PONTIFICIA **UNIVERSIDAD CATÓLICA** DEL PERÚ

PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

FACULTAD DE CIENCIAS E INGENIERÍA

MODELO DE ENTONACIÓN PARA UN SINTETIZADOR DE VOZ CON APLICACIÓN A UN SISTEMA DE INFORMACIÓN VÍA TELEFÓNICA EN CINES

Tesis para optar el Título de Ingeniero Electrónico

Presentado por:

CRISAIDA MARLIT FLORES ALVAREZ

Lima - Perú

2005

ÍNDICE

RESUMEN

Los sistemas de diálogo vía telefónica son desarrollos en los cuales la computadora emula el comportamiento humano para dar respuesta a la consulta del usuario. En este contexto, la entonación con la cual la computadora emita la respuesta, tiene un papel importante, pues es el factor que dotará de "naturalidad" al sistema. En tanto esté bien modelada, la entonación permitirá que para el usuario sea casi imperceptible la diferencia entre una voz de persona y una voz sintética.

Los medios de acceso para conocer información de horarios y tarifas de cine es limitado, pues la tecnología actual en nuestro país sólo permite que el flujo de información sea a través de un medio escrito, Internet o vía una llamada telefónica. Esta última opción es la preferida por los usuarios pues la consideran una forma ágil de transmisión de información, sin embargo está limitado a un horario de atención.

Por ello, dada la creciente demanda de los servicios de esparcimiento y diversión en nuestro país, en particular: lo actividad de ir al cine; y con ella la necesidad de los usuarios de acceder a la información de una forma ágil vía telefónica, entonces, la implementación de un sistema de diálogo usando síntesis de voz que utilice un modelo de entonación, permitirá obtener una alta calidad de voz sintetizada, así los defectos que la atención humana introduce se verán reducidos.

La investigación está desarrollada en cuatro capítulos. En el primer capítulo se analiza la problemática que envuelve el servicio de atención al cliente vía telefónica en los centros de cine. En el segundo capítulo, se muestra el estado actual de la tecnología sobre los sistemas de diálogo, las diversas técnicas para modelar la entonación y se definen los conceptos base que nos permitirán modelarla. Se elige el modelo de entonación en base a corpus, por ser el más adecuado para el sintetizador de voz que usa concatenación por selección de unidades. El tercer capítulo, muestra todo el procedimiento y metodología seguida en la implementación del modelo de entonación elegido: obtención y grabación de corpus, y segmentación de frases. En el cuarto capítulo se hace la evaluación del modelo de entonación implementado, para ello, se definen criterios de evaluación que permiten cuantificar la exactitud del modelo.

Como conclusión principal se obtiene que con el modelo de entonación en base a corpus implementado se alcanza una exactitud del 75%. Con esto, podemos concluir también que la metodología desarrollada para obtener el modelo de entonación en base a corpus representa una buena opción para este objetivo.

A mis padres:

por enseñarme a dar siempre lo mejor de mí.

INTRODUCCIÓN

De acuerdo a los estudios realizados por el INEI (Instituto Nacional de Estadística e Informática), la demanda por los servicios relacionados con la cultura y entretenimiento en nuestro país es creciente. Una de las actividades de ésta área es la relacionada con la proyección de películas en los cines. En los últimos cinco años, el número de salas de cine se ha incrementado, con el propósito que este servicio sea usado por todos los estratos económicos. Por ello, debido a que las empresas de cine requieren abarcar un mayor porcentaje del mercado, buscan satisfacer las necesidades de los espectadores de cine brindándoles alternativas para el manejo de la información sobre la cartelera.

Sin embargo, la tecnología actual en nuestro país, no permite alternativas de atención al cliente novedosas y que den alta calidad de servicio al usuario. Esta última característica es importante en tanto que determina que el usuario se sienta atraído por el servicio y lo use o se aleje de él.

Debido a la necesidad de mejorar la calidad de los sistemas de atención vía telefónica, éstos están siendo automatizados por los sistemas de diálogo. En otros países, como España y Estados Unidos (ver referencias), este sistema es usado para atender las consultas vía telefónica, en los cuales la persona llama y la consulta es atendida por

una computadora, que emulando el comportamiento humano, da respuesta a lo preguntado. En tal sentido, la entonación con la cual la computadora emita la respuesta tiene un papel importante, pues es el elemento que dota de naturalidad al sistema y permite una mejor comprensión del texto sintetizado. Diversos modelos tratan de aproximar esta característica, sin embargo la mejor aproximación de la entonación se logra en base al estudio del comportamiento de un corpus base (oraciones leídas y grabadas considerando diferentes elementos de entonación) adecuado a la aplicación a desarrollar.

En nuestro país, los medios de acceso para conocer información de horarios y tarifas de cine es limitado. El flujo de información se realiza a través de un medio escrito, Internet o vía una llamada telefónica, en la cual la operadora atiende la consulta del usuario. La vía telefónica es la opción preferida por los usuarios, pues constituye una forma ágil de transmisión de información, además de ser una fuente confiable. Sin embargo, ese servicio está limitado a un horario de atención y expuesto a ser un servicio de calidad variable, de acuerdo a la disposición de la operadora para atender la llamada.

La presente investigación plantea como hipótesis que dada la necesidad de los usuarios de los cines de acceder a la información de una forma ágil vía telefónica, sumada a los factores que hacen variable la calidad de servicio ofrecido por dichas operadoras así como el factor económico que implica mantener a una persona durante un horario fijo en el cargo, entonces, la implementación de un sistema de diálogo usando síntesis de voz que utilice un modelo de entonación, permitirá obtener una alta calidad de voz sintetizada, la que a su vez permitirá tener un sistema de información que evite los defectos que la atención humana introduce, con un mejor rendimiento de atención y una reducción de costos de mantenimiento para la empresa. Esto además introducirá al usuario en una cultura de adaptación a las nuevas tecnologías que rigen

los sistemas de información a nivel internacional, siendo esto una clave de mercado para las empresas de cine las que buscan brindar nuevas experiencias de cine; de esta forma el uso de esta nueva tecnología se constituye en una marca distintiva del servicio que se ofrece, por lo tanto en una estrategia de posicionamiento de mercado.

El objetivo de esta tesis es implementar un módulo, para el programa sintetizador de voz, que se encargue del modelado de entonación, esto contribuirá a que el producto final (voz sintetizada) posea naturalidad y brindará al usuario del sistema una experiencia novedosa y agradable de acceso a la información. Cabe indicar que toda esta investigación es parte de un proyecto mayor: “Sistema de Dialogo Hablado aplicado a una Recepcionista Telefónica Automática", aprobado por la DAI (Dirección Académica de Investigación) en el 2004. Dicho proyecto consiste en el desarrollo de un sistema de diálogo para brindar información en cines, por lo tanto, todos los módulos del proceso están en actual desarrollo. Con esta tesis, se pretende contribuir a la realización de dicho producto mayor.

La metodología empleada considera una profunda investigación sobre las técnicas empleadas en la actualidad para el modelado de la entonación, así como un estudio de algunos rasgos de la prosodia, métrica y lingüística en el español. Por tanto, la interacción con profesionales de otras áreas (lingüística, acústica y comunicaciones) ha sido importante para poder implementar esta tesis, sobre todo en la etapa de obtención de corpus base, pilar sobre el cual se plantea el modelo de entonación. Este conjunto de actividades se ha complementado con un diagnóstico de los sistemas actuales de información que dan las empresas de cine, realizado a través de entrevistas a usuarios del sistema y entrevista a profesionales de una empresa de cine (Multicines CinePlanet).

Dado que el tema de la presente investigación es de actual desarrollo, el principal medio de adquisición de información ha sido Internet. Se ha puesto especial cuidado en seleccionar artículos publicados por los grupos de investigación de tecnologías del habla (ver referencias) y que dicha información sea lo más actual posible.

La investigación está desarrollada en cuatro capítulos. En el primer capítulo se analiza la problemática que envuelve el servicio de atención al cliente vía telefónica en los centros de cine. El análisis se divide en tres partes: (i) análisis de la demanda de los servicios de entretenimiento y cultura en nuestro país, (ii) análisis del servicio de atención al cliente en los cines del país y (iii) análisis del servicio de atención al cliente en una cadena de cines en particular. Luego de ello, se analizan los factores internos al proceso de atención al cliente vía telefónica para cines.

En el segundo capítulo, se muestra el estado actual de la tecnología sobre los sistemas de diálogo y las diversas técnicas actuales para modelar la entonación. En la búsqueda del modelo más adecuado para la aplicación a desarrollar, se definen los conceptos básicos para poder abordar el tema. Con ello, el modelo teórico queda definido.

El tercer capítulo, muestra todo el procedimiento seguido en la implementación del modelo de entonación elegido. Dicha descripción está centrada en tres puntos: (i) la obtención del corpus base de unidades, (ii) el etiquetado de las unidades y (iii) el desarrollo del programa que dado un texto obtenga los parámetros de entonación por cada unidad de entonación elegida.

En el cuarto capítulo se hace la evaluación del modelo de entonación implementado, para ello, se elaboran tablas con criterios de evaluación que permitirán cuantificar la exactitud del modelo. Se hace una evaluación a nivel de sílaba y frase entregada.

Como conclusión principal, la presente investigación establece que con el modelo de entonación en base a corpus implementado en esta tesis se alcanza una exactitud del 75%. De igual forma, podemos concluir también que la metodología desarrollada para obtener el modelo de entonación en base a corpus representa una buena opción para este objetivo.

Para finalizar, quiero agradecer a mi familia por ser la motivación que me llevó a hacer esta tesis y por todo el soporte que me dieron durante este desarrollo; a los profesores Andrés Flores y Renzo Astorne, por sus enseñanzas y por acompañar con sus consejos todo el trabajo realizado para culminar esta tesis; y a todas aquellas personas que con su insistente aliento me ayudaron para hacer de esta tesis una realidad.

CAPÍTULO 1:

LOS SERVICIOS DE ATENCIÓN AL CLIENTE EN CINES DE LA CIUDAD DE LIMA

La demanda de cultura y entretenimiento en nuestro país es creciente, y con ella, el interés de las personas hacia las salas de cine va en aumento. En los últimos cinco años, el número de salas de cine se ha incrementado, llegando no sólo a los distritos considerados pudientes, sino también se ha extendido hacia los conos de la capital. Por ello, dado que las empresas de cine quieren llegar a más gente, buscan satisfacer las necesidades de los espectadores de cine, brindándoles alternativas para el manejo de la información sobre la cartelera.

En el presente trabajo, se analizará la problemática que envuelve el servicio de atención al cliente vía telefónica en los centros de cine, basado sobre la experiencia de la cadena Multicines CinePlanet. El análisis se divide en tres partes: (i) análisis de la demanda de actividades de esparcimiento en el país, (ii) análisis del servicio de atención al cliente en los cines del país y (iii) análisis del servicio de atención al cliente en una cadena de cines en particular. Conocido ello, se procede a analizar factores internos al proceso de atención al cliente vía telefónica. Luego de este análisis se puede resaltar la necesidad de la innovación tecnológica para atender de forma eficiente, con rapidez y calidad de servicio al creciente número de espectadores de cine que ha encontrado en la vía telefónica un medio para solicitar información de cartelera de cine.

1.1. La demanda de las actividades de esparcimiento en la ciudad de Lima

El área de esparcimiento y cultura es una de las áreas que capta mayor interés por parte de los consumidores. De acuerdo a los estudios realizados por el Instituto Nacional de Estadística e Informática [INEI 1996], el rubro más atractivo desde el punto de vista del incremento de la demanda de servicios ante un incremento en los ingresos familiares es el que corresponde a los Servicios No Mercantes prestados a los Hogares, el cual abarca las áreas de esparcimiento, diversión, servicios culturales y de enseñanza. Este incremento se da en los hogares de todos los estratos. De igual forma, el comportamiento del gasto en consumo en estos servicios es proporcional al número de miembros de los hogares.

Tal como lo muestra la publicación "Oferta y Demanda Global 1991-2002" [INEI 2003], el consumo en las áreas de esparcimiento y diversión se ha incrementado. Este comportamiento, ha dado lugar a que la inversión en el sector cultura se incremente en estos últimos años, siendo, una de las áreas: el cine.

El número de espectadores de cine en nuestro país ha ido en aumento, tal es así que actualmente se tiene cerca de 30 complejos de múltiples salas de cine en Lima y algunas ciudades del interior del país.

El interés de las personas por ir al cine, se ha incrementado debido al desarrollo tecnológico visto en los efectos especiales de las películas y a los avances en cuanto a desarrollo de video y audio mostrado en la proyección de las mismas. Estos son factores que envuelven al espectador en un mundo de magia y lo invitan a vivir la experiencia del cine.

Así también, la publicidad en torno a los estrenos que hacen las empresas productoras de películas de cine y las promociones que se dan en los cines (en cuanto a precios)

para combatir la piratería hacen que la oferta de ir al cine sea atractiva para los espectadores.

1.2. Actividad de atención al cliente en los cines de la ciudad de Lima

Así como el número de espectadores se ha incrementado, la necesidad de saber más sobre lo que ofrecen los cines y la información sobre cartelera de cine también se ha incrementado. Para conocer un poco más de cerca los factores dentro de los servicios de atención al cliente en los cines, se realizó una entrevista con la Srta. Carmen Muñiz, Jefa del Departamento de Atención al Cliente de la Empresa CinePlantet. Aproximadamente, dicha cadena de cines recibe por semana 5000 llamadas vía telefónica y se atiende a información sobre películas. Esto significa un 5% de los asistentes por semana al cine. Los espectadores, cuyas edades van entre los 16 y 25 años principalmente, buscan que su comunicación sea lo más ágil posible, pues valoran mucho el tiempo que puedan estar haciendo uso del teléfono. Una comunicación para este tipo de información dura aproximadamente 50 segundos, en los cuales el usuario además de recibir la información solicitada busca tener un trato agradable por parte del operador. La atención vía telefónica se desarrolla como una llamada común de teléfono.

Además de la atención ofrecida por un operador telefónico, algunos cines cuentan con la opción de atención vía telefónica vía marcado de tonos (CineMark), en el cual, el usuario va eligiendo el tipo de información al que quiere acceder por medio del marcado de tonos de teléfono. Esta opción no es bien recibida por los usuarios puesto que dilatan mucho la comunicación, lo cual desalienta la llamada y a largo plazo lleva a la inutilización del sistema, y posible reducción del número de asistentes al cine (aproximadamente un 5%). Sumado a ello, esta alternativa no da al usuario un trato personalizado, consideración importante cuando se habla de trato a personas.

Otra opción de manejo de información es a través de la página Web, la cual, en el caso de CinePlanet recibe un promedio de 12000 visitas semanales. El inconveniente de este medio es que la información no siempre está debidamente actualizada, por ello, el usuario que accede a este tipo de vía no tiene información en el momento preciso.

1.3. Los servicios de atención al cliente vía telefónica en los cines de la ciudad de Lima

El servicio de atención al cliente vía telefónica, en el caso de la cadena Multicines CinePlanet, se brinda desde hace 3 años. Este surgió debido a que cada local de dicha cadena de cine poseía una línea de teléfono, la cual recibía numerosas llamadas solicitando información sobre películas. Por ello, mantener a un grupo de personas por cada local de cine, hacía costoso el mantenimiento del servicio. A raíz de ello, se decidió tener una línea dedicada para consultas, donde se centralizara el servicio de información. Actualmente este servicio es atendido por cuatro personas con ayuda de una computadora en la cual van visualizando la información solicitada.

El servicio se da en el mismo horario de proyección de las películas (de 11:00am hasta 08:00pm) y los operadores rotan en la atención del mismo, excepto en el horario de la tarde (de 3:00pm a 7:00pm) en el cual, debido al gran volumen de llamadas, los cuatro operadores deben estar atendiendo.

El cine, con el fin de dar un servicio más ágil y rápido a sus potenciales clientes, elige como operadores a personas dentro de su personal, pues para ellos es importante mantener a las personas que ya estén inmersas en la cultura de la compañía. Como se ve, la empresa no busca cubrir el puesto con personas de alguna formación técnica en cuanto al manejo de un centro de llamadas, este conocimiento se adquiere de forma empírica en el transcurso del desarrollo de la persona en el puesto.

1.4 Opciones del usuario ante la necesidad de información sobre cartelera de cines

Frente a la iniciativa del usuario de consultar información, surgen como opciones: la vía telefónica, algún medio escrito (periódico, revista) o vía Internet.

En el caso de elegir por algún medio escrito, el usuario deberá gastar dinero y tiempo en la adquisición de su fuente de información. En el caso de buscar la información vía Internet, encuentra el inconveniente de que dicha información no siempre está debidamente actualizada.

Si el usuario se decide por llamar vía telefónica al establecimiento de cine para solicitar la información, buscará que la llamada dure lo menos posible. Este servicio está limitado a un horario de atención y la calidad del servicio que se ofrece es un factor variable, que disminuye a medida que las horas de trabajo y volumen de llamadas aumentan. El diagrama de flujo de este proceso puede apreciarse en la Figura N°1.

Actualmente, la atención vía telefónica en otro tipo de servicios se realiza a través de los call center, el cual busca dar un servicio inmediato al cliente por medio del teléfono. El avance tecnológico, ayuda considerablemente a reducir el tiempo de respuesta lo cual permite que el número de servicios ofrecido a través de este medio esté en aumento, pues abarca desde el marketing al servicio posventa, todo a través de un único número de teléfono.

Debido a que los usuarios día a día exigen un valor agregado al servicio que reciben, los call centers tradicionales se han convertido en lo que se conoce como contact centers, donde se integran diversos canales de interacción con la empresa como teléfono, fax, e-mail, con la misma sencillez y eficacia que proporciona una solución de

centro de atención telefónica y ofreciendo a los clientes un único punto de contacto para resolver sus necesidades.

Gracias a los avances tecnológicos, el call center ha dejado de ser un simple gestor de llamadas para convertirse en un importante elemento del sistema CRM (Customer Relationship Management) de las empresas. Por ello, las personas que están al frente de los call centers deben ser personas capaces de manejar simultáneamente y con gran destreza el teléfono, la informática y a la persona que está al otro lado, es decir, la persona debe estar calificada para el puesto.

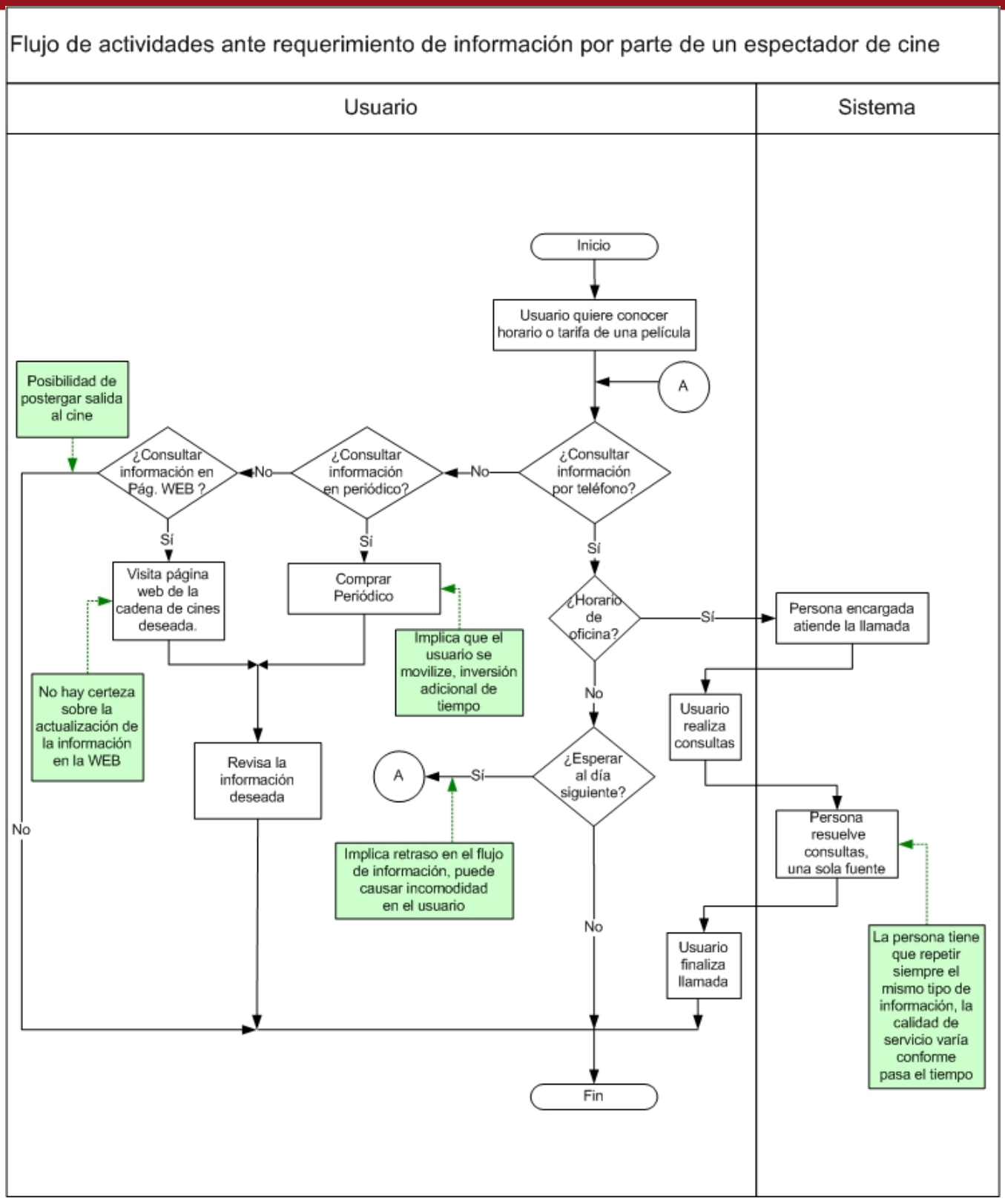

Figura N°1. Diagrama de Flujo del espectador de cine

ante la necesidad de información

1.5 Declaración del Problema

La demanda de las personas por los medios de entretenimiento, diversión y cultura tiene una tendencia de consumo ascendente. Para atender debidamente al volumen de consumidores de servicios de esta área, se debe tener un manejo de información que les permita conocer las posibilidades de esparcimiento que se ofrecen. Esta realidad, también afecta al mercado del cine, en el cual la gran variedad de películas que se ofrece, y el gusto de ver una película de alta calidad en video y audio en las salas de cine, hace que el número de espectares sea creciente.

Por ello, los espectadores experimentan la necesidad de información de carteleras de cine. Los medios de acceso para conocer información de horarios y tarifas de cine actualmente es limitado, pues la tecnología actual en nuestro país sólo permite que el flujo de información sea a través de un medio escrito, Internet o vía una llamada telefónica, en la cual la operadora atiende la consulta del usuario.

Los usuarios prefieren acceder a la información vía telefónica, pues es una forma ágil de transmisión de información; sin embargo, esta opción se ve limitada por el horario de atención en el cual se puede acceder a ella. Así también, se encuentra otro factor variable que afecta la eficiencia de calidad del sistema de atención al cliente: la monotonía que implica para el operador repetir el mismo tipo de información durante toda su jornada de trabajo. Las personas que atienden estos centros de atención no cuentan con una preparación técnica especializada. En contraste, los call centers, cuentan con operadores que además de tener una voz melodiosa tienen preparación para manejar teléfono, la informática y a la persona que está al otro lado. Así, las operaciones que maneja el sistema se realizan de una manera mucho más rápida. Dado que todo el sistema está expuesto al factor introducido por el comportamiento del usuario, la preparación del operador para el manejo de las situaciones que se le puedan presentar es importante para el buen rendimiento del sistema.

1.6 Conclusiones

Dado el incremento en el número de los espectadores de cine, es necesario brindar un sistema de información que satisfaga las necesidades de los usuarios.

La vía telefónica es forma usada por los clientes para conocer información sobre la cartelera de cine. Por ello, las empresas de cine están en la búsqueda de nuevas tecnologías que les permitan atender al creciente mercado que buscan llegar.

La dinámica generada en la transmisión de información vía telefónica cuando se solicita información de cine, debe ser ágil de tal forma que el usuario obtenga en poco tiempo la respuesta a su consulta. Esta característica hará que el usuario se sienta satisfecho con el servicio y lo use.

La falta tecnológica vista en los servicios de atención al cliente que se ofrecen actualmente en el país, refleja la necesidad de ingresar al mercado nuevas tecnologías para la atención al cliente, que con agilidad y calidad de trato sean capaces de atender los requerimientos de los usuarios.

CAPÍTULO 2:

MODELO DE ENTONACION POR CORPUS PARA UN SISTEMA DE INFORMACION DE CARTELERA DE CINES

Encontrada la necesidad de ofrecer otra alternativa para servicio de atención al cliente a las empresas de cine, se plantea un sistema de diálogo que permita automatizar el servicio de información.

El presente capítulo, se inicia presentado el estado actual de la tecnología sobre los sistemas de diálogo, mostrando que para la aplicación a realizar es necesario contar con una voz sintética lo más natural posible, por ello, también se presentan los modelos actuales para representar la entonación en español. En base a ello, se definen conceptos básicos para poder abordar el tema. Con ello, el modelo teórico queda definido.

La investigación realizada muestra que los sistemas de diálogo se presentan como buenas alternativas de automatización de atención al cliente. Por ello, en estos sistemas se debe tener especial cuidado con la calidad de voz a emitir, presentándose al modelo de entonación por corpus como aquel que ofrece mejores resultados debido a la precisión y variedad de unidades que conforman el corpus base.

2.1 Estado del Arte

2.1.1 Presentación del asunto de estudio

Los sistemas de diálogo son desarrollos que permiten al hombre interactuar verbalmente con la computadora. Actualmente, los sistemas de diálogo son utilizados en el extranjero en los sistemas de información vía telefónica, para automatizar la función de las operadoras telefónicas y mejorar así la calidad de servicio ofrecido. Para ello, está compuesto de varios módulos: (i) reconocimiento del habla, (ii) comprensión del habla y generación de respuestas, y (iii) síntesis de voz. Este último permite generar la voz sintética a partir de un texto de entrada, lo cual hará audible la respuesta de la computadora. En el desarrollo de estos sistemas, se busca que la calidad de voz generada posea naturalidad, por ello, en la medida en que el texto sintetizado posea una entonación adecuada será más comprensible y mejor recibido por el usuario.

En el presente estudio se muestra las tecnologías y modelos que actualmente se utilizan para enfrentar el modelado de la entonación dentro de los sistemas conversores de texto a voz.

Se empieza por presentar los modelos tradicionales usados para la entonación, finalmente se muestran los modelos y formas más modernas de modelado, que van de acuerdo a la técnica de síntesis de voz a usar.

De acuerdo a ello, se observa que el modelado de entonación en base a corpus es aquel que permite mejores resultados y está a la par con la técnica actualmente más empleada en síntesis de voz: concatenación por selección de unidades.

2.1.2 Estado actual de la investigación

Se puede considerar que hay dos tendencias en cuanto al modelado de la entonación se refiere. Estas tendencias están relacionadas básicamente al tipo de sintetizador de voz que use el modelo: los sintetizadores que no usan como referencia una base de datos de voz grabada (corpus) y los que sí cuentan con ella. Para el primer caso, son muy usados los modelos Pierrehumbert, IPO y Fujisaki.

Según Mari Cruz Amorós (AMORÓS 2005), la curva de entonación puede ser modelada en base a un sistema de niveles que identifican al acento tonal, tal como lo muestra el modelo Pierrehumbert. En este modelo, la entonación se presenta como una secuencia de tonos asociada a una simbología. Este modelo, uno de los pioneros en el modelado de la entonación en inglés, identifica dos tipos de tonos básicos: alto (H) y bajo (L), en cuya combinación forma diferentes tonos. Estas combinaciones se encuentran en el sistema de transcripción fonética TOBI (Tones and Break Indices) y constituye un marco de trabajo para diferentes grupos de investigación de diferentes lenguas (actualmente alrededor de once idiomas tienen algún sistema TOBI desarrollado) donde se establecen convenciones simbólicas para identificar los tonos que se encuentren en cada idioma. En el español, este modelo ha sido empleado por Sosa, quien en su libro “La entonación en el español” muestra que la transcripción usada en TOBI puede adaptarse también al español obteniendo buenos resultados, tal como se menciona en la revisión del libro de Sosa (PRIETO 2000). Otro trabajo desarrollado con este modelo para el español fue el presentado en “Modelos De Entonación Analítico Y Fonético-Fonológico Aplicados A Una Base De Datos Del Español De Buenos Aires” (GURLEKIAN et.al. 2004) donde se desarrolla un modelo para la entonación del español argentino en el cual concluyen que el modelo TOBI describe satisfactoriamente los contornos entonativos estudiados.

Sin embargo, tal como menciona Escudero Mancebo (2002) aplicar este modelo al español implica crear un amplio conjunto de reglas que permitan la correspondencia numérica entre los tonos y la evolución de F0 (frecuencia fundamental de vibración de cuerdas vocales al producir sonidos), por lo cual el modelo se torna un tanto complejo.

Por otro lado, escuelas de investigación han desarrollado su propio modelo de entonación. Tal es el caso del Instituto para la Investigación de la Percepción (IPO - Institute for Perception Research), parte del "Centro de Investigación de Sistemas de Interacción con el Usuario" de la Universidad Técnica de Eindhoven, en Holanda. El modelo creado es conocido como modelo IPO, plantea que sólo algunos movimientos del pitch (frecuencia fundamental) producen la sensación de entonación por lo cual sólo éstos deben ser identificados y modelados. Para implementar el modelo IPO se requiere:

- Estilizar el contorno de F0 mediante segmentos rectos y con ello crear una "copia cercana" de la muestra original.
- Revisar el comportamiento de los segmentos rectilíneos, en duración y amplitud.
- Crear en base a lo estudiado, una gramática que identifique y limite los movimientos tonales para una determinada lengua.

El punto débil de este modelo, está en que al realizar aproximaciones lineales se puede perder características microprosódicas del texto, las cuales son perceptibles por el oyente, tal como lo menciona Escudero Mancebo (2002).

Uno de los modelos más sólidos en la actualidad, es el obtenido en el modelo Fujisaki, el cual representa la entonación como una serie de comandos que siguen una secuencia en el tiempo (MARIÑO 2002). Se identifican dos comandos base: tipo frase y tipo acento, los cuales, se aplican a la entrada de un filtro que modela el tracto vocal

y cuya salida es el perfil de entonación deseado. Estos comandos responden, según (GURLEKIAN et.al. 2004), a los acentos tonales y acentos de frase. Sin embargo, resulta complicado identificar automáticamente si a una frase le corresponde un comando tipo frase o un comando tipo acento.

Dada la tendencia actual de generar voz sintética de mayor naturalidad, surgen los sintetizadores de voz que realizan la síntesis teniendo como referencia una base de datos de voz grabada (corpus). Este tipo de sintetizadores se conoce como "Sintetizadores de Voz por Selección de Unidades". En este caso, para modelar la entonación se define una serie de parámetros que caracterizarán a la unidad de entonación elegida, la cual es además identificada de acuerdo a algún tipo de análisis prosódico, generalmente basado en la métrica (considerando el número de sílabas de la frase a sintetizar). Los parámetros que definen la entonación son: pitch (F0), energía y duración; así como el grupo acentual y melódico al que pertenecen, tal como se menciona en "Spoken language processing: a guide to theory, algorithm, and system development" (HUANG 2001). La particularidad de este método es que al trabajar sobre la base de voz grabada, los resultados obtenidos son altamente favorables.

2.1.3 Síntesis sobre el asunto de estudio

- Debido a la necesidad de mejorar la calidad de los sistemas de atención vía telefónica, éstos están siendo automatizados por los sistemas de diálogo.
- La entonación tiene un papel importante en los sintetizadores de voz pues es el elemento que dota naturalidad al sistema y permite una mejor comprensión del texto sintetizado.
- Existen diversos modelos para la entonación, sin embargo la mejor aproximación de la entonación se logra mediante la parametrización de las unidades de un corpus base.

- Cada elemento del corpus debe estar etiquetado con información que defina su entonación, la información de la etiqueta debe contener los parámetros que definan la entonación de las unidades elegidas.

2.2 Los sistemas de diálogo

Los sistemas de diálogo son sistemas que reciben como entrada frases expresadas de forma oral y generan como salida frases expresadas también de forma oral. Su finalidad es emular el comportamiento inteligente de un ser humano de poder comunicarse. En la actualidad, estos sistemas son usados por diversas empresas para proporcionar información de forma automática (por ejemplo, horarios de salida de aviones, partes meteorológicos, estado de cuentas bancarias, etc.).

De acuerdo al Instituto Mexicano de Telemarketing (IMT) se puede distinguir diversos niveles de sistema de diálogo, los cuales se diferencian por el tipo de datos de entrada que recibe y por consiguiente por el tipo de respuesta que emite:

IVR Hablado (*Interactive Voice Response*): permite sostener un diálogo dirigido en base a la detección, por parte del sistema, de palabras clave. Es el nivel básico de comunicación oral entre la computadora y el usuario.

Sistema de lenguaje natural: recibe como entrada frases en lenguaje natural de forma continua, permitiendo una iniciativa mixta (tanto por parte del usuario como por parte del sistema) de conversación.

Sistema de diálogo natural: recibe como entrada frases en habla espontánea y permite un diálogo abierto. Su complejidad es mayor dado que se aproxima mucho más al tipo de comunicación que se da entre dos personas.

2.3 Síntesis por Selección de Unidades

Un sintetizador de voz permite convertir texto a voz, es decir, genera voz sintética. Se utilizan muchas formas para sintetizar voz, sin embargo, la técnica más usada en la actualidad es: Concatenación por Selección de Unidades (FLORES et.al 2001; GTS UVIGO; Natural Vox). La voz sintética es generada por la concatenación y modificación de las unidades, seleccionadas de un corpus, necesarias para componer la secuencia de sonidos que se quieren producir. Las unidades pueden ser de diferentes tipos: unidades menores, como demisílabas, hasta frases completas, todas ellas en un mismo corpus.

Para elegir la unidad del corpus que mejor represente a la unidad que se quiere producir, se evalúan dos tipos de costos:

1) Costo de concatenación: surge debido a que los segmentos de voz utilizados están condicionados por la coarticulación producida por el contexto de donde se extraen.
2) Costo de selección: debido a que las características prosódicas de los segmentos de voz del corpus pueden ser diferentes a la prosodia requerida para el habla sintética.

Para minimizar estos costos, se utiliza un corpus amplio. Se busca tener la mayor cantidad de unidades de voz grabada en diferentes condiciones de contexto fonético y prosódico. Estas condiciones están limitadas por la aplicación a desarrolllar.

La principal ventaja de este método es la mejora de calidad y naturalidad de la voz sintética, respecto a los otros métodos de síntesis. Este tiene un mejor desempeño cuando el corpus es diseñado para una aplicación determinada, pues las unidades

consideradas en el corpus responderán al contexto de desarrollo del trabajo. En este caso, la voz sintética generada podría ser prácticamente indistinguible de una voz natural. (VILLARRUBIA et.al. 2002)

Dado que el costo de selección pretende estimar la cercanía entre cada segmento de la base de datos y su homólogo de la secuencia que se quiere sintetizar, permitirá valorar la idoneidad de la elección de cada uno de los segmentos de forma aislada para representar a las unidades de la frase de la referencia.

2.4. Entonación en Base a Corpus

Este modelo de entonación está desarrollado para los sintetizadores de voz que usan concantenación en base a la técnica de "Selección de Unidades" debido a que en ambos casos se toma como referencia la información contenida en un corpus de voz grabado.

Para el caso de la entonación, se busca que el corpus contenga unidades de entonación que reflejen el tipo de habla que se quiere sintetizar. Cada unidad del corpus debe estar ubicada dentro de alguna clasificación métrica y etiquetada con los valores de los parámetros que definen su entonación (ELORRIETA et.al. 2004). La identificación de la unidad en base a la métrica de la frase que se quiere sintetizar será el puente que unirá la unidad objetivo con la unidad del corpus, pues en base a dicha identificación se realizará la asignación de parámetros. En la búsqueda de la unidad se puede considerar dos casos:

(i) Si la unidad objetivo se encontrara en el corpus, la asignación de parámetros es automática.

(ii) Si la unidad objetivo no se encontrara en el corpus, los parámetros entregados corresponden a lo aproximado para dicha unidad en base al estudio del comportamiento del corpus en cada tipo de unidades consideradas.

Con el método de entonación en base a corpus se busca reducir el costo de selección en el sintetizador, incluyendo las unidades del corpus base de entonación en el corpus de síntesis.

2.4.1 Unidades de entonación

Existe diversidad de opiniones sobre la elección de la unidad que mejor represente los parámetros de la entonación (GARRIDO 1996; LOPEZ 1993).

Sin embargo, el mayor consenso está alrededor de tres tipos de unidad:

(i) La sílaba: tiene asociada la información del acento, por ello, muestra los cambios de F0 en combinación con variaciones de otras caracterísitcas como duración, intensidad y calidad.

(ii) El grupo acentual: está conformado por una sílaba acentuada y por todas las no acentuadas que le preceden. Ese grupo se utiliza para describir los patrones de entonación a nivel local pues tiene asociada la información del ritmo.

(iii) El grupo de entonación: es la mínima frase con sentido, donde no se presenta ninguna ruptura prosódica importante. Por ello, sirve para planificar el discurso. La identificación del grupo de entonación es sencilla en el discurso hablado, debido a que es aquella frase que el locutor pronuncia de forma continua, sin ningún quiebre o ruptura. Sin embargo, en el discurso escrito, resulta un poco complicado delimitar el grupo de entonación. Por ello, se define que el grupo de entonación debe contener aproximadamente un promedio de ocho sílabas. (LOPE 1995).

Como se puede apreciar, las unidades de entonación mayores se forman y se definen en base a las sílabas, ya sea por su característica (tónica o no) o por su cantidad (dependiendo del número de sílabas podríamos limitar a los grupos de entonación).

Por ello, para propósitos de esta tesis, se elige como unidad básica de entonación a la sílaba, la cual, además de representar de forma adecuada los fenómenos de la entonación, nos permitirá tener flexibilidad para formar palabras.

2.4.2. Características Cualitativas

Existen diversas propuestas para la clasificación prosódica de las unidades de entonación (LOPEZ 1993; GARRIDO 1996; ESCUDERO 2002). Estas propuestas se asemejan en tomar a la silaba con unidad mínima de entonación y diferencian en el número de clasificaciones que se le asignan a dicha unidad.

Se ha visto conveniente utilizar para el desarrollo de la presente tesis la propuesta de clasificación de López (LOPEZ 1993) en la cual define la prosodia de la sílaba en base a tres tipos de características:

(i) Tipo de grupo de entonación: se considera que en el español se pueden encontrar nueve tipos de grupos de entonación: terminativo, continuativo, contrastibo, incidental, yuxtapuesto, vocativo, apelativo absoluto, apelativo pronominal y parentético. Cada uno de ellos, asociado a un tipo de oración, por ejemplo: el grupo de entonación terminativo está asociado a las oraciones enunciativas. El número de tipos de grupos de entonación a considerar en el corpus puede limitarse de acuerdo a la aplicación a desarrollar.

(ii) Posición del grupo acentual dentro del grupo de entonación: de acuerdo a la ubicación del grupo acentual dentro de la frase (grupo de entonación) se puede distinguir tres tipos: inicial, central y final.

(iii) Tipo de acento :definido por la posición de la sílaba tónica de la palabra: agudo (acento en la última sílaba), llano (acento en la penúltima sílaba) o esdrújulo (acento en la antepenúltima sílaba).

Para hacer un poco más específica la clasificación de las sílabas, se considera una de las características propuestas por Garrido (GARRIDO 1996):

(iv) Número de sílabas del grupo acentual: este indicador servirá para aproximar mejor las características de la sílaba dentro del grupo acentual.

2.4.3. Características Cuantitativas

Estas características permitirán la identificación plena de cada unidad y la evaluación del costo de selección. Por ello, por cada unidad del corpus se deberá extraer los siguientes parámetros:

(i) Duración: esta característica es importante pues tiene información de las condiciones pragmáticas y semánticas de la unidad de entonación.

(ii) Pitch o frecuencia fundamental: hace referencia a la frecuencia fundamental de vibración de las cuerdas vocales percibir de un sonido, por lo tanto es quien lleva la información de las características personales de la voz. El pitch cambia durante la emisión de la voz, pues depende del acento, entonación, emoción y ánimos del locutor cuando habla (DELLER et.al. 1993).

(iii) Energía: lleva información sobre la tonicidad de la sílaba, los sonidos de las sílabas tónicas se pronuncian con mayor energía que los de las sílabas átonas. La energía de la señal también varía dependiendo de si el tramo es sonoro, sordo o si es una pausa.

Dado que el sintetizador de voz tiene que evaluar el costo de concatenación de la unidad, es necesario que la unidad de entonación esté identificada con valores de sus parámetros laterales (FLORES 2001), finalmente la lista de parámetros que identificarían a la unidad sería la siguiente:

(i) Pitch al extremo izquierdo

(i) Pitch al centro

(ii) Pitch al extremo derecho

(iii) Energía al extremo izquierdo

(iv) Energía al centro

(v) Energía al extremo derecho

(vi) Duración del segmento

2.5. Segmentación de Unidades

Este proceso permite cortar o segmentar frases en unidades menores. Dado que, en el modelado de la entonación por corpus, se tiene un gran número de datos grabados los cuales deben ser etiquetados, resulta eficiente realizar como primera aproximación, una segmentación automática, es decir, por medio de algún programa. Luego, en base a dichos resultados, se procede con la revisión realizada por una persona, quien una a una va escuchando las grabaciones y afinando los límites de las unidades. Realizar una primera aproximación por medio del segmentador automático, resulta de mucha ayuda para quien se encarga de la segmentación manual.

Muchas de las propuestas revisadas para la segmentación automática (TORRE 2001; ADELL 2004), consideran como parte del desarrollo una etapa basada en técnicas de reconocimiento de voz. Sin embargo, de acuerdo al trabajo “Speech Segmentation without Speech Recognition” (WANG et.al 2003), se puede diferenciar entre una consonante, vocal y pausa evaluando el nivel de energía, los cruces por cero de la señal de voz y el nivel de ruido. El algoritmo propone que cada señal a evaluar, sea cortada en tramos de 20ms en los cuales se hallan las características mencionadas, luego, por cada una de ellas, se realiza una comparación de valores umbrales de nivel de ruido y cruces por cero, con lo cual se puede distinguir si el tramo analizado es una vocal, consonante o pausa.

2.6. Modelo Teórico

Los sistemas de diálogo consideran como proceso importante el proceso de síntesis de voz, pues permitirá la realización audible de la respuesta que se quiera emitir. Los sistemas de atención al cliente están considerando a los sistemas de diálogo (en sus diferentes tipos) como alternativa para automatizar dicho proceso.

Para que dicho sistema sea aceptado por el usuario, se debe lograr que la calidad de voz emitida sea lo más cercana posible a la voz natural de una persona. Por ello, se considera importante el tratamiento de la entonación que cada frase a sintetizar debe recibir.

El modelado de la entonación en base a corpus, considera una base de datos de voz cuyas unidades estén limitadas a la aplicación a realizar. Las unidades del corpus deben ser unidades de entonación, útiles también para el proceso de síntesis. Por ello, se elige como unidad básica del corpus, para el desarrollo de esta tesis, a la sílaba, por ser la unidad que mejor representa los fenómenos de la entonación y por ser la unidad mínima de formación de la palabra. Esto nos permitirá tener una mayor posibilidad de formación de palabras.

En la búsqueda de la naturalidad de la voz sintética, el costo de selección debe ser mínimo. Por ello, el corpus debe ser amplio y debe considerar el mayor número de unidades de acuerdo a la clasificación cualitativa ya mencionada. En busca de un corpus amplio, se tendrá un gran número de grabaciones de voz las cuales deben ser segmentadas en la unidad mínima elegida (para este caso: la sílaba). La segmentación a realizar en esta tesis será manual con ayuda de algún software de edición de sonidos, así se puede realizar el etiquetado de forma simultánea.

El resultado de este módulo, constituye la entrada del módulo sintetizador el cual evaluará el costo de selección de las unidades, para, en base a ello, formar la frase audible que resultará del sistema. El costo de selección debe ser mínimo pues ello dotará de inteligibilidad al sistema y tendrá un mejor desempeño en su interacción con el usuario.

En la Figura N°2 se puede apreciar el esquema del modelo teórico.

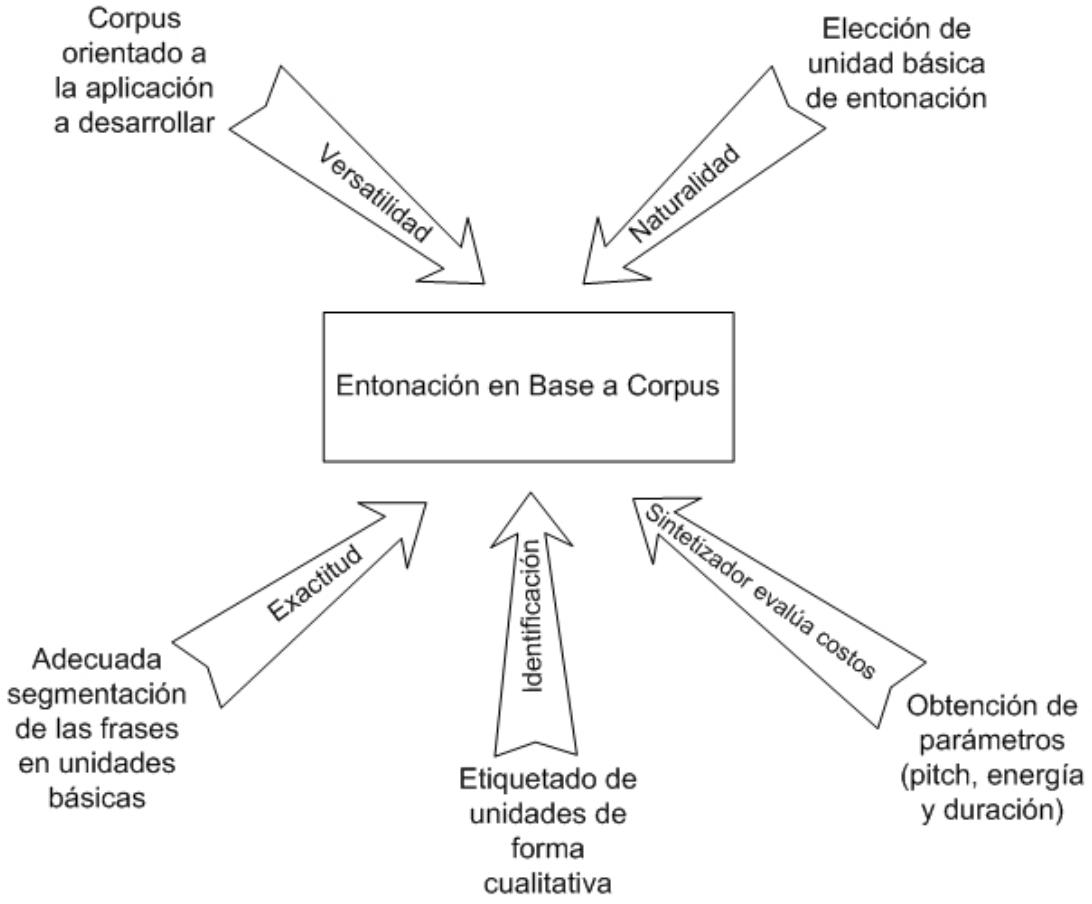

Figura N°2.- Modelo Teórico

2.7 Conclusiones

Los sistemas de diálogo se presentan como buenas alternativas actuales para automatizar el servicio de atención al cliente. Para cubrir el objetivo principal de estos sistemas de dar una atención adecuada a los usuarios, en estos sistemas se debe tener especial cuidado con la calidad de voz a emitir.

Para lograr naturalidad en la voz sintética se debe considerar un adecuado modelo de entonación. El modelo de entonación con mejores resultados actualmente es el realizado en base a un corpus, es decir, se parte del conocimiento que nos pueda dar el estudio de una base de datos de voz grabada. Para ello, se identifica el comportamiento de la energía, pitch y duración de todos los tipos de unidades considerados dentro de la aplicación a desarrollar.

La naturalidad e inteligibilidad de las frases sintetizadas son evaluadas por el sintetizador en el costo de selección y concatenación, por lo tanto, el módulo sintetizador de voz busca minimizarlos.

CAPÍTULO 3:

IMPLEMENTACION DEL MODELO DE ENTONACION

El modelo de entonación en base a corpus es el modelo que, en la actualidad, permite obtener voz sintética de gran naturalidad debido a que sus parámetros reflejan el comportamiento de un corpus real de voz.

En el presente capítulo, se describe en detalle el procedimiento seguido para la obtención del modelo de entonación en base a corpus. La descripción está centrada en tres puntos: (i) la obtención del corpus de unidades, (ii) el etiquetado de las unidades y (iii) el desarrollo de una función que dado un texto obtenga los parámetros de entonación por cada una de sus sílabas.

Se puede resaltar la importancia de contar con los servicios de una persona preparada para la lectura adecuada del corpus, pues el tener frases leídas de forma clara y bien vocalizada hace más sencilla la segmentación. Así también, se concluye que a mayor cantidad de unidades dentro del corpus, mejor la asignación de parámetros estándar por cada grupo acentual. Por ello, para procesar todo el corpus, sería adecuado contar con un programa que realice la segmentación de forma automática, lo cual haría más efectivo la segmentación del corpus, tanto por su efectividad para delimitar las sílabas de las frases como por su rapidez.

3.1 Obtención del corpus

3.1.1 Redacción del corpus

En la redacción de las frases que constituyen el corpus de entonación se tuvo cuidado en elegir el tipo de frases adecuadas para la aplicación a desarrollar. Dado que el sistema sintetizador de voz que usará este modelo está orientado a información de cines, se delimitó el tipo de información a entregar, de la siguiente forma:

- Saludos
- Información sobre horarios
- Información sobre tarifas
- Nombres de películas
- Información sobre sinopsis de películas

Para las tres primeras opciones, la información entregada es estática, es decir, es información conocida y su contenido es independiente de la película cuya información se quiera ofrecer, para este caso frases completas (es decir, aquellas frases que forman parte del diálogo ya conocido) serán grabadas en el corpus del sintetizador. Tal no es el caso de los nombres de las películas y sinopsis, en el cual, a medida que la cartelera se renueva esta información cambia. Por ello, el desarrollo de esta tesis se centra en la entrega este tipo de información.

En el análisis del tipo de texto que aparece en esta información, se observó que los textos son conjuntos de oraciones del tipo terminativas (enunciativas). Cada texto, tiene entre 3 y 5 grupos acentúales, cada uno con 6 sílabas como máximo, siendo frecuente sólo 3 o 4 sílabas por grupo acentual. Por ello, se optó por incluir en el corpus sólo oraciones del tipo terminativas, que cumplieran los criterios de clasificación mencionados en el capítulo 2. Así, se obtuvo un corpus de 50 oraciones terminativas, en las cuales se incluyeron grupos acentúales con las siguientes características:

- Grupos acentúales agudos iniciales de 1 a 6 sílabas.
- Grupos acentúales agudos centrales de 1 a 6 sílabas.
- Grupos acentúales agudos finales de 1 a 6 sílabas.
- Grupos acentúales graves iniciales de 1 a 5 sílabas (el caso de 6 sílabas es una combinación difícil de encontrar).
- Grupos acentúales graves centrales de 1 a 6 sílabas.
- Grupos acentúales graves finales de 2 a 6 sílabas (el caso de 1 sílaba es una combinación muy difícil de encontrar).
- Grupos acentúales esdrújulos iniciales de 1 a 4 sílabas (los casos de 5 y 6 sílabas son combinaciones muy difíciles de encontrar).
- Grupos acentúales esdrújulos centrales de 1 a 6 sílabas.
- Grupos acentúales esdrújulos finales de 3 a 5 sílabas (los casos de 1, 2 y 6 sílabas son combinaciones muy difíciles de encontrar).

Este corpus fue elaborado con ayuda del estudiante de lingüística Giancarlo Peña, y las oraciones están listadas en el Anexo N°1. La colaboración de un lingüista en esta etapa es importante pues por sus conocimientos en distintos niveles de la estructura de una lengua, en nuestro caso el español, tiene las nociones técnicas necesarias para elaborar el corpus con las mejores unidades.

3.1.2 Grabación del corpus

Luego de elaborar la lista de frases se procedió con la grabación de las mismas. Para ello, se contó con los servicios de una locutora, Srta. Rosario Lozano, pues por su formación poseía los conocimientos adecuados de modulación de voz, vocalización y respiración para obtener claridad en la reproducción hablada de los textos. El costo de este servicio fue de S/100.00 (cien nuevos soles) por hora de grabación. En total se registraron 60 minutos de grabación: 40 minutos para el corpus base y 20 minutos para el corpus de prueba (al cual se hace referencia en el capítulo 4).

Las grabaciones se realizaron en un estudio de grabación profesional, con el objetivo de tener un ruido de fondo muy bajo y contar con los equipos adecuados. Para ello se alquiló uno de los estudios de grabación de la empresa K'antaro Records. El costo de dicho ambiente fue de US$15.00 (quince dólares) por hora de uso.

Los equipos que se usaron fueron los siguientes:

- Una computadora PIII con tarjeta de sonido SoundBlaster, modelo Audigy 4, propiedad del estudio de grabación.
- Un micrófono dinámico AKG D3800, propiedad de Laboratorio de Procesamiento Digital de Señales. Este micrófono, además de su alta fiabilidad y buena respuesta en frecuencia, cuenta con un patrón hipercardioide polar el cual lo hace recomendable para grabaciones vocales o de lectura.
- Para conectar el micrófono con la computadora se utilizó un conector XLR hembra de 3 pines, que a través de un conector estándar se conecta a la PC.

Para registrar las señales de voz, se utilizó el software CoolEdit versión 6.0. Esta versión es la que se encontró en el estudio de grabación. Este software es propiedad de la empresa Syntrillium Software, y ha sido constantemente galardonada por ser una herramienta de mucha utilidad y de fácil uso para los profesionales de la producción de audio digital. Sin embargo, actualmente este software ya no se comercializa, pues la empresa Adobe Systems Incorporated adquirió los derechos de los programas desarrollados por Syntrillium Software (Mayo del 2003), como fruto de esta transacción la empresa compradora lanzó al mercado (Agosto del 2003) el programa Adobe Audition (una versión mejorada del Cool Edit), cuyas versiones son comerciales ahora.

Las grabaciones se realizaron a 16kHz y a 16 bits. Los archivos de audio fueron guardados en formato wav, lo cual permitía trabajarlos en cualquier otro programa para edición de sonidos.

3.2 Etiquetado del corpus base

Para el etiquetado de las frases grabadas se utilizó el programa Wavesurfer 1.8.3, el cual está disponible de forma gratuita en Internet (ver referencias). Este software ha sido desarrollado en el grupo de “Habla, Música y Audición” de la escuela de Ciencias Computacionales y Comunicaciones del Royal Institute of Technology de Estocolmo. Se eligió este software pues no sólo permite visualizar y manipular de forma sencilla los archivos de audio, sino también hacer anotaciones y transcripciones sobre las ondas de audio que visualiza.

Para poder hallar los parámetros por cada unidad, se tuvo que realizar en principio la segmentación de cada archivo de audio en sílabas, con ayuda de los paneles “Waveform”, “Transcription” y “Spectogram” del Wavesurfer. El panel “Waveform” permite visualizar la onda, el panel “Transcription” crea una línea de texto adjunta al panel “Waveform” y permite realizar anotaciones sobre cualquier parte de la onda que se está visualizando. En nuestro caso, en este panel se escribía el identificador de la sílaba, el guardarlo con la propiedad “TIMIT” permite abrir dicho texto en cualquier editor de textos, en el cual, se muestra el texto ingresado y el rango de la onda al cual hace referencia. El panel “Spectogram” muestra el espectrograma de la señal, el cual se obtiene aplicándole una ventana “Hamming” de 512 puntos. Este panel fue de ayuda para poder aproximar los límites laterales de la sílaba. La Figura N°3 muestra los paneles usados en la segmentación

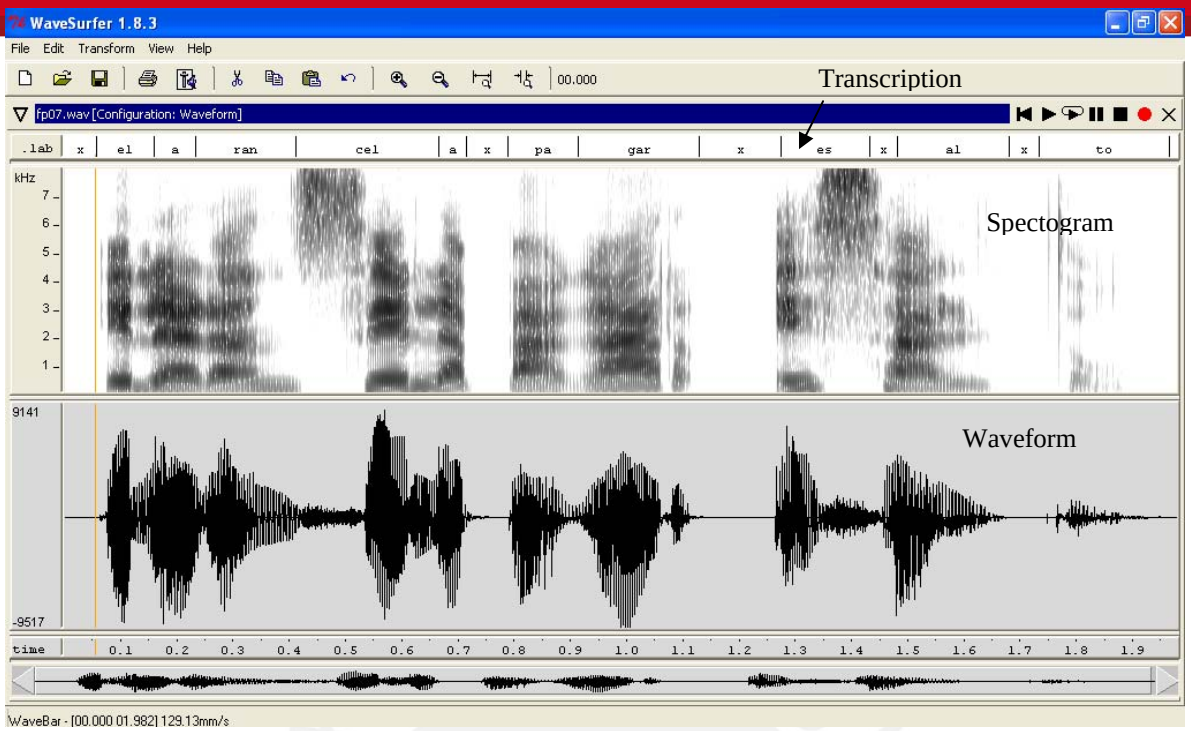

Figura N°3. Ejemplo de segmentación realizada.

Luego de segmentada la onda, se procedió a generar la curva de pitch y energía, con ayuda de los paneles “Pitch Contour” y “Power Plot” del Wavesurfer. En este caso, además de poder visualizar la curva requerida, el Wavesurfer permite obtener el valor del punto de la curva sobre el cual está situado el cursor. Este procedimiento puede ser observado en la Figura N°4. Así, por cada sílaba se obtuvo: pitch en el extremo izquierdo, pitch en el centro, pitch en el extremo derecho, energía en el extremo izquierdo, energía en el centro, energía en el extremo derecho y duración. La información obtenida fue almacenada en un archivo de Excel, siguiendo el formato que se muestra en el Anexo N°2.

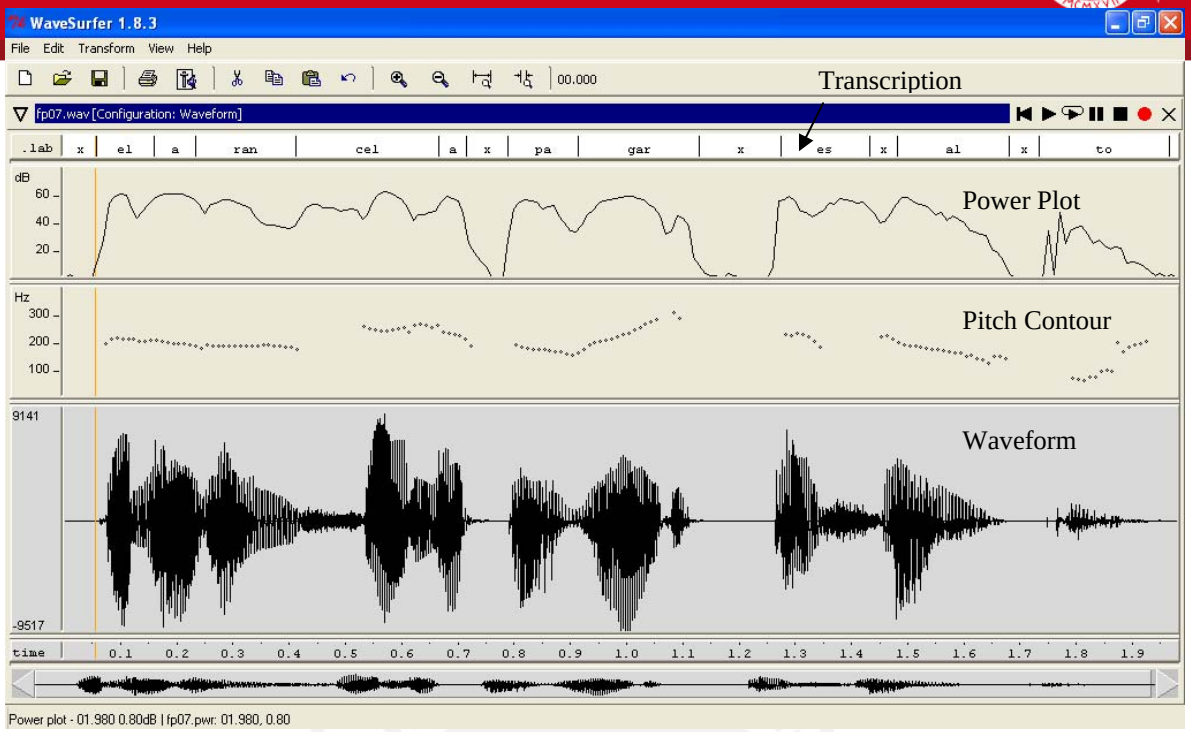

Figura N°4. Gráficas usadas en la identificación de las sílabas

De todas las frases segmentadas se obtuvo 600 unidades silábicas, de las cuales 554 resultaron diferentes. Sobre estas últimas se realiza el modelo de la entonación. Los parámetros estándar para pitch, energía y duración se hallaron mediante la observación de la tendencia de las curvas de pitch y energía de las frases del corpus. Así, se observó que para grupos acentúales iguales, la tendencia de la curva de pitch era semejante. Lo mismo ocurrió para el caso de la energía. En base a esta observación, se hallaron parámetros estándar para cada uno de los 47 tipos de grupos acentúales considerados en esta tesis.

Para almacenar los valores obtenidos por cada patrón, se crearon nueve archivos de Excel, donde se indica el valor del pitch (izquierda, centro, derecha), energía (izquierda, centro, derecha) y duración de cada sílaba de acuerdo al patrón al que pertenezcan. El orden de los archivos guardados sigue el esquema de la Figura N°5:

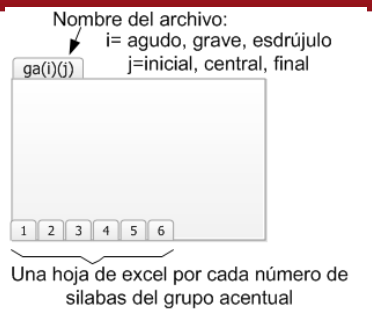

Figura N°5. Esquema de almacenamiento de Parámetros Estándar

En los Anexos N° 3, 4, 5, 6, 7, 8, 9, 10, 11 se muestran los parámetros asignados por cada tipo de grupo acentual, así como el formato utilizado.

Así también, las unidades obtenidas en la segmentación se encuentran en tres archivos de Excel adicionales donde están dividas en función al nombre de su grupo acentual (agudo, grave o esdrújulo). Dentro de cada archivo se cuenta con 3 hojas de cálculo que permiten la separación de las unidades de acuerdo a su posición dentro de la frase (inicio, centro, final). Para este caso, se sigue el esquema de la Figura N°6:

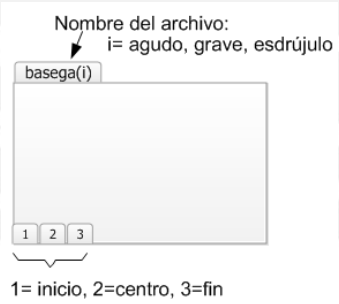

Figura N°6. Esquema de almacenamiento de Parámetros
de Unidades del Corpus

Este tipo de organización permitirá un manejo ordenado de la información y facilidad al momento en que la función implementada en esta tesis tenga que obtener los parámetros de la frase que se quiera sintetizar.

3.3 Implementación de la Función *entona()*

Para realizar la asociación de parámetros de cada sílaba, se programó una función que recibe como entrada un texto, halla las características cualitativas de cada sílaba que lo conforma de acuerdo al grupo acentual al que pertenezca y le asigna parámetros de pitch, energía y duración. Esta función se programó en Matlab 6.0, debido a que permite programar de forma sencilla y da la posibilidad de compilar el código .m para que las funciones creadas en Matlab se pueden ejecutar desde Visual C++.

La función creada es la función *entona()*, y su esquema de funcionamiento se muestra en la Figura N°7.

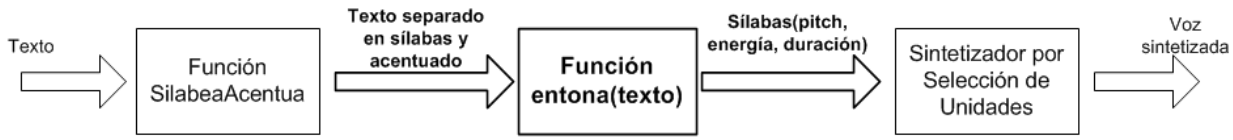

Figura N°7. Esquema de funcionamiento de la función *entona()*

La función *entona()* cuenta como parámetro de entrada a la oración, salida de la función *SilabeaAcentua()*. Esta función tiene como entrada un texto, y da como salida el texto de entrada separado en sílabas y acentuado siguiendo un formato definido, tal como se aprecia en la Figura N°8. Esta función es parte de la Tesis del Ingeniero Elí Segura (SEGURA 2003), quien implementó dicha función como parte de un sintetizador de voz usando demisílabas. Dado que dicha función constituye un módulo ya creado dentro del sintetizador que usará la función *entona()*, su desarrollo no se considera como parte de esta tesis, tan sólo se hace referencia para indicar que su salida es la entrada de nuestra función objetivo.

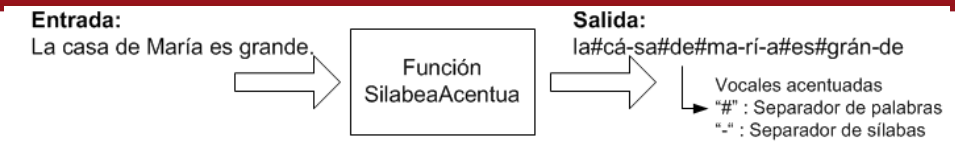

Figura N°8. Ejemplo de funcionamiento de la función *SilabeaAcentua()*

Los parámetros entregados por la función *entona()* son los parámetros que identificarán a las sílabas que el sintetizador buscará en su corpus para poder hacer la selección de unidades. Para el caso del sintetizador de voz, de encontrar en la base de datos una sílaba idéntica a la identificada por *entona()*, ésta será seleccionada y constituirá automáticamente una unidad a sintetizar. De no encontrar la sílaba exacta dentro del corpus, elegirá la sílaba del corpus que más se parezca a la entregada por entona, para en base a ella, formar la sílaba con los parámetros requeridos.

El diagrama de flujo seguido en la función entona es el que se muestra en la Figura N°9:

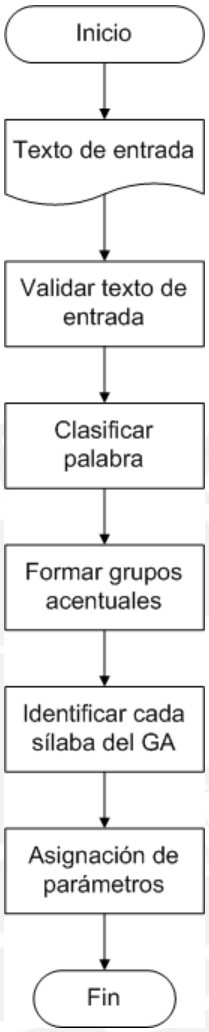

Figura N°9. Diagrama de flujo de la función *entona()*

La función principal *entona()* inicia con la declaración de una variable: continuar, la cual cambiará a medida que se desarrolle el programa. Esta variable tomará diferentes valores de acuerdo a los estados del programa, y será la llave que permita detener la ejecución del programa en caso de errores.

continuar = 0 → la función debe ejecutarse

continuar = 1 → se ha producido un error por ingresar un texto que tiene menos de tres palabras y más de 10.

continuar = 2 → se ha producido un error por ingresar un texto con alguna de las siguientes características: sólo monosílabos, un grupo acentual tiene más de 6 sílabas o la oración sólo está formada por un grupo acentual.

continuar = 3 → la ejecución de la función se ha completado con éxito, por lo tanto se deben mostrar los resultados.

Para evitar que el texto de entrada esté ligado a un sólo formato, se creó la función *cambiaformato()*, la cual recibe tres parámetros de entrada: el texto, el signo separador de palabra y el signo separador de sílaba, y da como salida el texto escrito de acuerdo a las reglas usadas en la función *entona()*. La Figura N°10 ilustra este funcionamiento. Así, si el módulo de la función *SilabeaAcentua()* podrá ser reemplazado por otro que cumpla la misma función y tenga signos separadores diferentes. En este caso, bastará indicarlos en la función *cambiaformato().*

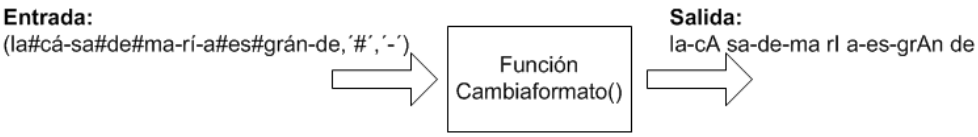

Figura N°10. Funcionamiento de la función *cambiaformato()*

Luego de normalizar el texto de entrada de acuerdo a los requerimientos de la función entona, se verifica que el texto de entrada tenga entre tres y diez palabras, debido a que la lectura de textos con más de 10 palabras de forma continua suena poco natural. De no cumplirse esta condición la función *entona()* finaliza y muestra en pantalla un mensaje de error.

Si el texto de entrada cumple los requerimientos establecidos, se procede con la identificación de cada palabra de acuerdo a la posición de su sílaba tónica, es decir, se indica si la palabra es: aguda(1), grave(2) o esdrújula(3). Para ello, se crea la función

clasificacion() la cual tiene dos entradas: la palabra que se quiere identificar y el número de sílabas de la misma, y da como salida un vector que tiene un uno en la posición de la sílaba tónica. En base a esta información y al número de sílabas por palabra se puede distinguir qué tipo de palabra es. El número de sílabas por palabra se cuenta con la función creada: *cuentasilabas().*

Una vez identificado el tipo de sílaba tónica de cada palabra, se procede con la formación de los grupos acentúales, los cuales están formados por una sílaba tónica y todas las no átonas que la preceden.

Cada grupo acentual formado, así como la posición que ocupa dentro de la frase, se ingresan a la función creada *ganalizar()*, la cual entrega como respuesta la clasificación de cada sílaba del grupo acentual, indicando: posición de la sílaba dentro del grupo acentual, número de grupo acentual al que pertenece, número total de sílabas dentro del grupo acentual, tipo de grupo acentual (1, 2 o 3) y si la sílaba es tónica o no (1 o 0).

Antes de seguir con las siguiente etapas se verifica que hasta el momento las características obtenidas estén dentro de las limitaciones de esta tesis, es decir, que no se hayan ingresado sólo monosílabos no acentuados, que los grupos acentúales no tengan más de 6 sílabas o que haya más de un grupo acentual por oración. Si la verificación es exitosa, se procede con la asignación de parámetros, sino, la función *entona()* termina su ejecución indicando en pantalla un mensaje de error.

Para la asignación de parámetros, primero se busca la sílaba en el corpus de unidades, para ello se crearon las funciones *buscabaseag()*, *buscabasegrave()*, *buscabasesd()* que de acuerdo al tipo de grupo acentual al que pertenezca busca la unidad. Estas funciones reciben como entrada a: la sílaba, el ubicación de la sílaba

dentro del grupo acentual, el número de grupo acentual dentro de la frase, el número total de sílabas dentro del grupo acentual y el número total de grupos acentúales dentro de la frase. En base a estos datos, se busca la correspondencia de la sílaba requerida dentro del corpus de unidades, de encontrarse la sílaba se le asignan los parámetros indicados en el corpus, de no encontrarse la sílaba, la salida de la función es cero.

Si la salida de la función que busca la sílaba en la base es cero, se procede con la asignación de parámetros estándar obtenidos del estudio del corpus, para ello, se utilizan las funciones creadas: *getparagudo()*, *getpargrave()*, *getparesdj()*, las cuales reciben como entrada: la posición de la sílaba dentro del grupo acentual, el número de grupo acentual dentro del a frase, el número total de sílabas dentro del grupo acentual y el número total de grupos acentúales dentro de la frase.

Una vez obtenidos los parámetros se sale del lazo principal, y se procede a afinar los resultados de la asignación de parámetros. Es decir, si la sílaba caracterizada no fue encontrada en la base de datos y se le asignaron parámetros estándar, se pregunta si es que la sílaba está formada por alguna de estas letras: s,c,z,f,j. De ser así, el pitch asignado se cambia por cero, pues como se explicó anteriormente, los parámetros estándar se hallaron para casos generales y no para casos puntuales como los son estas letras.

Finalmente la salida de la función entona es una matriz de “mx8”, donde m es el número de sílabas de la oración. Los ocho parámetros entregados por sílaba son: pitch a la derecha, pitch al centro, pitch a la izquierda, energía a la derecha, energía al centro, energía a la izquierda, duración e identificador (parámetro 8).

El parámetro 8 indica un 1 si la sílaba se encontró en el corpus y 2 si a la sílaba se le asignaron parámetros estándar. Este último parámetro es importante porque sirve como indicador para identificar qué silabas requerirían ser ajustadas por el sintetizador antes de ser sintetizadas.

3.4 Conclusiones

El corpus de trabajo debe ser leído por una persona preparada, con buena pronunciación pues esto permitirá que en la segmentación de unidades se encuentren límites más acertados.

La segmentación de las unidades es una parte importante para el modelado de la entonación en base a corpus debido a que a partir de ella se realiza el etiquetado en función de parámetros básicos (pitch, energía y duración).

Mientras más grande sea el corpus de trabajo, la asignación de parámetros estándar para cada tipo de unidad es mucho más robusta. De contar con un programa que realice la segmentación de forma automática, la segmentación se realizaría de forma más rápida, por lo tanto, el estudio del corpus podría incluir más unidades.

CAPÍTULO 4:

EVALUACIÓN DEL MODELO DE ENTONACIÓN IMPLEMENTADO

Los modelos de entonación pueden ser evaluados de forma cualitativa y de forma cuantitativa. El análisis cualitativo es aquel en el cual las personas evalúan la calidad de voz sintetizada usando el modelo de entonación descrito. A falta de un sintetizador de selección de unidades, para el cual está propuesto éste modelo de entonación, ésta prueba se omite en esta tesis.

Por su parte, las pruebas cuantitativas, consisten en, dada una frase, comparar los valores de los parámetros obtenidos del modelo de entonación con los valores reales medidos de dichos parámetros. Para tal efecto, la construcción de un corpus de prueba es necesaria.

En este capítulo, se describe el método de evaluación empleado para este modelo, así como las tablas de evaluación utilizadas para el modelo que nos permitirán obtener resultados cuantificables sobre la exactitud del modelo.

Finalmente, se obtiene como resultado principal que el modelo de entonación en base a corpus implementado en esta tesis tiene un 75% de exactitud para una frase entregada en modo texto. Este resultado puede mejorar si el corpus sobre el cual se obtienen los parámetros estándar es mayor al que se cuenta.

4.1 Descripción del Método empleado

La prueba realizada consistió en comparar los parámetros reales (pitch, energía y duración) de las oraciones de un corpus de prueba con los parámetros obtenidos con la función *entona()* para cada oración de dicho corpus.

Se elaboró un corpus de prueba, que contiene ejemplos de los tipos de grupos acentúales considerados dentro del estudio de esta tesis. El corpus de prueba está conformado por 25 oraciones (listadas en el Anexo N°12) grabadas en las mismas condiciones en las que se grabó el corpus de dónde se obtuvieron las unidades básicas de entonación.

Para ello, las oraciones del corpus de prueba, fueron segmentadas y etiquetadas una a una siguiendo el mismo método detallado en el capítulo 3. Toda la información obtenida se almacenó en un archivo de Excel, siguiendo el formato del Anexo N°13. Resultado de la segmentación se obtuvieron 272 sílabas etiquetadas.

Por otra parte, además de contar con el resultado de la función *entona()*, se crearon las funciones *grafpitch()* y *grafener()* que recibiendo como entradas los parámetros de salida de la función *entona()* y el arreglo donde se almacenan las sílabas que conforman la oración, se encargan de generar las curvas de pitch y energía aproximadas de la oración objetivo. Estas curvas se compararon con las generadas por el Wavesurfer para los archivos de audio que contienen las oraciones grabadas del corpus de prueba, y constituyó un medio más de validación de los resultados obtenidos.

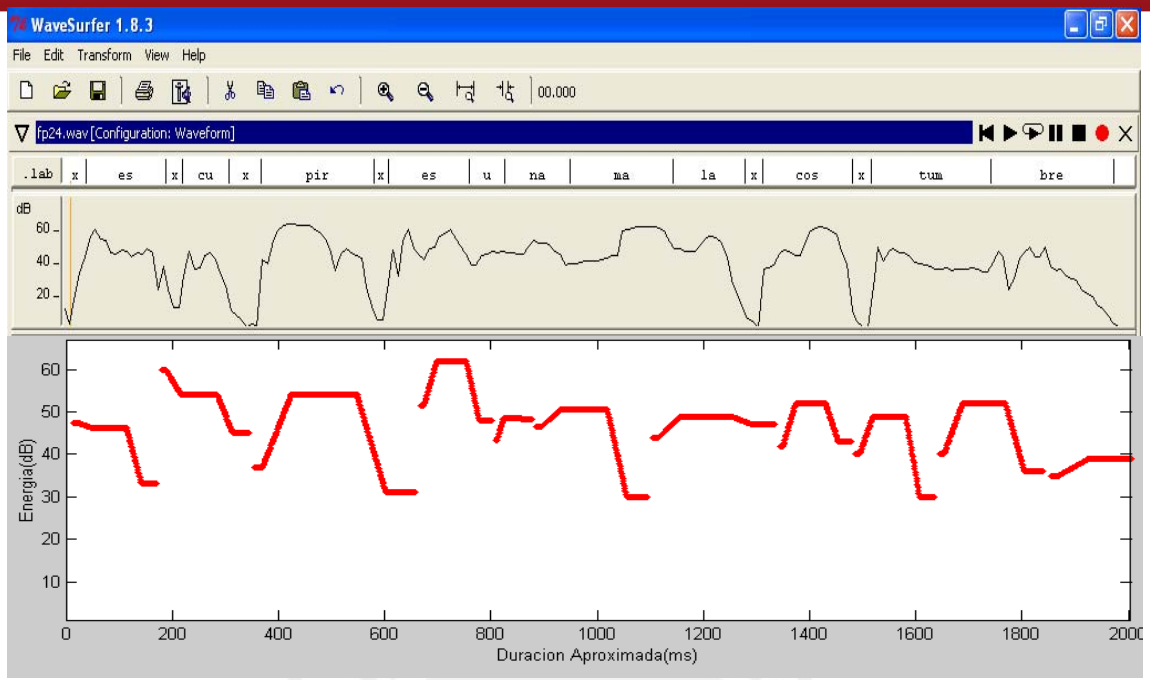

Figura N°9. Gráfica del Wavesurfer vs Gráfica obtenida en Matlab para la Energía, usando función *grafener()*

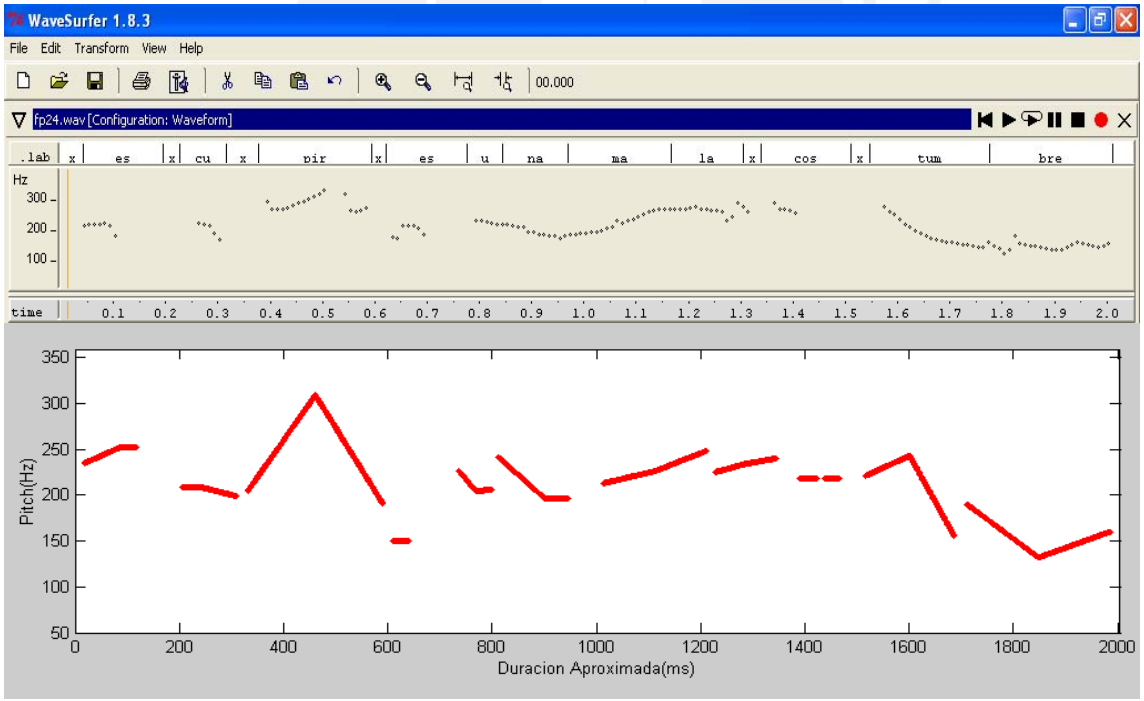

Figura N°10. Gráfica del Wavesurfer vs Gráfica obtenida en Matlab para la Energía, usando función *grafpitch()*

Los parámetros obtenidos con la función *entona()* para cada oración del corpus de prueba se almacenaron en la misma hoja de Excel donde se guardaron los resultados del etiquetado de las unidades de prueba, para poder realizar una comparación directa de ambos resultados.

Durante la evaluación se verificaron las tendencias de las curvas del pitch y entonación de cada sílaba. Dado que se tiene tres valores, tanto para el pitch como para la energía, se evaluó la tendencia de las curvas por cada tramo, es decir, primer tramo: comprendido entre el primer y segundo valor, segundo tramo: comprendido entre el segundo y tercer valor. Si la tendencia (ascendencia o descendencia de la curva) era correcta se le asignaba un puntaje de 1 a dicho tramo, de lo contrario el puntaje asignado era 0. Para el caso de la frecuencia y energía, por lo tanto, podremos alcanzar los siguientes valores:

Valor Primer Tramo	Valor Segundo Tramo	En Porcentaje	Significado
0	0	0%	Asignación de parámetros con 0% de exactitud
0	1	50%	Asignación de parámetros con 50% de exactitud
1	0	50%	Asignación de parámetros con 50% de exactitud
1	1	100%	Asignación de parámetros con 100% de exactitud

Tabla N°1. Tabla de Evaluación para Asignación de Pitch y Energía

De igual modo se preparó una tabla de evaluación para cuantificar el grado de correcta identificación de cada sílaba, se le asignó un puntaje igual a 1 si la tendencia para el tramo era la esperada y 0 si dicha condición no se cumplía.

Pitch		Energía		En Porcentaje	Significado
Valor Primer Tramo	Valor Segundo Tramo	Valor Primer Tramo	Valor Segundo Tramo		
0	0	0	0	0%	Asignación de parámetros con 0% de exactitud
1 0 0 0	0 1 0 0	0 0 1 0	0 0 0 1	25%	Asignación de parámetros con 25% de exactitud
1 1 0 1 0 0	1 0 1 0 1 0	0 1 1 0 0 1	0 0 0 1 1 1	50%	Asignación de parámetros con 50% de exactitud
1 1 1 0	1 1 0 1	1 0 1 1	0 1 1 1	75%	Asignación de parámetros con 75% de exactitud
1	1	1	1	100%	Asignación de parámetros con 100% de exactitud

Tabla N°2. Tabla de Evaluación para Asignación de Parámetros de la Sílaba

El parámetro correspondiente a la duración no fue evaluado de forma independiente, tal como se hizo para el pitch y la energía, debido a que es un parámetro que sirve de marco para la evolución de los dos parámetros ya mencionados.

Cabe indicar que las pruebas cualitativas del modelo de entonación no se pudieron realizar debido a que aun no está implementado el sintetizador de selección de unidades para el cual se ha desarrollado este modelo. Sin embargo, al contrastar los parámetros obtenidos con parámetros reales podemos decir que los resultados de este método de evaluación no serán lejanos de los que se pudieran obtener escuchando la voz sintética generada con este modelo de entonación.

4.2 Resultados Obtenidos

Los resultados de las evaluaciones realizadas para las 272 unidades segmentadas del corpus de prueba se muestran en el Anexos N° 14, 15 y 16.

En base a dichos resultados, observamos que de las sílabas del corpus de prueba, sólo 47 sílabas se encontraban en la base de datos de unidades, para el resto de unidades se aplicaron los parámetros estándar obtenidos en base al estudio del corpus de entonación descrito en el capítulo 3.

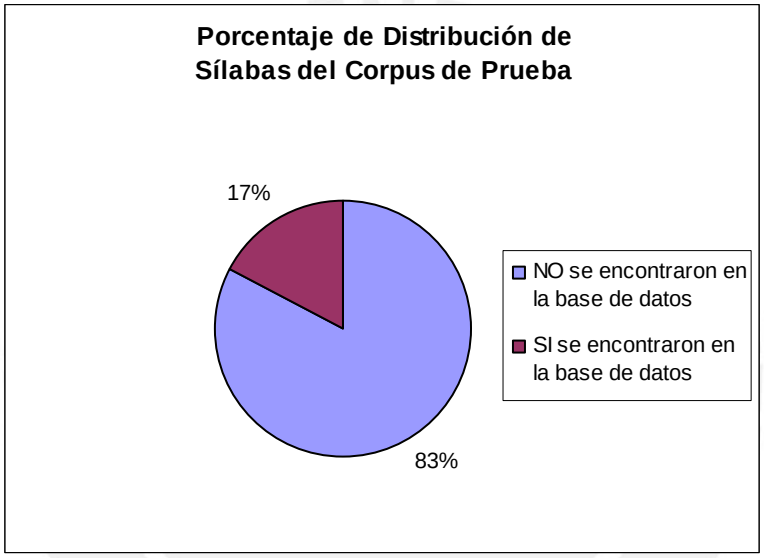

Figura N°11. Porcentaje de Distribución de las Sílabas del Corpus de Prueba

Luego de aplicar los criterios de evaluación mencionados se obtuvo que el 28.68% del número de sílabas del corpus de prueba fue identificado de forma correcta al 100% con sus parámetros. El 49.63% del corpus se logró identificar de forma correcta en un 75%, el 18.38% en un 50% y tan sólo un 3.31% en un 25%. Para este caso, todas las unidades pudieron ser identificadas correctamente en por lo menos 1 parámetro. Se observa que más del 50% del corpus fue identificado correctamente con más de 3 parámetros.

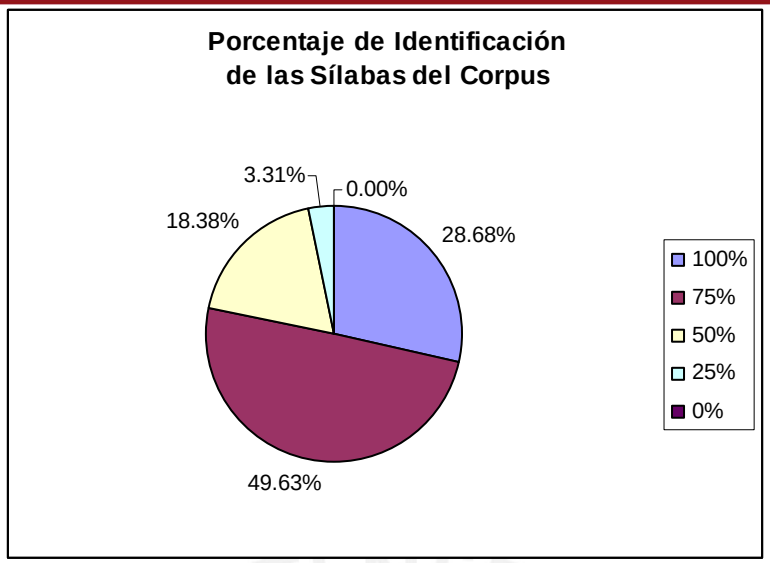

Figura N°12. Porcentaje de Identificación de las Sílabas del Corpus

Haciendo un análisis por separado del pitch y energía, obtuvimos que sólo un 2.57% de sílabas fue identificada con una tendencia incorrecta de pitch, la diferencia de porcentaje entre las sílabas que fueron identificadas de forma correcta al 100% es pequeña respecto al porcentaje de sílabas que fueron bien identificadas al 50%. Para el caso de la energía, las sílabas que fueron identificadas al 100% constituyen más de la mitad del total del corpus. Esto puede ser debido a que los valores de energía son valores continuos, a diferencia de los valores del pitch, cuya curva presenta discontinuidades. Por ello, los parámetros estándar considerados para la energía son más exactos que los considerados para el pitch.

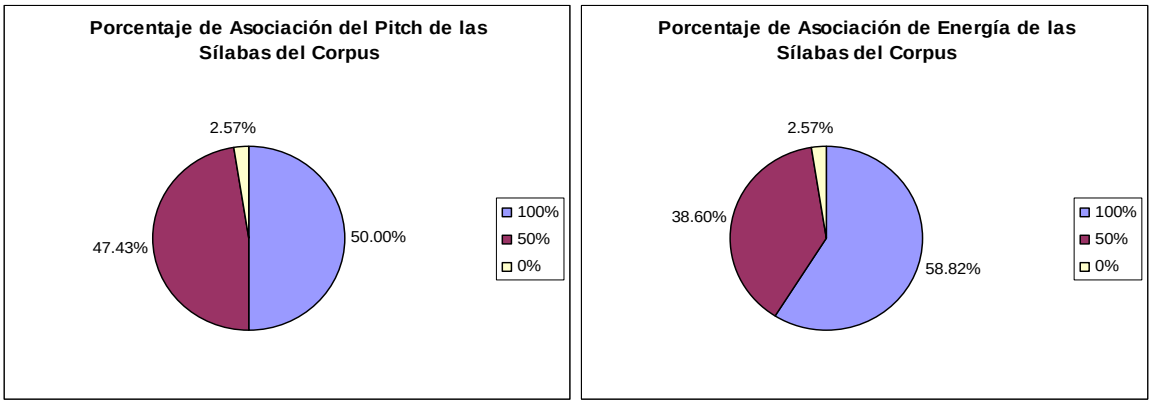

Figura N°13. Porcentaje de Identificación de Pitch y Energía de las Sílabas del Corpus

Para conocer la exactitud de los parámetros estándar asignados por cada tipo de grupo acentual, hicimos el mismo análisis considerando sólo a las sílabas que no se encontraron en la base de datos. Tanto para el pitch como para la energía más del 50% de las unidades ha sido identificado correctamente con todos sus parámetros.

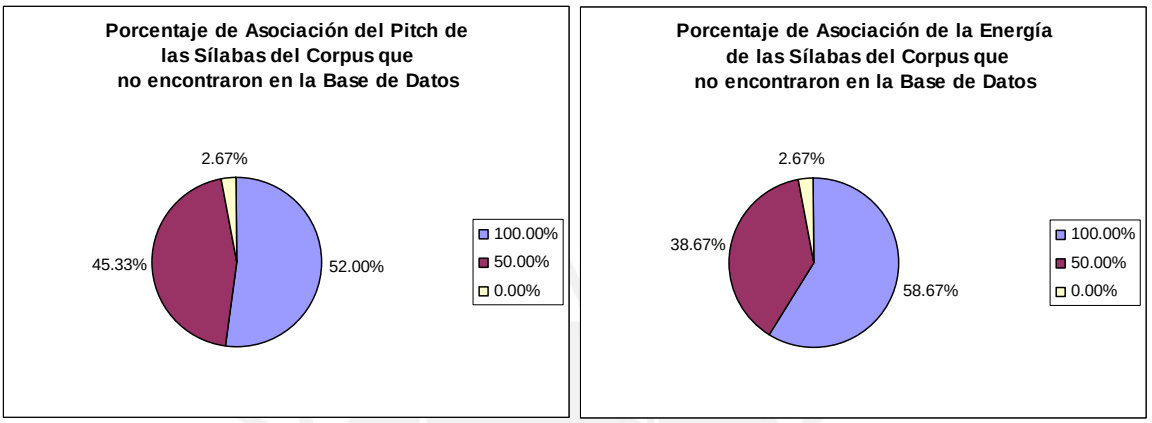

Figura N°14. Porcentaje de Identificación de Pitch y Energía de las Sílabas que no se encontraron en la Base de Datos

Para el caso de las duraciones, se hizo una comparación por cada oración del valor real versus el valor obtenido por el modelo, con ello se obtuvo que para la oración el valor total de la duración entregado por el modelo, para la mayoría de los casos, está en un rango del ± 10% de su valor real.

Finalmente, haciendo un análisis a nivel de oración, para las 25 oraciones que contiene el corpus de prueba, se puede apreciar que la máxima exactitud para representar la oración es de 84%, y la mínima obtenida es de 60%.

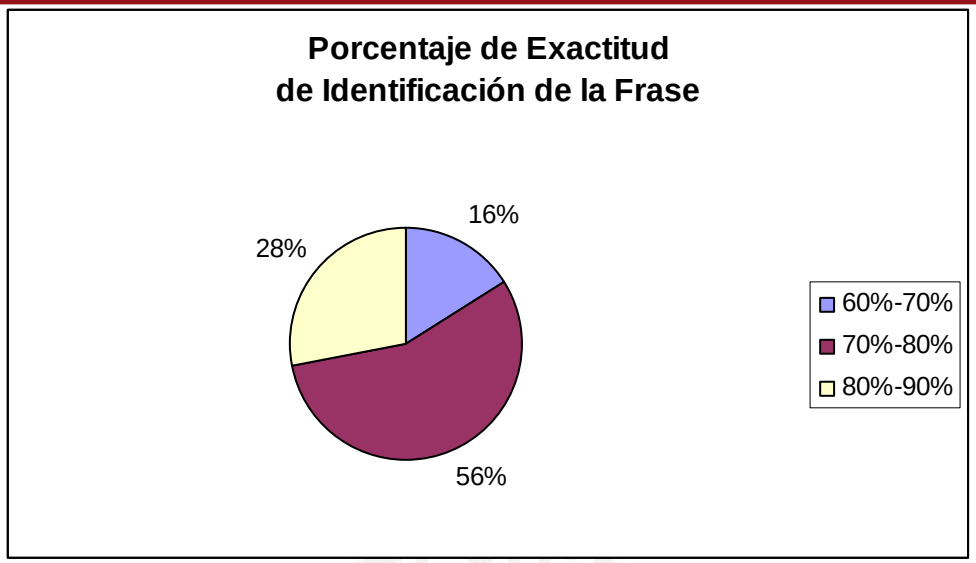

Figura N°15. Porcentaje de Exactitud de Modelado de Pitch y Energía para las frases del Corpus de Prueba

Por lo tanto, si tomamos un valor promedio de los porcentajes de exactitud hallados, obtenemos aproximadamente 75%, valor con el cual podríamos representar el correcto modelado de la entonación obtenido con la función *entona().*

4.3 Conclusiones

Para evaluar esta tesis, solo se considera el método cuantitativo debido a que el sintetizador por selección de unidades para el cual se está desarrollando el modelo de entonación en base a corpus aun está en desarrollo.

Los resultados calculados indican que el modelo de entonación en base a corpus implementado en esta tesis tiene un 75% de exactitud. Es decir, ese será el porcentaje con el cual una frase entregada será correctamente modelada.

Para conseguir mejores resultados, se tiene que contar con un corpus base de mayor cantidad de unidades, con lo cual los parámetros estándar modelarán mejor a las sílabas. Los valores de los parámetros entregados por la función *entona*() son sólo válidos si es que se quiere modelar la entonación de una persona, en este caso, de la locutora que leyó las frases.

Las tendencias por tramos de las curvas generadas a partir esos parámetros se mantiene pues es independiente del locutor, solo es dependiente del tipo de grupo acentual que se esté modelando.

CONCLUSIONES

1. Dado el incremento en el número de los espectadores de cine, es necesario brindar un sistema de información que satisfaga las necesidades de los usuarios.

2. La vía telefónica es la mejor forma que los usuarios han encontrado para solicitar información, sin embargo el volumen de llamadas que ingresa a la central telefónica puede ser tan alto que obstaculice el flujo de información. Las llamadas deben ser atendidas en el menor tiempo posible y el trato ofrecido al usuario debe ser atento y personalizado.

3. Ante la falta tecnológica vista en los servicios de atención al cliente que se ofrecen actualmente en el país, es necesario ingresar al mercado nuevas formas de atención, que con agilidad y calidad de trato sean capaces de atender los requerimientos de los usuarios.

4. Los sistemas de diálogo se presentan como buenas alternativas de automatización de atención al cliente vía telefónica. Por ello, en estos sistemas se debe tener especial cuidado con la calidad de voz a emitir. Para lograr naturalidad en la voz sintética se debe tener un adecuado modelo de entonación. El modelo de entonación con mejores resultados actualmente es el realizado en

base a un corpus, debido a que las unidades que conforman el corpus son unidades de entonación elegidas de acuerdo a la aplicación a realizar.

5. La naturalidad e inteligibilidad de las frases sintetizadas son evaluadas en el sintetizador de voz, mediante funciones definidas como costo de selección y concatenación, las cuales se busca minimizar.

6. Se debe contar con una persona preparada para leer el corpus del cual se obtendrán las unidades básicas, ya que al tener frases bien vocalizadas y con la entonación adecuada, resulta más sencillo encontrar los límites laterales de las unidades. Además de la preparación, es recomendable que el locutor encargado de la lectura del corpus no sea sometido a largas horas de grabación pues a medida que pasa el tiempo, su calidad vocal disminuye.

7. La segmentación de las unidades es una parte importante para el modelado de la entonación en base a corpus debido a que es a partir de ella que el etiquetado en función de parámetros básicos se realiza. Si la segmentación se realiza de forma manual por varias personas, es posible que para una misma unidad sus límites no coincidan exactamente, debido a que dicha delimitación depende de la percepción de cada persona.
8. Mientras más grande sea el corpus de trabajo, la asignación de parámetros estándar para cada tipo de grupo acentual es mucho más robusta. Para ello, no es sólo necesario contar con un corpus adecuado, sino también realizar una buena segmentación.

9. De contar con un programa que realice la segmentación de forma automática, la segmentación se realizaría de forma más rápida, por lo tanto, el estudio del corpus podría incluir más unidades.

10. Los modelos de entonación pueden ser evaluados de forma cualitativa y de forma cuantitativa. Para evaluar esta tesis, sólo se considera el método cuantitativo debido a que el sintetizador por selección de unidades aun está en desarrollo. Los resultados calculados indican que el modelo de entonación en base a corpus tiene un 75% de exactitud. Es decir, entregada una frase, ésta será identificada con sus parámetros correctos en dicho porcentaje.

11. Los valores de los parámetros entregados por la función *entona()* son sólo válidos si es que se quiere modelar la entonación de una persona, en este caso, de la locutora que leyó las frases, tanto del corpus base como del corpus de prueba. Sin embargo, las tendencias por tramos de las curvas generadas a partir esos parámetros se mantiene pues es independiente del locutor, sólo es dependiente del tipo de grupo acentual que se esté modelando.

RECOMENDACIONES

1. Para que el modelo de entonación en base a corpus sea más robusto y dé mejores resultados, el corpus base debe contener el mayor número de unidades posible. Es necesario que la aplicación en la cual el modelo será usado esté bien limitada. Con esto se logrará tener parámetros estándar más aproximados a los valores reales y el porcentaje de exactitud del modelo aumentará.

2. Para la grabación del corpus se debe contar con las condiciones adecuadas: equipos, ambiente y persona preparada para la lectura. En el caso del locutor, es recomendable que no sea sometido a largas jornadas de grabación, es decir, no grabar más de una hora seguida; el efecto del cansancio del locutor sobre la lectura no es percibido sino hasta el momento de realizar la segmentación en el cual al escuchar las grabaciones detenidamente sale a luz la omisión e inadecuada pronunciación de algunos sonidos. Esto podría ser una fuente de error al momento de establecer los parámetros estándar.

3. Las unidades de los corpus grabadas, deben ser cuidadosamente segmentadas y etiquetadas. Si la segmentación se realiza de forma manual por más de una persona, es necesario que esta actividad sea realizada bajo un mismo criterio para considerar, sobre todo, los límites de las unidades pues estas zonas son las más críticas durante la segmentación debido a que los sonidos de juntura suelen

confundirse, es decir, los límites de las sílabas tienen a capturar características de los sonidos adyacentes. Por ello, es recomendable el uso de programa de segmentación automática, así la tarea de segmentar se realizaría de forma mucha más efectiva y en menos tiempo, con lo cual daría la posibilidad de trabajar con un corpus de mayor número de unidades.

FUENTES

1. ADELL, Jordi y BONAFONTE, Antonio
   2004 Análisis de la Segmentación Automática de Fonemas para la Síntesis de Voz [en línea]
   <http://gps-tsc.upc.es/veu/research/pubs/download/ade_ana_04.pdf>

2. AMORÓS CÉSPEDES, Mari Cruz
   2005 Sincronización entre pico tonal y acento: resultados según posición métrica y morfológica [en línea]
   <http://www.iula.upf.edu/materials/050218cicres.pdf>

3. CALL CENTER MAGANIZE
   2005 Call Center Practices Around the World [en línea]
   <http://www.callcentermagazine.com/shared/article/showArticle.jhtml?articleId=165600383&classroom=>

4. CASTAÑEDA, Pablo
   2000 El lenguaje verbal del niño [en línea]
   <http://www.comunidadandina.org/bda/docs/PE-EDU-0003.pdf >

5. Center of Spoken Language Understanding – CSLU
   <http://cslu.cse.ogi.edu/>

6. CHIN-TENG, Lin et. al.
   2004 A novel prosodic-information synthesizer based on recurrent fuzzy neural network for the Chinese TTS system". Systems, Man and Cybernetics, Part B, IEEE Transactions. Pag.: 309- 324 Volume: 34, Issue: 1

7. CORDÓN, E. et.al.
   2002 Métodosubjetivo Para Evaluar La Entonación Sintética" [en línea]
   <http://bips.bi.ehu.es/ahoweb/files/publicaciones/URSI2002_N.pdf>

8. DONG WANG, Lie et. al.
   2002 Speech Segmentation without Speech Recongnition
   Departamento de Ingeniería Eléctrica y Electrónica de la Universidad de Tsinghua, Beijing

9. DRAGAN, Richard
   2004 Speak to Your Server [en línea]. PC Magazine Vol. 23, p58, 2p, 1c. EBSCO Academic Search Premier
   <http://search.epnet.com/login.aspx?direct=true&AuthType=cookie,ip,url,uid&db=aph&an=13278512&lang=es>

10. ESCUDERO MANCEBO, David
    2002 Modelado Estadístico de Entonación con Funciones de Bézier: Aplicaciones a la Conversión de Texto-Voz en Español. Tesis (Doc). Universidad de Valladolid. España. xx+ 183p

11. ELORRIETA, Gorka y ROMERA, Magdalena
    2004 Estudio Experimental De Las Unidades Prosódicas Del Discurso Y Sus Funciones [en línea]
    <http://www.ucm.es/info/circulo/no18/elorrome.htm>

12. FLORES TOSCANO, Leonardo
2001 Síntesis de Voz mediante la implementación del Unit Selection. Tesis (Ing) . Universidad de las Américas, Puebla, Mexico

13. FLORES, Leonardo et.al.
2001 Síntesis en Español Mexicano con el Método de Selección de Unidades de Longitud Variable [en línea]
<http://mailweb.udlap.mx/~ingrid/ingrid/articulo_33.pdf>

14. GARRIDO ALBIÑANA, Juan María
1996 Modelling Spanish Intonation for Text-to-Speech Applications. Tesis (Doc). Universidad Autónoma de Barcelona. España. 275p

15. GURLEKIAN, Jorge et.al.
2004 Modelos De Entonación Analítico Y Fonético-Fonológico Aplicados A Una Base De Datos Del Español De Buenos Aires [en línea]
<http://www.ub.es/labfon/XIII-15.pdf>

16. HUANG, Xuedong
2001 Spoken language processing: a guide to theory, algorithm, and system development, Upper Saddle River, NJ : Prentice Hall PTR, 980p

17. INSTITUTO NACIONAL DE CULTURA
Manifestaciones Artísticas Contemporáneas
<http://inc.perucultural.org.pe/patri5.shtml >

18. INSTITUTO NACIONAL DE ESTADÍSTICA E INFORMÁTICA
1996 Elasticidad de la Demanda de los Principales Bienes y Servicios Consumidos por las Familias en Lima Metropolitana [en línea]
<http://www.inei.gob.pe/biblioineipub/bancopub/Est/Lib0095/PRESEN.htm >

19. INSTITUTO NACIONAL DE ESTADÍSTICA E INFORMÁTICA
2001 Oferta y Demanda Global 1991 – 2000 [en línea]
<http://www.inei.gob.pe/biblioineipub/bancopub/Est/Lib0468/Libro.pdf>

20. INSTITUTO NACIONAL DE ESTADÍSTICA E INFORMÁTICA
2003 Oferta y Demanda Global 1991 - 2002 [en línea]
< http://www.inei.gob.pe/biblioineipub/bancopub/Est/Lib0556/Libro.pdf >

21. IMT - Instituto Mexicano de Telemarketing
¿Cómo evaluar aplicaciones de diálogo hablado en un Centro de Conacto? [en línea]
<www.imt.com.mx/recontact/37/evaluar.php>

22. LLISTERRI, Joaquim et.al.
1999 Fonética y Tecnologías del Habla [en línea]
<http://liceu.uab.es/~joaquim/publicacions/Fonetica_TecnolHabla.pdf>

23. LLISTERRI, Joaquim et.al.
2002 La conversión de texto en habla: aspectos lingüísticos [en línea]
<http://liceu.uab.es/~carme/CTH_FDS02.pdf>

24. LLISTERRI, Joaquim et.al.
2003 Entonación y tecnologías del habla [en línea]
<http://liceu.uab.es/~joaquim/publicacions/Ariel_Aplicaciones.pdf>

25. LOPE BLANCH, Juan M.
1993 Nuevos estudios de lingüística hispánica. México, D.F.: UNAM Instituto de Investigaciones Filológicas, 206 p. ; 23 cm.

26. LÓPEZ GONZALO, Eduardo
1993 Estudio De Técnicas De Procesado Lingüístico Y Acústico Para Sistemas De Conversión Texto-Voz En Español Basados En Concatenación De Unidades. Tesis (Doc). Universidad Politécnica de Madrid. España. 257p

27. LÓPEZ DE MÁNTARAS, Eduardo
2002 La Inteligencia Artificial Hoy: El referente de HAL en "2001 Una Odisea Espacial" [en línea]
<http://eia.udg.es/~blopez/dss/lectures/mantaras_mundocientifico.pdf>

28. MARIÑO ACEBAL, José
2002 Tecnologías del habla y el lenguaje para un asistente personal. Memoria Técnica [en línea]
<http://gps-tsc.upc.es/veu/aliado/memoriatecnica.pdf>

29. MINKER et. al.
2004 The SENECA spoken language dialogue system. Speech Communication [en línea] Vol. p89, 14p. EBSCO Academic Search Premier <http://search.epnet.com/login.aspx?direct=true&AuthType=cookie,ip,url,uid&db=aph&an=13878684&lang=es>

30. Massachusetts Institute of Technology (MIT) - SPOKEN LANGUAGE SYSTEMS GROUP
2003 Procesamiento de la información paralingüística. [en línea] <http://mit.ocw.universia.net/6.345/NR/rdonlyres/Electrical-Engineering-and-Computer-Science/6-345Automatic-Speech-RecognitionSpring2003/62BDE031-20B0-4058-872F-D6A3774567D6/0/lecture23.pdf>

31. Massachusetts Institute of Technology (MIT) - SPOKEN LANGUAGE SYSTEMS GROUP
Aplicaciones de los sistemas de diálogo [en línea]
<http://groups.csail.mit.edu/sls//applications/>

32. NATURAL VOX
Ejemplos de Sistemas que Usan Tecnologías del habla [en línea]
<http://www.natvox.es/clientes.html>

33. MUÑIZ GONZÁLEZ, Rafael
La atención al cliente en el siglo XXI. [en línea]
<http://www.marketing-xxi.com/los-call-centers-106.htm>

34. NAM SOO, Kim et. al.
2004 Discriminative training for concatenative speech synthesis.
Signal Processing Letters, IEEE Page(s): 40- 43 Volume: 11

35. PRASAD et al.
2004 Robots that can hear, understand and talk [en línea]
Advanced Robotics; Vol. 18 p533, 32p. EBSCO Academic Search Premier <http://search.epnet.com/login.aspx?direct=true&AuthType=cookie,ip,url,uid&db=aph&an=13269981&lang=es>

36. PRIETO, Pilar
2000 Revisión del Libro “La entonación del Español” - Sosa (1999) [en línea] <http://seneca.uab.es/pilarprieto/Review-Sosa.pdf>

37. SASSO, Len
2004 Voices from the Machine [en línea]. Electronic Musician; Vol. 20 p29, 6p, 1. EBSCO Academic Search Premier <http://search.epnet.com/login.aspx?direct=true&AuthType=cookie,ip,url,uid&db=aph&an=12443134&lang=es>

38. SPEECH TECHNOLOGY GROUP – DPTO OF ELECTRONIC ENGINEERING – TECHNICAL UNIVERSITY OF MADRID
2000 Estructura de un conversor de texto a voz [en línea] <http://lorien.die.upm.es/~juancho/pfcs/DPF/capitulo3.pdf>

39. SEGURA, Elí
2003 Modelado Prosódico-Lingüístico Para Un Sistema Conversor Texto A Voz Mediante Concatenación De Demisílabas. Tesis (Ing) Pontificia Universidad Católica del Perú. Perú. 152p

40. TECHNICAL UNIVERSITY OF MADRID - DPTO OF ELECTRONIC ENGINEERING - SPEECH TECHNOLOGY GROUP
2000 Sistemas de conversión de texto a voz. [en línea] <http://lorien.die.upm.es/~juancho/pfcs/AJP/cap1.pdf>

41. TODA et.al.
2004 Optimizing sub-cost functions for segment selection based on perceptual evaluations in concatenative speech synthesis. Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP '04). IEEE International Conference. Pag.: I - 657-60 vol.1 Volume: 1 ISSN: 1520-6149 Number of Pages: 5 vol. (cix+1045)

42. TORRE TOLEDANO, Doroteo
2001 Segmentación Y Etiquetado Fonéticos Automáticos. Tesis (Doc) Universidad Politécnica de Madrid. España.

43. UNIVERSIDAD DE OHIO – FACULTAD DE LINGÜÍSTICA
TOBI. [en línea] <http://www.ling.ohio-state.edu/~tobi>

44. UNIVERSIDAD DE SALAMANCA - DEPARTAMENTO DE LENGUA ESPAÑOLA
Glosario de Fonética [en línea] <http://web.usal.es/~joluin/comentariofilologico/glosariofonetica.pdf>

45. UNIVERSIDAD POLITÉCNICA DE CATALUNYA - GRUPO DE PROCESAMIENTO DE SEÑALES
2000 Operaciones en Centros de llamadas [en línea] <http://gps-tsc.upc.es/veu/basurde/propuestatecnica.pdf>

46. VILLARRUBIA GRANDE, Luis et.al.
2002 Tecnología del Habla para aplicaciones Multilingües, Multiservicio y Multiplataforma [en línea] <http://www.sec.upm.es/rsanse/pub%5Ctechmmm.doc>

47. WERNER et.al.
2004 Toward spontaneous speech Synthesis-utilizing language model information in TTS". Speech and Audio Processing, IEEE Transactions. Pag.: 436- 445 Volume: 12, Issue: 4.

48. (Autor Desconocido)
Características de Modelo IPO [en línea]
<http://www.ims.uni-stuttgart.de/projekte/mate/mdag/pd/pd_3.html>

49. (Autor Desconocido)
Explicación del Modelo IPO [en línea]
<http://liceu.uab.es/publicacions/MATED1.1.6Prosody/annex2/Ipo.html>

50. WAVESURFER – Página de descarga
<http://www.speech.kth.se/wavesurfer/download.html>

ANEXO N°1

ORACIONES DEL CORPUS BASE

1. La heroicidad está relacionada con la caridad.
2. La publicidad de los capulíes no me gustó.
3. La comicidad más graciosa está aquí.
4. La periodicidad de la onda no es constante.
5. Sé más.
6. El impulso que necesitas está en el jardín.
7. La mariposa voló alto.
8. Mariposa se fue a preparar ají.
9. El paralelismo de las rectas se comprobó.
10. Máximo esfuerzo debes realizar.
11. El árabe compró dos camellos.
12. El centésimo puesto fue ocupado por José.
13. Lo ví golpeándose contra la pared.
14. Te ví carcajeándote cuando te sirvieron el anís.
15. Sé árabe para poder ingresar al farol.
16. El puesto centésimo de los semánticos no fue cubierto.
17. Se fue a comprar el nuevo café máximo.
18. Para salir a divertirte está el sábado.
19. La lombriz juega en el túnel.
20. La aliteración consiste en sugerir imágenes con sentidos.
21. La caridad es una virtud que debemos cultivar.
22. Una aliteración combinada con paralelismo confundió al oyente.
23. Un onomástico se verá el semántico.
24. El versículo de la brújula tiene aguja.
25. El cañón tiene problemas
26. Ese señor es el que cobra.
27. El que cobra es ese señor.
28. Aquella estrella brillante es Venus.
29. Pedro compró esa corbata en París.
30. El gobierno ha desmentido el rumor.
31. Esa película dura tres horas.
32. Me gusta que actúe así
33. Que actúe así me gusta.
34. El ladrón saltó por la ventana.
35. Esa es la señora a quien visité
36. Esa es la puerta por lo cual se escapó
37. Le rogó que fuera con ellos
38. Nadie te creerá mintiendo así
39. Ese árbol es dónde chocó
40. No existe el número veintitrés.
41. La pancarta proclama justicia
42. El once es un número impar
43. Participar en deportes para tu salud
44. Empalme con esta ruta
45. Es capaz de decirlo todo
46. Pastillas para la gripe
47. Paisano déme una mano
48. Le hicieron una fiesta al paisano
49. El piano está apolillado
50. Le pegó al chico

ANEXO N°2

FORMATO PARA ALMACENAR INFORMACIÓN DE UNIDADES SEGMENTADAS Y ETIQUETADAS DEL CORPUS BASE

Sílaba	Número de Oracion	Posición de la sílaba en GA	Número de silabas del GA	Número de GA	Número Total de GA	F0 izq	F0 cen	F0 der	Energía izq	Energía cen	Energía der	Duración

Sílaba: sílaba analizada
Número de oración: identificador de la oración del corpus base
Posición de la sílaba en GA: posición de la sílaba dentro del grupo acentual
Número de silabas del GA: número total de sílabas del grupo acentual
Número de GA: número de grupo acentual dentro de la oración
Número Total de GA: número total de grupos acentúales de la oración
F0 izq: pitch en el extremo izquierdo de la sílaba
F0 cen: pitch en el centro de la sílaba
F0 der: pitch en el extremo derecho de la sílaba
Energía izq: energía en el extremo izquierdo de la sílaba
Energía cen: energía en el centro de la sílaba
Energía der: energía en el extremo derecho de la sílaba
Duración: duración (en segundos) de la sílaba

ANEXO N°3

PARAMETROS ASIGNADOS PARA GRUPOS ACENTUALES AGUDOS INICIALES

De una sílaba

Posición Sílaba	F0 izq (Hz)	F0 cen (Hz)	F0 der (Hz)	Energía izq (dB)	Energía cen (dB)	Energía der (dB)	Duración (s)
1	226	226	272	45	49	53	0.262

De dos sílabas

Posición Sílaba	F0 izq (Hz)	F0 cen (Hz)	F0 der (Hz)	Energía izq (dB)	Energía cen (dB)	Energía der (dB)	Duración (s)
1	235.0	211.0	205.0	47.0	60.0	34.0	0.140
2	187.0	256.0	209.0	33.0	59.0	43.0	0.125

De tres sílabas

Posición Sílaba	F0 izq (Hz)	F0 cen (Hz)	F0 der (Hz)	Energía izq (dB)	Energía cen (dB)	Energía der (dB)	Duración (s)
1	245	228	212	40	60	60	0.112
2	222	209	199	60	54	45	0.156
3	205	309	191	37	54	31	0.286

De cuatro sílabas

Posición Sílaba	F0 izq (Hz)	F0 cen (Hz)	F0 der (Hz)	Energía izq (dB)	Energía cen (dB)	Energía der (dB)	Duración (s)
1	257	216	155	25	66	12	0.153
2	201	194	186	49	60	42	0.118
3	173	184	169	42	44	37	0.115
4	190	226	286	46	62	22	0.160

De cinco sílabas

Posición Sílaba	F0 izq (Hz)	F0 cen (Hz)	F0 der (Hz)	Energía izq (dB)	Energía cen (dB)	Energía der (dB)	Duración (s)
1	269	221	198	32	60	58	0.1456
2	231	250	218	51	56	43	0.113
3	242	248	198	46	52	49	0.148
4	220	215	211	51	50	45	0.151
5	185	234	216	37	61	41	0.146

De seis sílabas

Posición Sílaba	F0 izq (Hz)	F0 cen (Hz)	F0 der (Hz)	Energía izq (dB)	Energía cen (dB)	Energía der (dB)	Duración (s)
1	245	216	207	13	59	55	0.142
2	228	211	192	57	61	57	0.083
3	194	197	83	47	53	22	0.123
4	84	192	174	22	55	42	0.110
5	190	188	193	51	54	43	0.129
6	175	237	166	47	58	30	0.221

ANEXO N°4

PARAMETROS ASIGNADOS PARA GRUPOS ACENTUALES AGUDOS CENTRALES

De una sílaba

Posición Sílaba	F0 izq (Hz)	F0 cen (Hz)	F0 der (Hz)	Energía izq (dB)	Energía cen (dB)	Energía der (dB)	Duración (s)
1	223	251	251	41	60	49	0.250

De dos sílabas

Posición Sílaba	F0 izq (Hz)	F0 cen (Hz)	F0 der (Hz)	Energía izq (dB)	Energía cen (dB)	Energía der (dB)	Duración (s)
1	199	198	151	56	46	29	0.205
2	285	229	210	55	59	44	0.106

De tres sílabas

Posición Sílaba	F0 izq (Hz)	F0 cen (Hz)	F0 der (Hz)	Energía izq (dB)	Energía cen (dB)	Energía der (dB)	Duración (s)
1	234	224	198	50	53	40	0.134
2	207	194	181	50	47	37	0.150
3	221	249	244	41	53	44	0.195

De cuatro sílabas

Posición Sílaba	F0 izq (Hz)	F0 cen (Hz)	F0 der (Hz)	Energía izq (dB)	Energía cen (dB)	Energía der (dB)	Duración (s)
1	264	274	252	44	50	26	0.119
2	237	250	205	28	51	32	0.161
3	163	190	166	43	49	35	0.151
4	194	199	191	44	50	39	0.177

De cinco sílabas

Posición Sílaba	F0 izq (Hz)	F0 cen (Hz)	F0 der (Hz)	Energía izq (dB)	Energía cen (dB)	Energía der (dB)	Duración (s)
1	248	270	234	35	54	36	0.107
2	247	216	73	45	52	31	0.112
3	211	193	181	44	48	41	0.176
4	169	182	168	41	48	32	0.129
5	195	224	221	40	59	44	0.159

De seis sílabas

Posición Sílaba	F0 izq (Hz)	F0 cen (Hz)	F0 der (Hz)	Energía izq (dB)	Energía cen (dB)	Energía der (dB)	Duración (s)
1	308	344	347	54	60	59	0.108
2	347	327	308	60	51	55	0.081
3	309	234	295	55	34	10	0.116
4	236	205	200	45	64	37	0.097
5	180	177	176	55	57	40	0.104
6	235	230	199	62	61	29	0.282

ANEXO N°5

PARAMETROS ASIGNADOS PARA GRUPOS ACENTUALES AGUDOS FINALES

De una sílaba

Posición Sílaba	F0 izq (Hz)	F0 cen (Hz)	F0 der (Hz)	Energía izq (dB)	Energía cen (dB)	Energía der (dB)	Duración (s)
1	243	141	120	51	46	35	0.451

De dos sílabas

Posición Sílaba	F0 izq (Hz)	F0 cen (Hz)	F0 der (Hz)	Energía izq (dB)	Energía cen (dB)	Energía der (dB)	Duración (s)
1	262	242	233	52	49	29	0.090
2	252	170	158	48	52	30	0.251

De tres sílabas

Posición Sílaba	F0 izq (Hz)	F0 cen (Hz)	F0 der (Hz)	Energía izq (dB)	Energía cen (dB)	Energía der (dB)	Duración (s)
1	260	242	241	47	51	47	0.107
2	274	263	229	49	48	39	0.114
3	242	193	163	45	46	23	0.291

De cuatro sílabas

Posición Sílaba	F0 izq (Hz)	F0 cen (Hz)	F0 der (Hz)	Energía izq (dB)	Energía cen (dB)	Energía der (dB)	Duración (s)
1	250	253	218	44	51	49	0.114
2	249	248	237	47	51	36	0.112
3	244	241	231	42	54	46	0.125
4	244	179	157	48	42	26	0.292

De cinco sílabas

Posición Sílaba	F0 izq (Hz)	F0 cen (Hz)	F0 der (Hz)	Energía izq (dB)	Energía cen (dB)	Energía der (dB)	Duración (s)
1	265	240	200	48	51	46	0.197
2	256	250	264	45	51	36	0.189
3	231	249	239	44	51	39	0.18
4	247	229	205	43	52	54	0.118
5	208	153	137	50	45	33	0.228

De seis sílabas

Posición Sílaba	F0 izq (Hz)	F0 cen (Hz)	F0 der (Hz)	Energía izq (dB)	Energía cen (dB)	Energía der (dB)	Duración (s)
1	241	284	312	43	52	32	0.139
2	172	206	184	50	52	31	0.096
3	187	165	163	40	42	30	0.19
4	170	200	185	45	46	42	0.089
5	206	214	196	47	49	28	0.138
6	201	216	93	43	45	29	0.18

ANEXO N°6

PARAMETROS ASIGNADOS PARA GRUPOS ACENTUALES GRAVES INICIALES

De una sílaba

Posición Sílaba	F0 izq (Hz)	F0 cen (Hz)	F0 der (Hz)	Energía izq (dB)	Energía cen (dB)	Energía der (dB)	Duración (s)
1	248	262	284	39	61	47	0.129

De dos sílabas

Posición Sílaba	F0 izq (Hz)	F0 cen (Hz)	F0 der (Hz)	Energía izq (dB)	Energía cen (dB)	Energía der (dB)	Duración (s)
1	217	207	157	27	56	21	0.139
2	198	266	200	42	55	35	0.214

De tres sílabas

Posición Sílaba	F0 izq (Hz)	F0 cen (Hz)	F0 der (Hz)	Energía izq (dB)	Energía cen (dB)	Energía der (dB)	Duración (s)
1	201	238	218	27	56	49	0.134
2	234	211	180	54	43	21	0.128
3	237	268	268	43	55	48	0.177

De cuatro sílabas

Posición Sílaba	F0 izq (Hz)	F0 cen (Hz)	F0 der (Hz)	Energía izq (dB)	Energía cen (dB)	Energía der (dB)	Duración (s)
1	106	208	187	37	55	50	0.155
2	185	179	170	45	56	43	0.161
3	171	157	141	45	37	22	0.083
4	206	206	205	24	46	42	0.117

De cinco sílabas

Posición Sílaba	F0 izq (Hz)	F0 cen (Hz)	F0 der (Hz)	Energía izq (dB)	Energía cen (dB)	Energía der (dB)	Duración (s)
1	126	221	173	57	53	22	0.163
2	200	198	196	45	63	36	0.079
3	196	197	189	37	61	47	0.102
4	189	185	183	46	54	50	0.105
5	185	207	210	49	52	47	0.253

De seis sílabas

No se considera dentro del análisis.

ANEXO N°7

PARAMETROS ASIGNADOS PARA GRUPOS ACENTUALES GRAVES CENTRALES

De una sílaba

Posición Sílaba	F0 izq (Hz)	F0 cen (Hz)	F0 der (Hz)	Energía izq (dB)	Energía cen (dB)	Energía der (dB)	Duración (s)
1	201	222	245	37	58	49	0.181

De dos sílabas

Posición Sílaba	F0 izq (Hz)	F0 cen (Hz)	F0 der (Hz)	Energía izq (dB)	Energía cen (dB)	Energía der (dB)	Duración (s)
1	234	218	184	47	50	41	0.179
2	213	226	248	44	49	47	0.218

De tres sílabas

Posición Sílaba	F0 izq (Hz)	F0 cen (Hz)	F0 der (Hz)	Energía izq (dB)	Energía cen (dB)	Energía der (dB)	Duración (s)
1	216	238	215	37	49	34	0.146
2	215	211	207	43	49	34	0.14
3	216	224	202	40	53	42	0.176

De cuatro sílabas

Posición Sílaba	F0 izq (Hz)	F0 cen (Hz)	F0 der (Hz)	Energía izq (dB)	Energía cen (dB)	Energía der (dB)	Duración (s)
1	210	207	207	42	52	41	0.114
2	205	202	180	45	50	45	0.108
3	201	172	163	41	42	27	0.167
4	226	260	246	44	56	41	0.159

De cinco sílabas

Posición Sílaba	F0 izq (Hz)	F0 cen (Hz)	F0 der (Hz)	Energía izq (dB)	Energía cen (dB)	Energía der (dB)	Duración (s)
1	220	290	280	52	51	42	0.160
2	221	220	211	56	57	43	0.111
3	211	191	188	37	51	43	0.148
4	180	170	175	45	47	38	0.160
5	222	257	268	44	54	44	0.182

De seis sílabas

Posición Sílaba	F0 izq (Hz)	F0 cen (Hz)	F0 der (Hz)	Energía izq (dB)	Energía cen (dB)	Energía der (dB)	Duración (s)
1	226	221	183	50	57	15	0.116
2	219	195	121	9	43	20	0.220
3	180	184	198	6	55	35	0.100
4	198	174	163	35	53	47	0.116
5	164	167	168	42	49	48	0.081
6	171	272	270	45	55	55	0.247

ANEXO N°8

PARAMETROS ASIGNADOS PARA GRUPOS ACENTUALES GRAVES FINALES

De una sílaba

No se considera dentro del análisis

De dos sílabas

Posición Sílaba	F0 izq (Hz)	F0 cen (Hz)	F0 der (Hz)	Energía izq (dB)	Energía cen (dB)	Energía der (dB)	Duración (s)
1	259	228	155	56	43	35	0.188
2	140	137	85	49	36	22	0.165

De tres sílabas

Posición Sílaba	F0 izq (Hz)	F0 cen (Hz)	F0 der (Hz)	Energía izq (dB)	Energía cen (dB)	Energía der (dB)	Duración (s)
1	185	213	208	20	52	34	0.140
2	225	240	145	35	56	28	0.249
3	204	150	168	40	32	8	0.190

De cuatro sílabas

Posición Sílaba	F0 izq (Hz)	F0 cen (Hz)	F0 der (Hz)	Energía izq (dB)	Energía cen (dB)	Energía der (dB)	Duración (s)
1	225	234	240	42	52	43	0.128
2	220	218	213	40	49	30	0.141
3	221	243	156	40	52	36	0.188
4	190	132	160	35	39	25	0.304

De cinco sílabas

Posición Sílaba	F0 izq (Hz)	F0 cen (Hz)	F0 der (Hz)	Energía izq (dB)	Energía cen (dB)	Energía der (dB)	Duración (s)
1	251	247	100	51	50	48	0.138
2	255	234	217	46	48	38	0.102
3	252	242	230	48	53	28	0.099
4	271	231	181	51	54	45	0.141
5	173	139	167	39	39	17	0.227

De seis sílabas

Posición Sílaba	F0 izq (Hz)	F0 cen (Hz)	F0 der (Hz)	Energía izq (dB)	Energía cen (dB)	Energía der (dB)	Duración (s)
1	281	296	271	45	50	50	0.144
2	239	263	212	43	52	53	0.173
3	181	292	269	40	56	51	0.160
4	236	229	224	57	50	38	0.200
5	246	221	155	51	52	37	0.186
6	113	119	163	37	36	25	0.233

ANEXO N°9

PARAMETROS ASIGNADOS PARA GRUPOS ACENTUALES ESDRÚJULOS INICIALES

De una sílaba

Posición Sílaba	F0 izq (Hz)	F0 cen (Hz)	F0 der (Hz)	Energía izq (dB)	Energía cen (dB)	Energía der (dB)	Duración (s)
1	199	250	281	14	63	28	0.207

De dos sílabas

Posición Sílaba	F0 izq (Hz)	F0 cen (Hz)	F0 der (Hz)	Energía izq (dB)	Energía cen (dB)	Energía der (dB)	Duración (s)
1	239	224	203	60	51	55	0.137
2	203	210	217	55	65	43	0.143

De tres sílabas

Posición Sílaba	F0 izq (Hz)	F0 cen (Hz)	F0 der (Hz)	Energía izq (dB)	Energía cen (dB)	Energía der (dB)	Duración (s)
1	261	247	218	52	58	42	0.156
2	183	232	218	41	54	40	0.172
3	190	181	184	48	54	45	0.137

De cuatro sílabas

Posición Sílaba	F0 izq (Hz)	F0 cen (Hz)	F0 der (Hz)	Energía izq (dB)	Energía cen (dB)	Energía der (dB)	Duración (s)
1	265	253	224	28	52	48	0.165
2	221	203	199	46	55	46	0.068
3	199	182	187	45	55	44	0.105
4	186	215	220	43	50	30	0.233

De cinco sílabas

No se considera dentro del análisis.

De seis sílabas

No se considera dentro del análisis.

ANEXO N°10

PARAMETROS ASIGNADOS PARA GRUPOS ACENTUALES ESDRÚJULOS CENTRALES

De una sílaba

Posición Sílaba	F0 izq (Hz)	F0 cen (Hz)	F0 der (Hz)	Energía izq (dB)	Energía cen (dB)	Energía der (dB)	Duración (s)
1	293	271	210	58	64	52	0.141

De dos sílabas

Posición Sílaba	F0 izq (Hz)	F0 cen (Hz)	F0 der (Hz)	Energía izq (dB)	Energía cen (dB)	Energía der (dB)	Duración (s)
1	221	194	196	55	50	43	0.066
2	199	233	282	43	60	49	0.194

De tres sílabas

Posición Sílaba	F0 izq (Hz)	F0 cen (Hz)	F0 der (Hz)	Energía izq (dB)	Energía cen (dB)	Energía der (dB)	Duración (s)
1	264	258	224	45	48	26	0.133
2	197	189	183	50	50	45	0.121
3	210	248	265	37	53	43	0.149

De cuatro sílabas

Posición Sílaba	F0 izq (Hz)	F0 cen (Hz)	F0 der (Hz)	Energía izq (dB)	Energía cen (dB)	Energía der (dB)	Duración (s)
1	215	214	161	52	63	38	0.163
2	185	184	185	7	55	45	0.118
3	280	280	276	45	45	62	0.147
4	306	296	288	57	54	51	0.118

De cinco sílabas

Posición Sílaba	F0 izq (Hz)	F0 cen (Hz)	F0 der (Hz)	Energía izq (dB)	Energía cen (dB)	Energía der (dB)	Duración (s)
1	204	205	223	24	46	47	0.086
2	218	213	189	48	54	51	0.088
3	192	195	193	39	56	53	0.084
4	196	186	149	43	56	45	0.117
5	167	230	257	30	39	40	0.194

De seis sílabas

Posición Sílaba	F0 izq (Hz)	F0 cen (Hz)	F0 der (Hz)	Energía izq (dB)	Energía cen (dB)	Energía der (dB)	Duración (s)
1	290	290	286	45	46	46	0.187
2	272	297	303	46	51	53	0.163
3	280	282	241	46	53	47	0.102
4	217	210	210	49	40	42	0.204
5	207	168	179	39	48	40	0.071
6	188	252	253	42	57	31	0.214

ANEXO N°11

PARAMETROS ASIGNADOS PARA GRUPOS ACENTUALES ESDRÚJULOS FINALES

De una sílaba
No se considera dentro del análisis.

De dos sílabas
No se considera dentro del análisis.

De tres sílabas

Posición Sílaba	F0 izq (Hz)	F0 cen (Hz)	F0 der (Hz)	Energía izq (dB)	Energía cen (dB)	Energía der (dB)	Duración (s)
1	242	214	210	48	57	4	0.223
2	155	155	157	40	57	44	0.179
3	156	119	167	40	32	9	0.213

De cuatro sílabas

Posición Sílaba	F0 izq (Hz)	F0 cen (Hz)	F0 der (Hz)	Energía izq (dB)	Energía cen (dB)	Energía der (dB)	Duración (s)
1	179	186	0	54	40	51	0.117
2	290	290	290	53	39	51	0.186
3	277	184	151	49	46	43	0.165
4	143	134	70	37	34	11	0.159

De cinco sílabas

Posición Sílaba	F0 izq (Hz)	F0 cen (Hz)	F0 der (Hz)	Energía izq (dB)	Energía cen (dB)	Energía der (dB)	Duración (s)
1	235	289	277	57	49	48	0.171
2	280	280	282	49	51	52	0.153
3	254	184	153	50	42	28	0.219
4	195	192	153	6	32	7	0.148
5	125	123	190	13	31	9	0.219

De seis sílabas
No se considera dentro del análisis.

ANEXO N°12

ORACIONES DEL CORPUS DE PRUEBA

1. Ese bolígrafo no es mío.
2. Fue atropellado por un camión.
3. Nunca te contaré la verdad.
4. Ese señor me avisó del peligro.
5. La policía defiende nuestra seguridad.
6. El público lanzó almohadillas al árbitro.
7. El arancel a pagar es alto.
8. Esa niña es alta.
9. Le pidió que viera por ellos.
10. El trabajo costoso fue preparar tu anís.
11. El dálmata loco muerde por gusto
12. El diácono se asocio con el diablo
13. Debes denunciar los malos manejos
14. El dentista recetó dentífrico
15. Perder así no es justo
16. Lo despidió de una patada
17. El edicto municipal no se cumple
18. El catedrático dictó una conferencia magistral
19. La escuadra edil volvió a la segunda división
20. Esa novela fue escrita a dos manos
21. Lo cogió del pescuezo
22. Impidió tu entrada
23. Cupido está enamorado
24. Escupir es una mala costumbre
25. Repón lo dañado

ANEXO N°13

FORMATO PARA ALMACENAR UNIDADES SEGMENTADAS Y ETIQUETADAS DEL CORPUS DE PRUEBA

Sílaba	Número de Oracion	Posición de la sílaba en GA	Número de silabas del GA	Número de GA	Número Total de GA	Tipo de GA	F0 izq	F0 cen	F0 der	Energía izq	Energía cen	Energía der	Duración	Tend F0 1erT	Tend F0 2doT	Tend Ener 1erT	Tend Ener 2doT

Sílaba: sílaba analizada
Número de Oración: identificador de la oración del corpus base
Posición de la sílaba en GA: posición de la sílaba dentro del grupo acentual
Número de silabas del GA: número total de sílabas del grupo acentual
Número de GA: número de grupo acentual dentro de la oración
Número Total de GA: número total de grupos acentúales de la oración
Tipo de GA: agudo, grave o esdrújulo
F0 izq: pitch en el extremo izquierdo de la sílaba
F0 cen: pitch en el centro de la sílaba
F0 der: pitch en el extremo derecho de la sílaba
Energía izq: energía en el extremo izquierdo de la sílaba
Energía cen: energía en el centro de la sílaba
Energía der: energía en el extremo derecho de la sílaba
Duración: duración (en segundos) de la sílaba
Tend F0 1erT: tendencia de los valores del pitch en entre el valor de la izquierda y centro
Tend F0 2doT: tendencia de los valores de pitch en entre el valor del centro y derecha.
Tend Ener 1erT: tendencia de los valores de energía en entre el valor de la izquierda y centro.
Tend Ener 2doT: tendencia de los valores de energía en entre el valor del centro y derecha.