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

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:

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

- **V**: [e];
- **CV**: [fe], [ke], [me], [re], [se], [te];
- **CCV**, sólo con [k], [t] o [f] seguido de la consonante líquida [r], es decir: [fre], [kre], o [tre];
- **VC**, **CVC** y **CCVC**, resultantes de agregar [s] final a los tres tipos anteriores: [es], [fes], [fres], etc.

Además, en L hay dos restricciones fonotácticas: el sonido r no puede ser inicial en una frase, ni tampoco puede suceder a una [s].

Este 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. 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 caracter ‘?’, 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. No puede contener espacios en blanco ni caracteres distintos de “eEfkmrst?” (usamos el caracter ‘r’ para representar al fono [r]). Por ejemplo, las siguientes secuencias son entradas válidas: “tetE”, “trEs?”, “sEestremEse”, “merEcetrEse”, “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)

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 (de ser necesario) modifique su prosodia.
    - El programa debe funcionar en modo batch (no interactivo), recibiendo como únicos argumentos la secuencia de fonos a sintetizar y 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
        i. Reproducir automáticamente el audio
        ii. 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 (ejemplo: _e, _E, es, Es, re, rE, 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 realiza por moodle. En el caso de un integrante debe ponerse como subject "TP1 apellido1 y apellido2". En En el caso de tres integrantes debe poner como subject TP1 apellido1 y apellido2 primer integrante, apellido1 y apellido2 segunto integrante,….
- Además se debe adjuntar un archivo comprimido "apellido1-apellido2…..zip" con:
    - 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 **miércoles 14 de diciembre-clase práctica**.

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 7 puntos. Los restantes 3 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 1 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 "mememe" donde los difonos me y em están repetidos, no puedo juntarlos. Yo pensaba que si los agregaba en orden, o sea:

select Sound -m

plus Sound me
plus Sound em
plus Sound me
plus Sound em
plus Sound me
plus Sound e-
Concatenate recoverably
debería montarse lo que necesito, pero eso me genera solo "meme".

Respuesta: El problema es con la selección de los objetos:

select Sound -m
plus Sound me -->selecciona el primer 'Sound me'
plus Sound em -->selecciona el primer 'Sound em'
plus Sound me -->el primer 'Sound me' ya está seleccionado, no hace nada
plus Sound em -->el primer 'Sound em' ya está seleccionado, no hace nada
plus Sound me -->el primer 'Sound me' ya está seleccionado, no hace nada
plus Sound e-
Para resolver este problema, tenés que renombrar los sonidos al abrirlos. Por ejemplo, después de abrir el difono "-m", renombralo como "difono1"; después de abrir el primer "me", 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 'EE' 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 EE es válido. Podríamos pedirle al sistema que sintetice "sEEke", por ejemplo. Aclaro que también son difonos válidos "eE", "Ee", "ee".¿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: También está la duda si para el difono 'ss' la palabra 'cessette' está bien para ser grabada.
Respuesta: No tienen por qué elegir palabras de español para grabar, el lenguaje L es inventado. Si les resulta más fácil pensar ejemplos parecidos al español, serviría "tres semestres" ("trEssemEstres" en L). Pero para grabar 'ss' yo eligiría una frase portadora más simple, como "messEme", por ejemplo.

Pregunta: No nos queda claro cómo grabar el difono _k. Por ejemplo, si quiero montar la sílaba [kres] y hago _k+kr+re+es+s_, nos parece que la k se va a escuchar "dos veces", porque si en el _k ya se escucha un sonido y en la kr también, al juntarlo va a quedar como "kkres", como si fueran dos "ataques" en vez de uno. Lo mismo con el fono [t]. ¿Cómo se puede solucionar esto?
Respuesta: El difono _k tiene que terminar en el silencio correspondiente a la obstrucción del aire en la oclusiva [k]. El difono kr tiene que *empezar* en ese silencio. Entonces, al pegar _k + kr, la primera mitad del fono [k] proviene de _k, y la segunda mitad proviene de kr.

La frase portadora de _k podría ser _keme, por ejemplo. El difono _k en sí mismo es muy sutil: algo así como una aspiración muy breve (porque es un fono sordo) seguida de un sonido gutural correspondiente al comienzo de la obstrucción del aire en el velo del paladar. Con auriculares debería ser audible, pero insisto que es muy sutil.

Pregunta: Al grabar el difono 'sk', lo pronuncié como [xk] (como en "mosca"), y entonces no se pega bien con otros difonos, como 'es'. ¿Puedo pronunciarlo como [sk], o eso sería hiperarticular?
Respuesta: Lo más conveniente es pronunciarlo como [sk], lo cual no necesariamente lleva a una hiperarticulación, que consiste en exagerar la articulación.

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 "?".