-end de un sistema TTS. El sistema debe recibir como entrada un string con la secuencia de fonos, y generar como salida un archivo de audio conteniendo el habla sintetizada correspondiente a dicha secuencia. En la secuencia de fonos pueden marcarse dos aspectos prosódicos:

- Si una vocal debe acentuarse, se introduce en mayúscula (‘E’); en caso contrario, con minúscula (‘e’).
- La secuencia de entrada puede terminar en el carácter ‘?’, en cuyo caso la salida deberá tener la prosodia de una pregunta (cómo es dicha prosodia es parte del problema a resolver).

La entrada debe representarse como una cadena de caracteres ASCII. Dicha cadena no puede contener espacios en blanco ni caracteres distintos de "aAbflmts?" (usamos el caracter 'm' para representar al fono [m]). Por ejemplo, las siguientes secuencias son entradas válidas: "tAta", "flAs?", "sAasblamAsa", "masAbatlAsa", "EsemeketrEfe?".

En los archivos adjuntos se incluyen dos ejemplos de alumnos de años anteriores, para otro lenguaje similar: "mamAsalAlapApa" y "papAsakAlakAma". Además se adjuntan scripts y otros archivos para praat que os servirán de ayuda (Proporcionado por Agustín Gravano, profesor de la Universidad de Buenos Aires – Argentina)

Trabajo a realizar

Las tareas a realizar consisten en:

1. Diseñar y grabar el inventario de sonidos (difonos), en mono, 16kHz, 16 bits.
2. En Praat, etiquetar los difonos en una capa de intervalos (*interval tier*) en un archivo TextGrid.
3. Recortar los difonos y generar un archivo wav para cada uno.
4. Crear un programa que, dada una secuencia de fonos, concatene los archivos de los difonos correspondientes, genere un archivo wav y, si fuera necesario, modifique su prosodia:
   - El programa debe funcionar en modo batch (no interactivo), recibiendo únicamente dos parámetros. El primero será la secuencia de fonos a sintetizar y el segundo el nombre del archivo wav a crear. Ejemplo:

```
python tts.py EsemeketrEfe? /tmp/output.wav
```

   - La salida debe guardarse como un archivo wav (mono, 16kHz, 16 bits).
   - El programa tendrá además dos opciones:
     - Reproducir automáticamente el audio
     - no reproducir automáticamente el audio.

Sugerencias

- En las grabaciones, hablar normalmente, sin hiperarticular ni sobreenfatizar los acentos.

- No recortar a mano los archivos de cada difono. En cambio, puede emplearse el script de Praat *save_labeled_intervals_to_wav_sound_files.praat* para generar un archivo wav para cada intervalo marcado en un TextGrid. En la opción "Margin (seconds)" usar 0.0001.

- Para concatenar los archivos wav, usar la opción "Combine sounds - Concatenate recoverably" de Praat, que permite ver en un TextGrid los archivos originales. Esto es muy útil para encontrar y rastrear errores en las síntesis realizadas.

- Para el programa del punto 4, usar el lenguaje de scripting de Praat para algunas cosas, y un lenguaje más manejable (por ejemplo, Python) para otras.

- Grabar las vocales acentuadas y no acentuadas como difonos distintos, por ejemplo: _a, _A, as, As, ba, bA, etc.

- No generar la prosodia de pregunta grabando difonos especiales. En este caso, modificar el pitch track del archivo wav generado. Por ejemplo, para ello pueden usarse los scripts provistos en la carpeta manipular-pitch de los archivos adjuntos (leer el archivo README incluido).

Modalidad de entrega

El trabajo se puede realizar individualmente o como mucho en grupos de **tres integrantes**.

- La entrega se realizará por moodle. En caso de realizar el trabajo individualmente se deberá entregar como subject "TP1 apellido1 y apellido2". En el caso de grupo de hasta tres integrantes se debe poner como subject "TP1 apellido1 apellido2 primer integrante, apellido1 y apellido2 segundo integrante, apellido 1 y apellido 2 tercer integrante".
- Además se debe adjuntar un archivo comprimido de nombre "apellido1-apellido2… .zip" con el siguiente contenido: ▪ inventario de sonidos (difonos).
- scripts necesarios para ejecutar el sintetizador, con el código bien comentado.
- archivo README.txt con cualquier aclaración adicional que sea necesaria, incluyendo una breve descripción de la forma en que decidieron modificar la prosodia, y mencionar con qué versión de Praat trabajaron (ej: 6.0.04).

La fecha límite de entrega es aproximadamente el **domingo 1 de diciembre**.

Modo de evaluación

El TP tiene una nota máxima de 10 puntos. Cumplir con los objetivos mínimos (es decir, que funcione y haga lo pedido) otorga 6 puntos. Los restantes 4 puntos corresponden a la calidad del habla sintetizada: 2 puntos por la limpieza de los sonidos y la ausencia de artefactos (clics y otros ruidos) y 2 puntos por la naturalidad en la prosodia generada.

Preguntas frecuentes

Pregunta: No me ha quedado claro si tenemos que grabar aparte los difonos acentuados o vamos a generar los acentos prosódicos artificialmente.
Respuesta: Tienen que grabar los difonos acentuados y los no acentuados por separado.

Pregunta: Para sintetizar una entrada nueva, ¿qué cosas deberían hacerse en Praat y cuáles no?
Respuesta: Una solución posible es que en Python (o similar) procese la secuencia de entrada y construya un script de Praat con los comandos necesarios: abrir los archivos wav de los difonos a sintetizar, seleccionar todos los objetos, concatenar, guardar el resultado. Después el mismo Python ejecuta el script de Praat.

Pregunta: Cuando tengo que repetir un difono, por ejemplo "mamama" donde los difonos ma y am están repetidos, no puedo juntarlos. Yo pensaba que, si los agregaba en orden, o sea:

> *select Sound -m*
> *plus Sound ma*
> *plus Sound am*
> *plus Sound ma*
> *plus Sound am*
> *plus Sound ma*
> *plus Sound a-*
> *Concatenate recoverably*

debería montarse lo que necesito, pero eso me genera solo "mama".
Respuesta: El problema es con la selección de los objetos:

> *select Sound -m*
> *plus Sound ma -->selecciona el primer 'Sound ma'*
> *plus Sound am -->selecciona el primer 'Sound am'*
> *plus Sound ma -->el primer 'Sound ma' ya está seleccionado, no hace nada*
> *plus Sound am -->el primer 'Sound am' ya está seleccionado, no hace nada*
> *plus Sound ma -->el primer 'Sound ma' ya está seleccionado, no hace nada*
> *plus Sound a-*

Para resolver este problema, tenéis que renombrar los sonidos al abrirlos. Por ejemplo, después de abrir el difono "-m", renombralo como "difono1"; después de abrir el primer "ma", renombralo como "difono2", etc. Entonces después, para concatenar, tenéis que hacer "select Sound difono1; plus Sound difono2; plus Sound difono3; ...".

Pregunta: Teníamos una duda sobre si el difono 'AA' era válido o no, ya que no se nos ocurre ninguna palabra en español en la que podamos encapsularlo para hacer las grabaciones.
Respuesta: Sí, el difono AA es válido. Podríamos pedirle al sistema que sintetice "sAAma", por ejemplo. Aclaro que también son difonos válidos "aA", "Aa", "aa". ¿Pero por qué tendrían que ser en español las palabras? L es un lenguaje inventado, y la noción de "palabra" no está definida en el lenguaje L, solo tiene secuencias de sílabas, sin significado ni conexión con el español.

> **Pregunta**: No nos queda claro cómo grabar el difono _b. Por ejemplo, si quiero montar la sílaba [blas] y hago _b+bl+la+as+s_, nos parece que la b se va a

escuchar "dos veces", porque si en el _b ya se escucha un sonido y en la bl también, al juntarlo va a quedar como "bblas", como si fueran dos "ataques" en vez de uno. Lo mismo con el fono [t]. ¿Cómo se puede solucionar esto?
Respuesta: El difono _b tiene que terminar en el silencio correspondiente a la obstrucción del aire en la oclusiva [k]. El difono bl tiene que *empezar* en ese silencio. Entonces, al pegar _b + bl, la primera mitad del fono [b] proviene de _b, y la segunda mitad proviene de bl.

Pregunta: ¿La cadena de entrada puede tener cualquier longitud?
Respuesta: Puede esperarse que la cadena de entrada tendrá una longitud máxima de 30 caracteres, o 31 si termina en "?".