Capitulo III. UNIT SELECTION

3.1 UNIT SELECTION

¿Que es Unit Selection?

Es un método por el cual se pueden concatenar formas sonoras con diferentes estructuras gramaticales, como lo son los fonemas, di fonemas, trifonemas, esto es para poder obtener un mejor resultado en el proceso de sintetizacion, así como en la producción de sonidos naturales.

En este proyecto, tomamos como base el trabajo realizado por Leonardo Flores, dentro del cual considera las características de la prosodia para la implementación de su proyecto, es decir, de identidad fonética y de preferencia para incluir unidades fonéticas de diferentes tamaños.

En nuestro caso utilizamos Unit selection, para lograr una mejor adecuación en cuanto a la selección de la mejor unidad para la construcción de las palabras, esto nos dará una mayor naturalidad en el proceso de sintetización como ya se había mencionado anteriormente.

También contamos con un nuevo corpus, grabado y previamente etiquetado por parte de Martín, otro compañero, el cual desarrollo el corpus para mejora del que se estaba utilizando, este corpus contiene más de 1300 archivos de voz, entre los cuales encontraremos grabaciones tanto de hombre como de mujer.

Estos archivos los encontraremos etiquetados y balanceados en su mayoría, en el cual se encuentran fonemas, frases, palabras, y oraciones, divididos particularmente en archivos (.wrd, .phn, .txt y .wav), esto nos ayudara para tener mayor claridad en lo que nuestro trabajo necesita para obtener mejores resultados en cuanto a claridad y naturalidad de la voz.

La técnica de Selección de Unidad, es la búsqueda a través de un corpus de voz, de las unidades que tengan particularidad con la frase que se vaya a sintetizar [Hunt, 1996]. Debemos tomar en cuenta que para que se realice una buena selección de las unidades, es necesario tomar los costos que se necesiten para evaluar la calidad de la unidad que se va a concatenar para la formación de la palabra.

3.2 Costos de Unit Selection

Para poder tomar una buena función, es necesario conocer y entender el costo de la unidad esto es, $C^t(u_i,t_i)$, un aproximado del resultado de la diferencia de las unidades $\{u_i\}$que están registradas en el corpus $\{t_i\}$con el cual trabajaremos. Las unidades a sintetizar en nuestro proyecto, como lo mencionábamos con anterioridad, serán las palabras y los fonemas, por lo tanto las "secuencias de unidades que componen la frase objetivo, podrá determinar las características, la identidad, y las unidades que le anteceden y en su caso las que son contiguas a esta." [Hunt, 1996].

Los pesos (W), son calculados como la suma de las diferencias de los candidatos que forman parte de la palabra a sintetizar, estas diferencias son representadas por $C_j{}^t(ti,ui)(j=1..., p)$. que es un subcosto de selección.

Entonces podremos decir que la función del costo para la selección adecuada estará determinada por:

$$C^{t}(ti,ui) = \sum_{p} W_j^{t} C_j^{t}(ti,ui) \qquad [1]$$

Para lograr una buena concatenación es necesario tomar en cuenta la siguiente expresión $C^{c}(u_{i-1},u_i)$ la cual es una estimación para la calidad en la unión entre dos unidades que son consecutivas (u_{i-1},u_i) . También la estimación de los subcostos de la concatenación esta determinada por la siguiente expresión q: $C_j^{c}(ui-1,ui)(j=1..., q)$ esto quiere decir que se da por el calculo dado en la caracterización de de *ui-1 y ui* que a su vez es proporcionada por la forma de procesar las unidades.

Las unidades que son consideradas como contiguas, se les asigna un peso de cero, ya que estas unidades se concatenan de manera natural. Los costos de concatenación cuando los pesos W_j^{c} son dados, estarán definidas de la siguiente manera [Hunt, 1996]; [Beutnagel, 1999]:

$$C^{c}(u_i\text{-}1,u_i) = \sum_{j=1} W_j^{c} C_j^{c}(u_i\text{-}1,u_i) \qquad [2]$$

En las fórmulas dadas [1] y [2] se deduce el costo total para definir la secuencia de unidades (n), y que nos dara la suma para los costos de la selección y concatenación, para el caso de los silencios los definimos de esta manera: $C^{c}(S,u_1)$ y $C^{c}(u_n,S)$, esto nos define el inicio y el fin, que son dados para lograr la concatenación de las unidades que contienen silencio en su estructura [Hunt, 1996]:

$$C(t_1^{n},u_1^{n}) = \sum_{i=1}^{n} C^{t}(t_i,u_i) + \sum_{i=2}^{n} C^{c}(u_i\text{-}1,u_i) + C^{c}(S,u_1) + C^{c}(u_n,S) \qquad [3]$$

Esto quiere decir que los pasos para que Unit Selection pueda ser una implementación, tenemos que determinar un conjunto, en donde encontremos las unidades u_1^{-n} que esta dado para el costo dentro de la ecuación [3] se pueda minimizar.

$$u_1^{-n} = \min_{u_1 \ldots u_n} C(t_1^n, u_1^n) \qquad [4]$$

3.3 Los sistemas de texto a voz y Unit Selection

Para los conceptos de los sistemas de texto a voz y de unit selection, se da de manera distinta a lo que es una arquitectura del sistema de TTS. Esto es por que no nos hemos dado a la tarea de tomar en cuenta la entonación que se le tiene que dar a cada fonema, así mismo pasa con su entonación. Dentro de unit selection encontramos di fonemas, trifonemas, pero en nuestro caso trabajaremos con puros fonemas y palabras, para agilizar un poco mas nuestra búsqueda, ya que consideraremos los contextos de cada fonema. Dadas las pronunciaciones y con las entonaciones que contiene nuestro corpus, se podrá evitar el Generador Prosódico, así el procedimiento para acceder al modulo de texto – fonemas será directo al modulo de sinterización, si es que se trabaja con fonemas, en el caso de las palabras o frases, el proceso será directo entre los módulos de procesamiento y el de síntesis, en el caso de que no se encuentre la palabra daremos paso al modulo de texto a fonemas, para la elaboración de la palabra.

En el caso de no pasar por el Generador Prosódico, no quiere decir que la concatenación de los fonemas para la producción de la voz, se dejen de aplicar las técnicas de suavización de voz, tomando en cuenta las fronteras y la regulación de la entonación.

3.4 Concatenación (Unit Selection)

En la actualidad, el trabajo realizado sobre unit selection fue incorporado a un sintetizador llamado CHART de ART (Interpreting Telecomunications Research, Labs.) por Andrew Hunt y Alan Black quienes implementaron la técnica mencionada, ya que contaban con la capacidad para realizar lo que llamaron “realizaciones naturales” [Hunt,1996].

Para la realización de la selección de unidades, se requiere de la definición, y del entrenamiento de los costos de selección y de concatenación. Andrew y Alan mencionan que es necesario que la base de las unidades a sintetizar sea vista como una red de estados de transición, ya que el costo de ocupancia, podría ser mapeado o copiado al costo de selección, esto consiste en un procedimiento en el que los estados almacenan la información que hace referencia a la unidad a sintetizar, ya que esta se encontraba dentro de la base de datos.

En cuanto al costo de transición, dado entre los estados, se da por medio del costo de concatenación, el cual es una estimación de la calidad para unidades que se presentan consecutivamente.

Black y Hunt usaban los fonemas como una unidad de síntesis, esto se daba cuando se quería producir un sonido de voz para la oración, esto es la incorporación del trabajo que se menciona unas líneas arriba, en la incorporación a CHART en donde se da la transformación de la entrada de la palabra objetivo. La cual definía una cadena que contenía los fonemas a sintetizar. La unidad podía relacionarse con cualquiera que existiera en el vocabulario usado.

En el modelo que realizan Balck & Hunt, donde encontramos la red de los fonemas en una base de datos, se representan las unidades que van a ser sintetizadas, dan un ejemplo en el cual la palabra a sintetizar es (synthesize --> /s/ /I/ /n/ /th/ /e/ /s/ /ai/ /z/) “los estados (cajas), representan los fonemas de la base de datos y sus transcripciones (líneas) para todas las posibles secuencias de concatenación”. Ya que obtenia el objetivo, que era representado por la transcripción fonética con un valor asignado de t_1^8 y como cada fonema contiene características de prosodia, se tenia que localizar la ruta para la secuencia de las unidades que minimizarían los costos para unit selection, esto se daba por los cálculos que eran realizados dentro del algoritmo de Veterbi, que se daba por un espacio de búsqueda por pesos o por un entrenamiento regresivo [Hunt, et. al.,1996].

En 1999, se propone realizar unas modificaciones a la técnica propuesta por Hunt & Black, los cuales consisten en un preprocesamiento de los costos de concatenación, Beutnagel [Beutnagel et. al., 1999]. Este trabajo es realizado e incorporado al sistema TTS de AT&T. Es aplicado a los di fonemas que aun sigue ingresando a un vector, cuyas características, son las de involucrar los elementos de la prosodia. Este trabajo hace mas eficaz los TTS de AT&T comparado con el CHART de ART, ya que es el resultado de cómo ha ido creciendo el sistema.

Los trabajadores, junto con Beutnagel, instituyen como característica necesaria, el procesar los costos de concatenación antes de realizar la sinterización, en vez de usar funciones costosas de cálculo así como el algoritmo de Viterbi.
Dentro de la Universidad de Colorado en Boulder, se da una línea de investigación que es aportada por Byan Pellom, miembro del CSLR (Center for Spoken Language Research).

Bryan Pellom, colabora con un dominio que es restringido a lo que es básicamente su aplicación, (reservación de vuelos), en el cual usa diferentes unidades para la concatenación de las unidades, dentro de las palabras o frases, la información es almacenada dentro de una lista ligada.
http://communicator.colorado.edu/examples/example.html

En el caso de que la frase u oración se encuentra dentro del corpus utilizado, se anexa completamente, en el caso contrario, se construirá a partir de los fonemas que hayan sido seleccionados a sintetizar. El trabajo de sus transcripciones, como lo son las alineaciones, es realizado bajo los modelos ocultos de Markov (HMM) [Internet 7].

En nuestro trabajo nos enfocaremos a crear las palabras a partir de los fonemas, ya que no contamos con un vector de características, como lo usaban Black & Beutnagel dentro del costo de selección. Para que este caso se diera tendríamos que conocer al por mayor sus proyectos.

Lo que se realiza en este proyecto es incluir fonemas y palabras, ya que se integran valores para las unidades seleccionadas, además de los contextos fonéticos. Tampoco daremos o crearemos una función para los costos de concatenación, pero si creamos un proceso en el cual se integran unidades que sean del mismo archivo de audio, así mismo les daremos prioridad a los pesos que hayan sido especificados.

Este proyecto es realizado a partir del trabajo anterior, el cual a su vez toma una evaluación de una heurística que se ha sido revisada en el laboratorio de la Universidad (TLATOA).

3.5 Resumen

En este capitulo comprendimos el proceso que realiza Unit Selection para la concatenación de elementos fonológicos. Dentro del estudio que se ha venido haciendo, nos dimos cuenta de que es necesario tomar en cuenta los costos de selección y concatenación para realizar una buena selección de la unidad a concatenar.