Estudio de la mejora de la calidad de voz para un sintetizador en idioma castellano usando el método de Autómatas Adaptativos (5 Enero 2009)

R. Caya y C. Zapata, *Member IEEE*

***Resumen*— El proyecto en desarrollo, presentado en este artículo, realiza un estudio de la mejora de la calidad de voz del motor de síntesis de voz *FESTIVAL* mediante la adición de autómatas adaptativos. Dichos autómatas están diseñados para dar solución a algunos problemas fonológicos identificados para el castellano, y se basan en reglas lingüísticas dependientes del contexto las cuales también son materia de investigación del proyecto. El sistema modificado será sometido a pruebas formales de calidad de voz como MOS y frases psicoacústicas de Harvard adaptadas al castellano, para evaluar su mejora.**

Palabras clave—Síntesis del habla, Interfaces de lenguaje Natural, Autómata Adaptativo.

I. NOMENCLATURA

- AA: Autómata adaptativo.
- TTS: Texto to speech, en español texto a voz.
- Elementos supra-segmentales: estructuras mayores a la palabra léxica propias del discurso oral.
- Inteligibilidad: la facilidad para comprender la señal de voz producida.
- Naturalidad: un indicador de la semejanza de los sonidos producidos artificialmente con los naturales.
- PLN: procesamiento de lenguaje natural
- Coarticulación: fenómeno lingüístico mediante el cual un fonema se ve influenciado por su contexto.
- Alofonemas: variaciones de un fonema en un idioma.
- Utterance: unidad del discurso ubicada entre dos pausas evidentes denotadas por signos de puntuación

II. INTRODUCCIÓN

ACTUALMENTE los avances en el campo de la tecnología del habla nos permiten encontrar sistemas de síntesis de voz muy avanzados en diversos idiomas. Sin embargo éstos aún presentan problemas que impiden el logro de la naturalidad deseada por los usuarios, entendida, de acuerdo con [9], como la similitud lograda entre la voz artificial generada y el discurso humano, respecto a las características de un idioma determinado.

El presente proyecto tiene como objetivo identificar y formular las reglas dependientes del contexto que puedan dar solución a algunos de los problemas encontrados en un sintetizador de voz en particular, y que permitan la mejora en la naturalidad de la voz sintética producida utilizando técnicas adaptativas. Su realización se encuentra justificada por el impacto que significaría para la sociedad actual contar con una mejor calidad de la síntesis de voz.

III. MARCO CONCEPTUAL

La búsqueda de un medio que permita una interacción más natural con el ordenador, tal como menciona [1], constituye una de las principales razones para la investigación en la rama denominada HCI por sus siglas en inglés (Human-Computer interaction). En este contexto, la síntesis de voz parece ser un gran aliado, debido a que los sistemas de síntesis de voz permiten enriquecer la interacción humano-computador desde el punto de vista de la respuesta del ordenador.

Los sistemas de síntesis son comúnmente denominados TTS por sus siglas en inglés (Text-to-speech), sin embargo su información de entrada puede ser tanto texto como una representación lingüística simbólica. Según la flexibilidad de la cadena de entrada que reciben, se pueden clasificar, como indica [11], en: sistemas de respuesta de voz y conversores de texto a voz. A estos últimos corresponde el sistema de síntesis con el que se trabaja en este proyecto, pues presentan mayores problemas en la calidad de voz al permitir trabajar con cualquier texto de entrada.

Todo sintetizador implementa un método de síntesis. En la actualidad los principales métodos utilizados son: síntesis concatenativa y síntesis por formantes. Tal como señala [17], el primer tipo se refiere a la capacidad del sintetizador de producir un discurso fluido, en tiempo real, en base a la consecución de unidades del habla, pregrabadas y etiquetadas en una base de datos. En este tipo de síntesis la selección del tamaño de las unidades a almacenar presenta una complejidad alta, con influencia en la calidad de voz obtenida, como se menciona en [14]. La síntesis por formantes, en cambio, propone la generación de la voz sintética a partir de un modelo acústico de resonadores electrónicos que simulen las características físicas del aparato fonológico humano [17].

Con respecto a los temas de investigación en el área, la mayoría busca integrar al proceso de síntesis el concepto de prosodia, con la finalidad de analizar y representar

Este trabajo ha sido apoyado por la Pontificia Universidad Católica del Perú a través del curso de Tesis que se imparte en la especialidad de Ingeniería Informática de la Facultad de Ciencias e Ingeniería.

C. Zapata imparte docencia en el Departamento de Ingeniería, Sección de Informática de la Pontificia Universidad Católica del Perú, Av. Universitaria 1801, San Miguel, Lima 32, Perú (correo e.: zapata.cmp@pucp.edu.pe).

R. Caya es alumna de pregrado en la Especialidad de Ingeniería Informática de la Facultad de Ciencias e Ingeniería de la Pontificia Universidad Católica del Perú, Av. Universitaria 1801, San Miguel, Lima 32, Perú (correo e.: rosalia.caya@pucp.edu.pe).

formalmente las características de los elementos supra-segmentales en la expresión oral. De acuerdo con [8], es la prosodia la encargada de dotar al discurso de expresividad y actitud que permiten que el receptor comprenda el mensaje transmitido.

IV. Descripción de la solución

En este proyecto se propone el estudio de la mejora de la calidad de un sintetizador de voz para el castellano, utilizando el método de AA, para la representación de relaciones gramaticales dependientes del contexto que afecten la calidad de la voz sintética. Dicha calidad puede ser evaluada desde dos puntos: inteligibilidad y naturalidad. Sin embargo, es importante resaltar que no se ha implementado un nuevo sintetizador de voz, sino que se ha modificado uno existente.

El sintetizador con el cual se trabaja es Festival Speech Synthesis System, una plataforma de software libre dedicada por completo a la síntesis, creada por el Centro de Investigación de Tecnologías del Lenguaje de la Universidad de Edimburgo [4]. Su elección corresponde a que se trata de un software libre, cuenta con documentación oficial, presenta flexibilidad en el trabajo con varios lenguajes de programación y ha sido utilizado anteriormente por otros trabajos de investigación con resultados satisfactorios. Debido a estas razones su uso académico es bastante viable.

El proyecto se viene realizando en dos fases: investigación de las reglas fonológicas del español y desarrollo del aporte producto de la investigación. Los objetivos establecidos para la primera fase son: identificar los factores que impiden la producción de una voz sintética natural y formular las reglas dependientes del contexto usando autómatas adaptativos que solucionen algunos de los factores identificados en el punto anterior. La fase de desarrollo del aporte tiene como objetivos: implementar los autómatas adaptativos en la plataforma de síntesis elegida; diseñar y realizar experimentos con los autómatas adaptativos implementados en casos reales de síntesis de voz para el castellano. Actualmente el proyecto está finalizando la primera fase.

A. Sustento de la solución

La síntesis de voz es realizada en FESTIVAL, por medio de la aplicación de reglas que buscan modelar sus características. Así mismo, varias de estas reglas presentan dependencias del contexto propias del idioma que modelan. Bajo estas consideraciones se eligió trabajar con un AA debido a que su principal objetivo es el análisis del contexto por medio de la formulación de reglas que le permitan adaptarse al mismo. Por lo tanto, los AA encuentran un escenario idóneo en el análisis de las características supra-segmentales del castellano.

Por otra parte, la formación de palabras fonológicas y el análisis de los modelos melódicos de puntuación, constituyen en castellano conceptos que modelan características supra-segmentales que enriquecen la expresión oral, como son: las pausas, la melodía y el tono. Así mismo, el análisis para ambos conceptos es viable dentro del alcance del presente proyecto al encontrarse normado su uso en la gramática española como se demuestra en [12].

Como se menciona, actualmente se emplean diversas técnicas en los sintetizadores de voz y hasta el momento la calidad de la síntesis en términos de naturalidad va asociada con el grado de complejidad del método usado para llevarla a cabo. Algunos de estos métodos son: Modelos ocultos de Markov, redes neuronales, TD-PSOLA, MEL-CEPSTRAL, entre otros. El inconveniente al utilizar estos métodos es que el manejar los problemas dependientes del contexto aumenta rápidamente la complejidad de la solución, restringiendo los proyecto a un ámbito reducido del dominio del problema para poder acceder a una calidad de voz aceptable. En estos casos no solo se tiene que tratar con la complejidad teórica del algoritmo sino que a ello se suman la propia complejidad del problema y la adaptación del algoritmo para solucionarlo. Tal como menciona [11], actualmente los sistemas de síntesis del habla demandan la maximización de la calidad de voz, mientras minimizan su requerimiento de espacio en memoria y la complejidad del algoritmo que manejan.

Un AA, tal como se menciona en [3], provee un modo claro y simple de manejar sintácticamente las características dependientes del contexto. Esta dependencia se maneja mediante modificaciones en la estructura del propio AA y que son llevadas a cabo por reglas que se disparan al detectarse una condición predefinida en el contexto de ejecución del autómata.

Es relevante, en este punto, señalar que las aplicaciones actuales de síntesis de voz, como FESTIVAL, utilizan reglas para el manejo de las condiciones de dependencia del contexto en la síntesis de voz, por lo que el usar AA se presenta como una alternativa muy natural, pero con una ventaja: la capacidad representativa de los AA, esto es, se espera que sea más sencillo de representar e implementar reglas que manejen condiciones de dependencia del contexto usando AA que con los métodos actuales.

B. Resultados esperados

El aporte práctico del proyecto es la implementación de una mejora de la naturalidad en la voz sintética producida por FESTIVAL para el castellano, lograda mediante AA.

El resultado esperado general es: La mejora de la calidad de voz con AA disminuyendo la complejidad de la solución requerida hasta ahora en para la solución de problemas de la síntesis de voz en relación con la naturalidad y la prosodia.

En relación con los objetivos específicos del proyecto los resultados esperados son:

1. La documentación de los factores identificados que impiden la producción de voz sintética natural y su orden de importancia.
2. El diseño de AA que permitan dar solución a algunos de los problemas antes mencionados para el castellano.
3. Los resultados de la evaluación realizada al sintetizador modificado con AA, respecto a la naturalidad resultante.

V. Apoyo teórico

Diversos modelos de autómatas han sido aplicados exitosamente en diversas ramas del PLN para afrontar múltiples problemas. Así la teoría de autómatas provee

herramientas eficientes y convenientes para la representación del fenómeno lingüístico y encuentra en el PLN su campo de mayor aplicación [1]. En los TTS, la teoría de autómatas ya ha sido empleada para las tareas de transformación del texto de entrada en unidades manejables por el sistema.

En consecuencia, tanto la teoría sobre autómatas como las aplicaciones realizadas hasta el momento sostienen que existe un subconjunto de reglas de la gramática del lenguaje natural que pueden ser representadas por medio de autómatas.

Los trabajos realizados hasta el momento respecto a TTS, señalan que una posible mejora en la calidad de voz se lograría al analizar y representar las características contextuales que presenta el texto de entrada del sistema. De este modo, fenómenos naturales como la coarticulación y la selección de alofonemas en el habla puedan ser simulados. Actualmente los TTS más avanzados han realizado mejoras en su calidad adicionando modelos estadísticos altamente complejos para analizar las características prosódicas inmersas a nivel léxico. Los resultados han sido diversos, presentando un mejora significativa en lenguas en las que algunas características prosódicas poseen representación léxica, como en el caso del francés y el castellano. Para el caso del castellano la prosodia es parcialmente representada en grafemas por medio de signos de puntuación, interrogación, y admiración, además del uso de la tilde para denotar la fuerza de voz en alguna sílaba y, en general, convenciones ortográficas y gramaticales que permiten aplicar diversas características al texto que indiquen cuál debe ser su representación oral.

El contar con una representación prosódica parcial en el nivel tipográfico del castellano permite identificar, de modo evidente, la relación entre sintaxis y prosodia. Dicha relación se presenta por medio de construcciones gramaticales, que en el campo lingüístico son reconocidas desde hace ya varios años.

Los autómatas adaptativos son aplicados en la actualidad a diversos campos. Tal como se menciona en [3], la tecnología adaptativa es aplicable, en general, en aquellos entornos en los cuáles se presente dependencia del contexto. La aplicación de dicha tecnología para algunas tareas en los procesos de síntesis de voz no sólo es posible sino que además es recomendada en [18], desde el punto de vista de un análisis del contexto para lograr una mejor representación del discurso humano tal como se postula en [16].

Los autómatas adaptativos presentan además la posibilidad de trabajar desde el nivel sintáctico, con una complejidad de solución mucho menor respecto a las técnicas utilizadas hasta el momento.

En resumen, la teoría de AA nos permite representar algunas características del discurso del castellano por medio de una gramática finita, clara y exacta, en especial cuando las reglas que constituyen dicha gramática presentan una fuerte dependencia del contexto en el que se encuentran.

VI. Métodología aplicada

El proyecto de investigación se encuentra guiado por la metodología propuesta en [4] para proyectos de investigación. La adaptación de las líneas generales descritas en [4] para este proyecto dieron como resultado las siguientes etapas:

1. Identificación de los parámetros presentes en la sintaxis que afectan la naturalidad del discurso.
2. Identificación de casos en los cuales se encuentra presente la dependencia del contexto.
3. Identificación de las reglas que solucionan los casos seleccionados.
4. Construcción de los autómatas adaptativos que implementan dichas reglas.
5. Selección y realización de pruebas que evalúen la calidad del sintetizador.
6. Análisis de resultados.
7. Elaboración de conclusiones.

VII. Diseño e implementación de los Autómatas Adaptativos

Para este proyecto se diseñarán e implementarán dos AA, cada uno con una función específica. El primero de ellos se dedicará a la formación de palabras fonológicas a partir del texto ingresado. El segundo se concentrará en la asignación de pausas y patrones melódicos característicos de los signos de puntuación en el castellano. Cada uno de ellos se detalla a continuación.

A. Autómata Adaptativo para la formación de palabras fonológicas

El autómata adaptativo que se diseña a continuación tiene como función el reconocimiento y la formación de las palabras fonológicas para el texto de entrada del sintetizador. La palabra fonológica es un concepto clave en la prosodia, la acentuación y la entonación dentro de la gramática española. Su inclusión dentro del procesamiento del texto en un sintetizador de voz permitirá elevar el análisis realizado al nivel de frase dentro de la jerarquía fónica.

La formación de la palabra fonológica en el castellano sigue las siguientes reglas [10]:

1. La concatenación se realiza con palabras léxicas completas, por lo tanto una palabra fónica contendrá por lo menos una palabra léxica.

2. Se concatenan las palabras átonas presentes, desde el inicio del texto a analizar, con la primera palabra que contenga la sílaba tónica, junto con la cual constituirán la palabra fónica.

Para el siguiente ejemplo extraído de [10]:

"El tigre se sube en el árbol"

Según la regla 1 y 2 su representación sería:

"Eltigre sesube enelárbol"

Donde la primera palabra fónica está formada por las palabras: El + tigre; la segunda, por: se + sube, y la tercera por: en + el + árbol.

Siguiendo la notación formal descrita en [18], el autómata adaptativo encargado de la construcción de palabras fonológicas, nombrado como M_F, se define formalmente como una 8-tupla:

M =(Q, $\sum$, q_0, F, δ, Q,Γ,π), donde:

Q: {a_0, a_f}

$\sum$: {átona, tónica, ε}

q_0 : {a_0}

F : { a_f }

δ:

- $P_0 : a_0$, átona$\rightarrow a_0$
- $P_1 : a_0, \varepsilon \rightarrow a_f$

Q : {a1, a2, a3,...}
Γ: (-, a, ε, a_f)
(+, a, ε, a')
(+, a', átona, a')
(+, a', tónica, a'')
(+,a'', ε, a_f)
π : ((a_0, átona, a_0)) = {(-, a, ε, a_f), (+, a, ε, a'),(+, a', átona, a'), (+, a', tónica, a''), (+,a'', ε, a_f)}, para todo a, a_f que existen y para cada a' y a'' que no existe en Q.

La cadena inicial del AA en su estado inicial está denotada por ω, y corresponde a la lista de palabras léxicas, enmarcadas por los símbolos separadores indicados para el idioma (en FESTIVAL son el espacio en blanco y los símbolos de puntuación presentes en cada idioma). Dicha lista es extraíble, dentro de la arquitectura de FESTIVAL, desde la relación WORD que posee la unidad del discurso (utterance) construida desde el texto proporcionado por el usuario. Por lo tanto, tenemos:

$$\omega = \text{utt.WORD}$$

Además, en Festival se tiene acceso a las listas de palabras átonas del castellano previamente identificadas y definidas en una lista de nombre "atona", con las cuáles se trabajará en el autómata.

La operación de M_F ocurre de la siguiente forma: se parte de la configuración inicial del autómata descrita en la Fig. 1

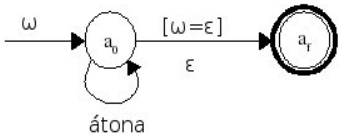

Fig. 1. Configuración Inicial del AA M_F

Luego, al recibir la cadena de entrada dicha configuración cambia a la señalada en la Fig. 2 en la cual se leen n palabras átonas antes de hallar una palabra tónica, y cada vez que se identifica dicha palabra se analiza si ya se acabó de procesar ω. Si fue así, se acaba el proceso. Si por otro lado ocurrió $\omega \neq \varepsilon$, entonces el autómata sustituirá la única transición condicionada en vacío presente por una secuencia de cuatro transiciones: La primera en vacío, la segunda consumiendo x palabras átonas, la tercera consumiendo una palabra tónica y una última transición condicionada hacia el estado final. Esta última transición garantiza la existencia de una transición condicionada en cada paso de la operación del autómata.

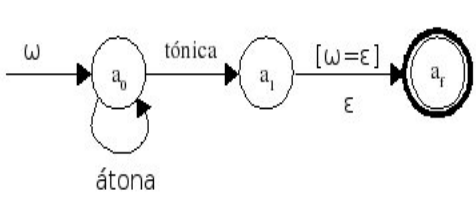

Fig. 2. Configuración del AA M_F luego de realizar la primera transición adaptativa

Luego de i transiciones adaptativas la configuración del autómata será como en la Fig. 3.

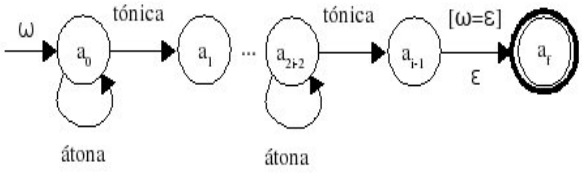

Fig. 3. Configuración del AA M_F luego de i transiciones adaptativas

De esta manera M_F se dedicará así a construir palabras fonológicas que presenten la forma correspondiente con su definición en castellano: [átona ^{n} tónica] ^{k}

B. Autómata Adaptativo para análisis de los signos de puntuación

Para el análisis de las pausas y entonaciones característicos de los signos de puntuación en el castellano se realizarán un conjunto de autómatas adaptativos, cada uno de los cuales analizará el caso particular de la utilización de un signo, y finalmente trabajarán en conjunto de acuerdo a la estructura jerárquica fónica de frases del idioma [6]. Los signos de puntuación que manifiestan pausas y entonación melódica son: "."(punto), ";"(punto y coma), ","(coma), "…"(puntos suspensivos) y ":"(dos puntos); sin embargo también se analizarán aquellos signos que se manifiestan sólo como cambios en el patrón melódico: "¡!"(signos de admiración), "¿?"(signos de interrogación) y "()"(paréntesis). Los autómatas encargados del análisis serán definidos individualmente para luego ser unificados al nivel de la oración.

El mecanismo encargado del análisis de la pausa y entonación características del punto, nombrado M_p, se define formalmente según [18]:
M =(Q, ∑, q_o, F, δ, Q,Γ,π), donde:
Q: {a_0, a_f}
∑: {l, [.], [ ], ε}
q_0 : {a_0}
F : { a_f }
δ:

[6] $P_0 : a_0, l \rightarrow a_0$
[7] $P_1 : a_0, \varepsilon \rightarrow af$
[8] $P_2 : a_0, [.] \rightarrow af$

Q : {a1, a2, a3,...}
Γ: (-, a, ε, a_f)
(+, a, [.], a')
(+, a', [ ], a'')
(+,a'', ε, a''')
(+,a''', l, a''')
(+,a''', [.], a_f)
π : ((a_i, [.], a_f)) = {(-, a, ε, a_f), (+, a, [.], a'), (+, a', [ ], a''), (+,a'', ε, a'''), (+,a''', l, a'''), (+,a''', [.], a_f)}, para todo a, a_f que existen y para cada a', a'', a''' que no existe en Q.

Donde l es cualquier símbolo diferente a un signo de puntuación: letras, números, o caracteres especiales. Así mismo es importante notar en adelante el reconocimiento del espacio en blanco que sigue a un signo de puntuación en el castellano, normado por la Real Academia Española en [12], con lo cual se elimina el problema de los separadores numéricos, como se menciona formalmente en [12] para el acápite de números.

La cadena de entrada de M_p está representada por ω y corresponde a la lista de caracteres ingresados para la síntesis, se analiza a este nivel debido a que Festival trabaja en los demás niveles sin signos de puntuación. Dicha lista es extraíble dentro de la arquitectura de FESTIVAL desde la relación TEXT que posee el utterance:

ω=utt.TEXT

Además se tiene acceso a la lista de signos de putuación del castellano previamente identificados y definidos en una lista de nombre punc, con las cuáles se trabajará en el autómata.

El funcionamiento de Mp es similar al autómata de formación de la palabra fonológica, pero en este caso además de reconocer el caracter punto [.], el segundo estado creado se encarga del reconocimiento del espacio en blanco que sigue al signo de puntuación, y de establecer una pausa normal(#) dentro de la estructura del utterance. Se establece esta pausa debido a que cada punto reconocido dentro del utterance corresponderá obligatoriamente a un punto seguido, a excepción del último el cual será un punto final o punto a parte(los cuales son fonológicamente idénticos), por ello en el estado af se establecerá una pausa larga (##) correspondiente a este tipo de puntuación.

Adicionalmente con cada pausa establecida se asignará el patrón melódico de acuerdo a las normas del castellano, para el caso del punto una inflexión descendente en cadencia absoluta(⇓),

La configuración inicial de Mp y su cambio con cada transición adaptativa se representa en la Fig. 4.

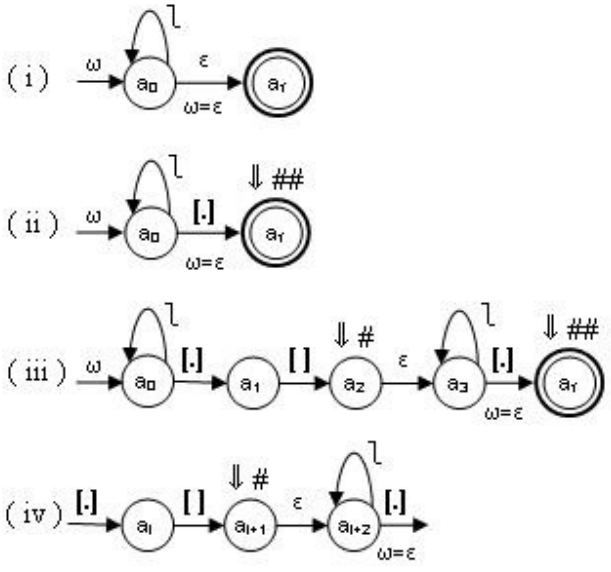

Fig.3. (i) Configuración del AA Mp. (ii) Mp luego de realizar la primera transición adaptativa. (iii) Mp luego de una segundo transición adaptativa. (iv) Cambios en Mp luego de cada transición adaptativa.

Para el caso de la coma, el mecanismo encargado de su análisis, nombrado Mc, se define formalmente:
M =(Q, ∑, q_0, F, δ, Q,Γ,π), donde:
Q: {b_0, b_f}
∑: {l, [,], [ ], ε}
q_0 : {b_0}
F : { b_f }
δ:

- P_0 : b_0, l → b_0
- P_1 : b_0, ε → bf
- P_2 : b_0, [,] → bf

Q : {b1, b2, b3,...}
Γ: (-, b, ε, b_f)
(+, b, [,], b')
(+, b', [ ], b'')
(+,b'', ε, b''')
(+,b''', l, b''')
(+,b''', ε, b_f)
π : ((b_i, [.], b_f)) = {(-, b, ε, b_f), (+, b, [,], b'), (+, b', [ ], b''), (+,b'', ε, b'''), (+,b''', l, b'''), (+,b''', ε, b_f)}, para todo b, b_f que existen y para cada b', b'', b''' que no existe en Q.

Este AA permite el reconocimiento de comas en todas las formas permitidas gramaticalmente: unitaria (como separación entre oraciones), vocativos y enumeraciones de acuerdo con [4].

Tal como en el autómata anterior, Mc reconoce las comas, asigna una pausa mínima(#) y establece la entonación correspondiente, en este caso una inflexión descendente (↓). Sin embargo este autómata no trabajará de modo aislado sino que se creará y obtendrá su entrada a partir del reconocimiento de la coma por M_p en alguno de sus estados de lectura, así su estado final retornará el control al estado correspondiente de M_p La descripción de su funcionamiento se presenta en la Fig. 5.

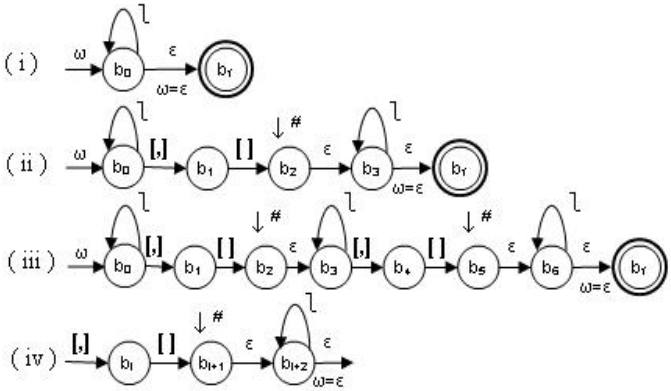

Fig.5. (i) Configuración del AA Mc. (ii) Mc luego de realizar la primera transición adaptativa. (iii) Mc luego de una segundo transición adaptativa. (iv) Cambios en Mc luego de cada transición adaptativa.

En el caso del punto y coma, el mecanismo encargado de su análisis, nombrado M_{pc}, se define formalmente:
M =(Q, ∑, q_0, F, δ, Q,Γ,π), donde:
Q: {c_0, c_f}
∑: {l, [;], [ ], ε}
q_0 : {c_0}
F : { c_f }
δ:

- P_0 : c_0, l → c_0
- P_1 : c_0, ε → cf
- P_2 : c_0, [;] → cf

Q : {c1, c2, c3,...}

Γ: (-, c, ε, c_f)
(+, c, [;], c')
(+, c', [ ], c'')
(+,c'', ε, c''')
(+,c''', l, c''')
(+,c''', ε, c_f)
π : ((c_i, [;], c_f)) = {(-, c, ε, c_f), (+, c, [;], c'), (+, c', [ ], c''), (+,c'', ε, c'''), (+,c''', l, c'''), (+,c''', ε, c_f)}, para todo c, c_f que existen y para cada c', c'', c''' que no existe en Q.

Según [10], el punto y coma indica una pausa superior a la marcada por la coma e inferior a la señalada por el punto, para el caso de este proyecto representada por (#). Así mismo, el punto y coma es usado para separar, como el punto enunciados completos y siempre representa una inflexión melódica descendente (↓), sin llegar a la cadencia con la cual se representa al punto.

Al igual que en el caso de la coma, M_{pc} trabajará en coordinación con M_p, conservando el análisis en el nivel de oración.

La representación del funcionamiento de M_{pc} se prensenta en la Fig.6.

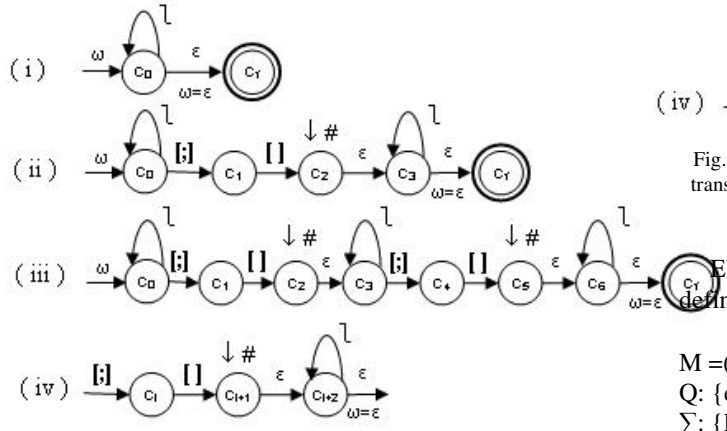

Fig. 6. (i) Configuración del AA Mpc. (ii) Mpc luego de realizar la primera transición adaptativa. (iii) Mpc luego de una segundo transición adaptativa. (iv) Cambios en Mpc luego de cada transición adaptativa.

Para el caso de los dos puntos, el mecanismo encargado de su análisis, nombrado M_{dp}, se define formalmente:

M =(Q, ∑, q_0, F, δ, Q,Γ,π), donde:
Q: {d_0, d_f}
∑: {l, [:], [ ], ε}
q_0 : {d_0}
F : { d_f }
δ:

- P_0 : d_0, l → d_0
- P_1 : d_0, ε → df
- P_2 : d_0, [:] → df

Q : {d1, d2, d3,...}
Γ: (-, d, ε, d_f)
(+, d, [:], d')
(+, d', [ ], d'')
(+, d'', ε, d''')
(+, d''', l, d''')
(+, d''', ε, d_f)
π : ((d_i, [:], d_f)) = {(-, d, ε, d_f), (+, d, [:], d'), (+, d', [ ], d''), (+, d'', ε, d'''), (+,d''', l, d'''), (+,d''', ε, d_f)}, para todo d, d_f que existen y para cada d', d'', d''' que no existe en Q.

Los dos puntos detienen el discurso para llamar la atención sobre lo que sigue, debido a ello en [10] y [4] se los representa con una inflexión descendente (↓), acompañada de una pausa pequeña ($^{\#\#}$), que puede ser similar a la correspondiente a una coma.

Al igual que en los casos anteriores éste autómata trabaja en coordinación con M_p, conservando el análisis en el nivel de oración.

En la Fig. 7 se presenta su configuración inicial y estados adaptativos.

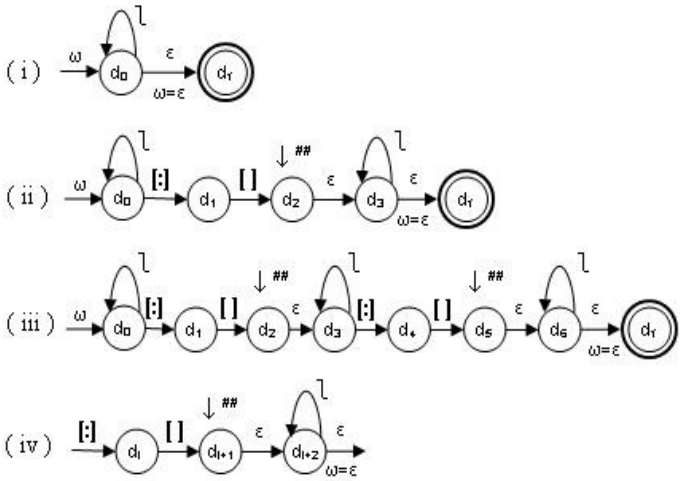

Fig. 7. (i) Configuración del AA Mpc. (ii) Mpc luego de realizar la primera transición adaptativa. (iii) Mpc luego de una segundo transición adaptativa. (iv) Cambios en Mpc luego de cada transición adaptativa.

El caso de los puntos suspensivos se nombra M_{ps} y se define formalmente como sigue:

M =(Q, ∑, q_0, F, δ, Q,Γ,π), donde:
Q: {e_0, e_f}
∑: {l, […], [ ], ε}
q_0 : {e_0}
F : { e_f }
δ:

- P_0 : e_0, l → e_0
- P_1 : e_0, ε → ef
- P_2 : e_0, [:] → ef

Q : {e1, e2, e3,...}
Γ: (-, e, ε, e_f)
(+, e, [:], e')
(+, e', [ ], e'')
(+, e'', ε, e''')
(+, e''', l, e''')
(+, e''', ε, e_f)
π : ((e_i, […], e_f)) = {(-, e, ε, e_f), (+, e, […], e'), (+, e', [ ], e''), (+, e'', ε, e'''), (+,e''', l, e'''), (+,e''', ε, e_f)}, para todo e, e_f que existen y para cada e', e'', e''' que no existe en Q.

Los puntos suspensivos tienen usos bastante estandarizados y se representan siempre como una inflexión ascendente en anticadencia (⇑), típica del final de una frase. Así mismo, suponen una interrupción de la oración o un final impreciso, lo cual se manifiesta a través de una pausa normal (#), según lo indica Navarro en [10].

Una representación de su funcionamiento se muestra en la Fig. 8.

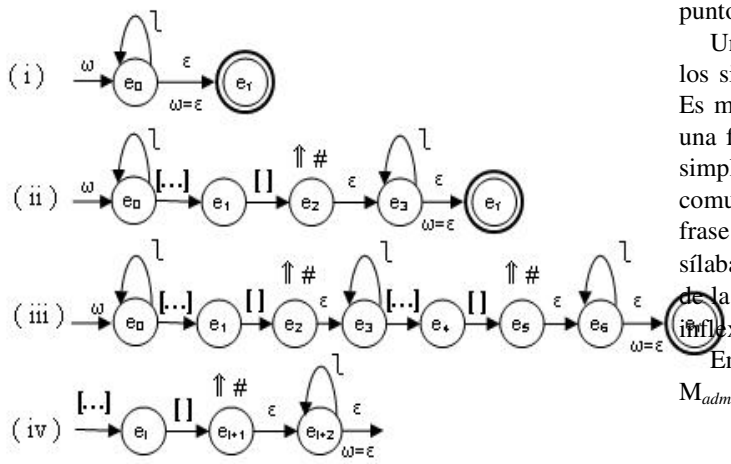

Fig. 8. (i) Configuración del AA Mps. (ii) Mps luego de realizar la primera transición adaptativa. (iii) Mps luego de una segundo transición adaptativa. (iv) Cambios en Mps luego de cada transición adaptativa.

En los casos de signos compuestos por apertura y clausura, tales como los signos de interrogación(¿ ?), admiración(!¡) y paréntesis(), se realizará un autómata por cada uno para garantizar las reglas de apertura y cierre características de los mismos.

A diferencia de los autómatas anteriores, el análisis de signos de interrogación, admiración y paréntesis debe tener en cuenta que estos signos encierran oraciones o frases con sentido completo y pueden contener otros signos de puntuación. Por lo tanto es necesario incluir un análisis a nivel de oración dentro de estos autómatas, agregando las características melódicas correspondientes de cada caso.

Para los signos de interrogación, el autómata M_{int} inicia reconociendo el signo de apertura (¿) y condicionando su llegada a un estado final con la lectura del signo de clausura (?). Luego un "estado" intermedio se encargará del análisis de la oración contenida entre los signos, este "estado" es en realidad una instancia del AA unificado de los signos de puntuación mostrados hasta ahora, nombrado como M_{uni}.

Con el reconocimiento del signo de apertura se asigna el patrón melódico correspondiente. Así las preguntas se manifiestan con un arranque en anticadencia (⇑) en la primera sílaba tónica de la frase y finalizan con una inflexión circunfleja (↑↓) desde la última sílaba tónica, según lo descrito por Alcoba en [4] y que podemos apreciar en la Fig. 9.

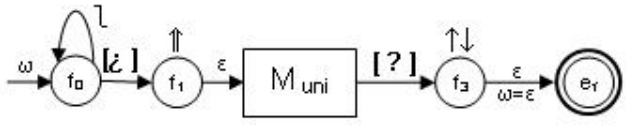

Fig. 9. Configuración general de.M *int*

M_{int} también trabaja en cooperación con M_p, asi se puede crear una jerarquía de oraciones, en la cual una instancia de M_p se encarga del análisis de la oración principal, y M_{int} puede analizar un segundo nivel de oración mediante otra instancia de M_p, la diferencia entre ambas instancias será que la responsable de la oración principal finaliza siempre con un punto (.) y la encapsulada en M_{int} no.

Un trabajo muy similar es el que se realiza con M_{adm}, para los signos de admiración, con la diferencia en la entonación. Es muy difícil tratar de sistematizar las distintas melodías de una frase contenida con signos de admiración, por ello se ha simplificado la tarea y se seleccionando aquellos rasgos comunes entre las diversas melodías. Bajo esta premisa, una frase admirativa se manifiesta con un ascenso hacia la primera sílaba tónica (⇑), un cuerpo tonal correspondiente al contenido de la frase, hasta la última sílaba tónica donde se presenta una inflexión de cadencia (⇓).

En la Fig. 10 se puede apreciar la configuración general de M_{adm}.

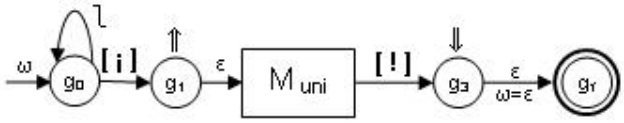

Fig. 10. Configuración general de.M *adm*.

Para el caso de los paréntesis en M_{par}, según [4] debido a que son signos que encierran elementos incidentales o aclaratorios intercalados en un enunciado, ambos se representan con una inflexión descendente (↓), tal como se puede ver en la Fig.11.

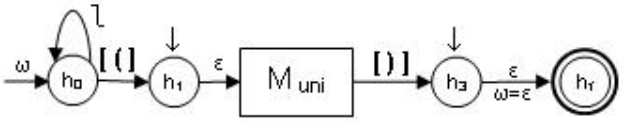

Fig. 11. Configuración general de.M *par*

El mecanismo unificado de puntuación, M_{uni}, está conformado por todos los AA mencionados anteriormente, y organizados de modo que cada uno de ellos pueda desarrollarse de manera independiente, pero manteniendo un trabajo colaborativo. En la Fig. 12 se presenta la configuración inicial de M_{uni}, la cual se limita a la lectura de caracteres del texto, y aunque una frase no contenga ningún signo de puntuación podrá ser analizada.

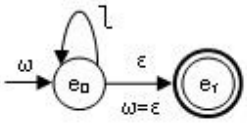

Fig. 12. Configuración inicial de.M *uni*.

La primera transición adaptativa creará la instancia del autómata correspondiente al signo de puntuación reconocido, y pasará el control sobre la cadena de entrada al mismo, el autómata respectivo asignará la pausa y patrón entonativos adecuado, luego seguirá consumiendo signos de ω hasta que encuentre un nuevo signo de puntuación, si este no le

corresponde entonces devolverá el signo a la cadena y el control regresará al estado correspondiente del autómata que realizó la transición adaptativa inicial.

La Fig. 13 muestra la configuración del autómata M_{uni} para el caso de una cadena de entrada que contenga todos los signos de puntuación analizados.

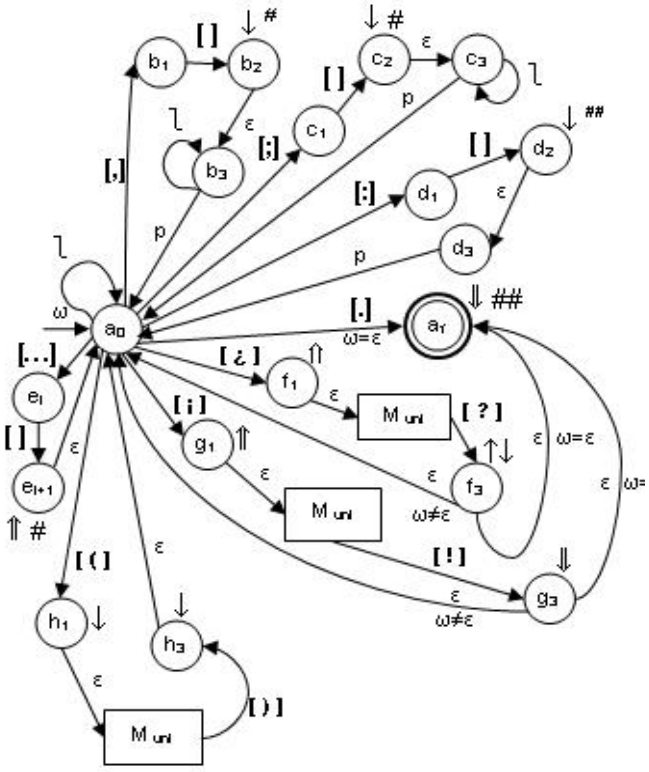

Fig. 13. Configuración de.M *uni* para un ω con todos los signos de puntuación analizados.

Las formas en las que puede ampliarse M_{uni} son variadas, por ejemplo puede crecer al interior de M_c, M_{pc}, M_{dp},o M_{ps}, tal como crecería cada uno de ellos de manera individual. Sin embargo también puede desarrollarse al nivel de la oración, en M_p, formando la estructura de análisis de un párrafo. La Fig. 14 nuestra esta última variante.

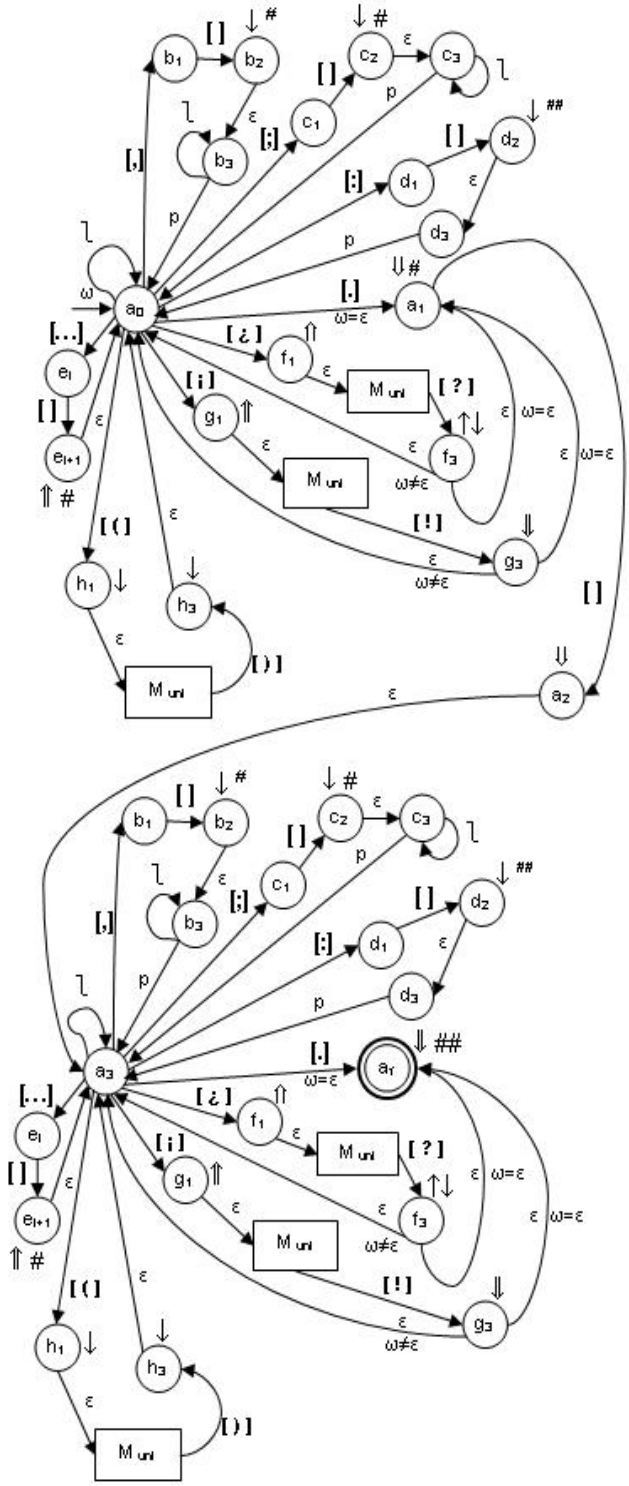

Fig. 14. Configuración de.M *uni* con crecimiento en M*p*.

VIII. IMPLEMENTACIÓN

La implementación de M_F que se presenta a continuación se realizó utilizando el entorno de desarrollo DrScheme, de uso libre para el lenguaje de programación Scheme.

```
;; Concatena los elementos de Word y los devuelve en un string
   ( define (concat_elem Word)
     (   if(zero? (length Word))
         ;;si es tónica
         ""
         ;; si es átona
         (string-append (car Word) (concat_elem (cdr Word)))
      )
   )

;; Forma las palabras fonológicas correspondientes de acuerdo
con  la lista  Word y las devuelve en forma de lista en pal_fon

   (define formar_fonologica
    ( lambda (Word pal_atona pal_fon)
      ( if(zero? (length Word))
          pal_fon
          ( if (not (member (car Word) atonas))
            (formar_fonologica (cdr Word) (list )
           ((append pal_fon (list ( string-append (concat_elem
                                          pal_atona)(car Word)))
              )
          )
        (formar_fonologica (cdr Word) (append pal_atona (list
                          (car Word))) pal_fon)
         )
       )
    )
  )
```

La implementación de M_F se encuentra insertada en el archivo spanlex.scm de FESTIVAL ubicado por defecto en la carpeta correspondiente a la voz en español

IX. Evaluación de la Calidad

Las siguientes secciones exponen la factibilidad de esta evaluación y el método elegido.

A. *Factibilidad de la evaluación de la calidad*

La calidad de un sintetizador del habla es un término multidimensional y puede ser evaluada desde tres perspectivas diferentes: su inteligibilidad, su naturalidad, y su adaptación para una aplicación (Klatt 1987). Cada una de estas perspectivas cobra diversos grados de importancia en determinados contextos, dependiendo del tipo de información que se necesita. La calidad es usada en comparaciones de elementos similares cuando no existen características cuantitativas disponibles, o estas son muy difíciles de utilizar o no deseadas.

El término "auditoría de la calidad" se refiere a la percepción humana del sonido, y sugiere que la comparación cualitativa debe ser realizada por oyentes quienes juzgarán de acuerdo con sus propias preferencias respecto a lo que mejor calidad significa. Sin embargo, algunas conclusiones pueden ser señaladas acerca de lo que es aceptado o no de manera general. Un dispositivo de medida puede ser utilizado en algunos casos, si se conoce la correspondencia entre la cantidad medida y la calidad de percepción respectiva.

La evaluación de la calidad es una tarea subjetiva, en la cual los jueces comparan lo que oyen contra lo que aprecian como perfecto. Las comparaciones entre las características de diferentes sistemas se pueden hacer, y el sistema con las características que están más cercanas a lo que la intuición del individuo califica como perfecta, tendrá la mejor calidad.

B. *Métodos de evaluación a utilizar en este trabajo*

Los métodos para la evaluación de la calidad de voz de un sintetizador, utilizados en la actualidad, se examinaron en el contexto de las necesidades del presente proyecto: la evaluación de la mejora de la calidad de voz respecto a la naturalidad, particularmente en aspectos de supra-segmentación y de puntuación.

Así mismo, se tomó en cuenta la viabilidad de aplicar las diversas pruebas respecto a la disponibilidad de los oyentes, el trabajo necesario para la preparación de las pruebas y la existencia de adaptaciones de cada prueba para el idioma castellano y la posibilidad de su obtención.

El método seleccionado fue MOS, que permite calificar la voz en sus diversas características al asignarle un valor en una escala [17]. Para el desarrollo de las pruebas se utilizarán algunas frases provenientes de diálogos cotidianos y otras, de la adaptación de las frases psicoacústicas de la Universidad de Harvard, adaptadas para el castellano, mencionadas en [14]. Debido a que se tiene especial interés en probar el desempeño de los modelos melódicos seguidos para puntuación, se efectuarán pruebas enfocadas hacia este punto por medio de la lectura de textos especialmente preparados por asesores lingüísticos.

X. Trabajos Futuros

El proyecto continuará con la implementación de los autómatas adaptativos necesarios para el tratamiento del análisis y representación de los patrones melódicos para los signos de puntuación, caso en el cual también se presenta dependencia del contexto y que podría enriquecer la calidad de la síntesis de voz.

El proyecto evaluará formalmente la calidad del sintetizador obtenido. Se realizará la comparación de la "naturalidad" entre la versión original del motor de síntesis FESTIVAL y la versión optimizada, por medio de los autómatas adaptativos implementados.

Para cada una de las pruebas, se decidió utilizar oyentes ingenuos o no preparados, con lo cual su primer contacto con la voz sintética permitirá recoger de modo más real su percepción sobre la misma. La referencia que se presentará a cada uno de los oyentes será la voz de un hispanohablante masculino nativo. Todos los participantes de las pruebas serán miembros de la comunidad universitaria de la Pontificia Universidad Católica del Perú e hispanohablantes nativos. Las pruebas se efectuarán en un ambiente cerrado, silencioso, de dimensiones medianas, y la emisión de la voz se realizará por medio de altoparlantes. Cada una de las pruebas a realizarse será guiada por una persona administradora del sistema.

XI. Conclusiones

Como resultado del trabajo realizado hasta hoy, se pueden

obtener las siguientes conclusiones:

1) La investigación en temas de síntesis de voz no es más una actividad electiva en el ámbito académico como lo eran años atrás. Dado que hoy en día su aplicación impacta el ámbito comercial alrededor de todo el mundo es necesario emplear recursos en la investigación de métodos que permitan solucionar los problemas inherentes a ella.

2) Los trabajos de síntesis para el idioma castellano, son pocos comparados con otros idiomas, como el inglés, y su naturalidad aún necesita ser mejorada en diversos aspectos.

3) La técnica de autómatas adaptativos presenta claridad, facilidad y comodidad para modelar los aspectos de la síntesis de voz dependientes del contexto en el idioma castellano tratados en el proyecto.

4) El lenguaje natural es definitivamente un tópico complejo en el cual existe mucha información a nivel lingüístico, pero menor a nivel informático. Se necesita de investigadores que analicen dicha información y la trasladen a especificaciones computables.

AGRADECIMIENTOS

Los autores reconocen las contribuciones de la profesora Beatriz Mauchi y al profesor Jorge Pérez del Departamento de Humanidades de la Pontificia Universidad Católica del Perú, por su aporte en el aspecto lingüístico del proyecto. Asimismo, al Doctor Joao José Neto de la Universidad Politécnica de Sao Paulo (Brasil), y al Doctor Joaquim Llisterri de la Universidad Autónoma de Barcelona (España) por su colaboración en el tema de métodos de evaluación de TTS.

REFERENCIAS

Periodicals (Artículos de revista):

[1] M. Mohri,, "On Some Applications of Finite-State Automata Theory to Natural Language Processing" *Journal of Natural Language Engineering*, vol. 2, n°1, p.p. 1-20, 1996. [Online] Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.36.5122

[2] B. A. Myers, "A Brief History of Human Computer Interaction Technology," *ACM interactions*, vol. 5, n°2, p.p. 44-54, March,1998. [Online].
Available:http://www.cs.cmu.edu/~amulet/papers/uihistory.tr.html

[3] J.J. Neto, "Adaptive Automata for Context-Sensitive Languages", *ACM Sigplan Notices*, vol. 29, n° 9, pp. 115-124, 1994.

Books (Libros):

[4] Alcoba, S., La expresión Oral, 1ra ed., Barcelona:Ariel, 2000.

[5] A. Black, P. Taylor y R. Caley, Festival Speech Synthesis System, 1.4 ed., University of Edinburh, 2002. [Online] Available: http://www.cstr.ed.ac.uk/projects/festival/

[6] J. Cantero, M. Marti and J. Llisterri, Teoría y análisis de la entonación, Barcelona: Edicions Universitat Barcelona, 2002.

[7] E. Da Silva and E. Muszkat, Metodologia da Pesquisa e Elaboração de Dissertação, 3th ed., UFSC - Universidade Federal de Santa Catarina, Florianópolis, Santa Catarina, 2001.

[8] X. Huang, A. Acero y H. W. Hon, Spoken Language Processing: A guide to Theory, Algorithm, and System Development, 1rst ed., New Jersey: Prentice-Hall, 2001, p.p. 689-953.

[9] D. Jurafsky and J. H. Martin, Speech and language processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, 2nd ed. New York: Prentice-Hall, 2000, p.p. 30-50.

[10]T. Navarro, Manual de la entonación Española, 2nd ed., New York, Hispanic Institute in the United States, 1954, pp. 1-60.

[11]D. O'Shaughnessy, Speech Communications: Human and Machine, IEEE Press Marketing, 2nd ed., 2000.

[12]Real Academia Española, Ortografía de la Lengua Española, Madrid: Espasa, 1999.

[13]Real Academia Española y Asociación de Academias de la Lengua Española, Diccionario Panhispánico de dudas, Madrid: Santillana, 2005.

Published Papers from Conference Proceedings (Artículos presentados en conferencias publicados):

[14]L. Aguilar, J. M. Fernández, J. M. Garrido, J. Llisterri, A. Macarrón, L. Monzón and M. A. Rodríguez, "Diseño de pruebas para la evaluación de habla sintetizada en español y su aplicación a un sistema de conversión de texto a habla" Actas del X Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural, Córdova,España, July 1994. [Online]. Available: http://liceu.uab.cat/~joaquim/publicacions/Aguilar_et_al_94_Evaluacion_Texto_Habla_Espanol.pdf

[15]J. Kominek, C. L. Bennett and A. W. Black,, "Evaluating and Correcting Phoneme Segmentation for Unit Selection Synthesis," in Proc. of the 8th European Conference on Speech Communication and Technology, pp. 313-316, Ginebra,Suiza, Sep.,2003.

[16]R. Sproat, M. Ostendorf, and A. Hunt, "The Need for Increased Speech Synthesis Research", in Report of the 1998 NSF Workshop for Discussing Research Priorities and Evaluation Strategies in Speech Synthesis , March, 1999.

Dissertations (Tesis doctorales):

[17]S. Lemmety, "Review of Speech Synthesis Technology", Master's thesis, supervised by M. Karjalainen, Helsinki University of Technology, University, Helsinki, March 1999.

[18]H. Pistori, "Tecnologia adaptativa en Engenharia de computação: estado da arte e aplicações", Dissertação de doutorado, orientada por J. J. Neto, Departamento de Engenharia de Comparação e Sistemas Digitais, Escola Politecnica da Universidad de Sao Paulo, 2003.

Rosalia Edith Caya Carhuanina nació en Lima, Perú, el 27 de febrero de 1986. Se graduó en el colegio privado Beata Ana María Javouhey y estudia en la Pontificia Universidad Católica del Perú.

Entre sus campos de interés están el procesamiento de lenguaje natural y las interfaces humano-computador.

Claudia Zapata Del Río (M'2007) nació en Lima, Perú, el 03 de octubre de 1978. Se graduó en el colegio Cristo Rey, se recibió de Bachiller en Ciencias con mención en Ingeniería Informática de la Pontificia Universidad Católica del Perú y concluyó en la misma los estudios de Maestría en Ciencias de Computación.

Obtuvo el título profesional de Ingeniero en Informática mediante trabajo de tesis y es miembro del Colegio de Ingenieros del Perú.

Ejerció profesionalmente en Synopsis S.A. y actualmente es profesor auxiliar de la Pontificia Universidad Católica del Perú