UNIVERSIDAD CENTRAL "MARTA ABREU" DE LAS VILLAS
VERITATE SOLA NOBIS IMPONETUR VIRILIS TOGA. 1948

Facultad de Ingeniería Eléctrica

Centro de Estudios de Electrónica y Tecnologías de la Información.

TRABAJO DE DIPLOMA

Conformación de una Base de Datos de Voz Cubana para Reconocimiento del Locutor.

Autor: **Dania Heredia Ruíz.**

Tutor: **Dr. Julián Cárdenas Cabrera.**

Ing. Rogelio Silverio Martínez.

Santa Clara

2007-2008

"Año 50 de la Revolución"

Universidad Central “Marta Abreu” de Las Villas

Facultad de Ingeniería Eléctrica

Centro de Estudios de Electrónica y Tecnologías de la Información.

TRABAJO DE DIPLOMA

Conformación de una Base de Datos Cubana para Reconocimiento del Locutor.

Autor: **Dania Heredia Ruíz.**

Tutor: **Dr. Julián Cárdenas Cabrera.**

Profesor Auxiliar. CEETI.

Facultad de Ingeniería Eléctrica. UCLV.

E-mail: julian@uclv.edu.cu

Ing. Rogelio Silverio Martínez.

Santa Clara

2007-2008

“Año 50 de la Revolución"

Hago constar que el presente trabajo de diploma fue realizado en la Universidad Central "Marta Abreu" de Las Villas como parte de la culminación de estudios de la especialidad de Ingeniería en Telecomunicaciones, autorizando a que el mismo sea utilizado por la Institución, para los fines que estime conveniente, tanto de forma parcial como total y que además no podrá ser presentado en eventos, ni publicados sin autorización de la Universidad.

_____

Firma del Autor

Los abajo firmantes certificamos que el presente trabajo ha sido realizado según acuerdo de la dirección de nuestro centro y el mismo cumple con los requisitos que debe tener un trabajo de esta envergadura referido a la temática señalada.

_____ _____

Firma del Autor

Firma del Jefe de Departamento donde se defiende el trabajo

_____

Firma del Responsable de Información Científico-Técnica

PENSAMIENTO

Hay una fuerza motriz más poderosa que el vapor,
la electricidad y la energía atómica: La voluntad.

Albert Einstein.

DEDICATORIA

A mis padres
A mi hermana
A mi esposo
A mi hijo

AGRADECIMIENTOS

A: Mis padres, que tanto se han preocupado por educarme, inculcándome los mejores valores morales y guiándome por los mejores caminos de la vida, brindándome amor y confianza.

A: Mi hermana por su cariño y apoyo incondicional.

A: Mi esposo por inculcarme fuerza de voluntad para llegar a ser una profesional.

A: La profesora Maria Esperanza por dedicarme su tiempo y hacer posible la culminación exitosa de este trabajo.

A: Mis tutores Julián y Rogelio por el apoyo y guía en este trabajo.

A: La revolución por haberme dado la oportunidad de realizar mis sueños.

A: Todos los profesores que han contribuido a mi formación.

A: Mi hijo Alejandro por solamente existir.

A: Mis compañeros de curso Delvis, Maykel, Héctor, Orlando, Richar, Yassel, Osmany y Alberto que siempre los tengo presente.

A: Mi suegro por estar siempre dispuesto a brindarme su ayuda.

A: Baby por su paciencia y amabilidad en todos estos años de carrera.

A: Mi familia, amigos y compañeros de trabajo que han confiado en mi y me han apoyado en todo momento.

A todos muchas Gracias.

TAREA TÉCNICA

1. Estudio de las diferentes metodologías y técnicas aplicadas en la creación de las bases de voces más empleadas en el Reconocimiento del Locutor.
2. Estudio del universo de locutores que conformaran el banco de voces.
3. Conformación de la base de voces.
4. Confección y Presentación del informe.

Firma del Autor

Firma del Tutor

RESUMEN

El Reconocimiento del Locutor es el proceso de automáticamente reconocer a una persona, con base en la información presente en su voz. Con el desarrollo actual de las tecnologías de la informática y las comunicaciones, existen un sin número de aplicaciones como por ejemplo: control de acceso por líneas telefónicas a cuentas bancarias, bases de datos, computadoras remotas, dejando abierto el camino al surgimiento de nuevas aplicaciones. Estos sistemas han tenido un importante desarrollo en la última década, gran cantidad de investigadores han dedicado sus esfuerzos en crear sistemas cada vez mas robustos ante las posibles degradaciones de la voz en situaciones reales. Un rol fundamental en la prueba de dichos sistemas lo juegan los bancos de voces, que permiten desarrollar tales sistemas.

Este trabajo presenta la conformación de un banco de voces contentivas del español que se habla en Cuba, el cual cuenta con un universo de 39 locutores, 22 hombres y 17 mujeres en un rango de edades de 18 a 59 años, representando a las tres regiones de nuestro país. Para la creación del mismo se realizó un estudio de los principales bancos creados en el mundo, aplicables al reconocimiento automático del locutor, determinándose los parámetros necesarios según nuestras especificidades.

TABLA DE CONTENIDOS

INTRODUCCIÓN

El habla es, sin duda, el método de comunicación más natural, intuitiva y eficiente para los seres humanos. El intercambio de información mediante el habla juega un papel fundamental en nuestras vidas. Las estructuras lingüísticas y acústicas de la voz son reconocidas desde antaño como estrechamente relacionadas con la habilidad de los humanos para reconocernos unos a los otros. Por ello, no es de extrañar que durante décadas la idea de crear sistemas automáticos, capaces de reconocer a las personas por su voz haya fascinado a ingenieros y científicos.

El reconocimiento automático del Locutor (RAL) ha sido objeto de estudio durante varias décadas. Los esfuerzos y logros realizados por los investigadores en este campo durante estos años han llevado al desarrollo de diferentes aplicaciones. No obstante, hay que precisar que aún nos encontramos lejos de lograr un sistema capaz de reconocer a una persona por el habla natural en cualquier ambiente. El reconocimiento es, pues, un campo con gran potencial de desarrollo futuro.

Los estudios y trabajos realizados en el campo del procesamiento digital de señales y de manera particular en el procesamiento de voz, así como en todo lo relacionado con la producción de voz de forma natural, han permitido la realización de investigaciones que elevan la eficiencia y minimizan los errores de estos sistemas. Factor clave en esto es el desarrollo de potentes bases de datos de voz. La creación de estos con elevado número de locutores y en condiciones lo mas reales posibles, ha permitido grandes progresos, en los últimos diez años, en

las investigaciones relacionadas con el RAL. Los bancos de voces estándares han sido utilizados como forma de medir el estado del arte en el desarrollo de las investigaciones realizadas, permitiendo además determinar las principales deficiencias y dificultades, así como las líneas futuras de investigación.

Ejemplo de lo anterior son los concursos del Instituto Nacional de Estándares de la Tecnología (NIST) de E.U. Anualmente realiza una convocatoria (NIST Speaker Recognition Evaluation Plans, 2008) en la cual se presentan los sistemas que se encuentran en desarrollo en los diferentes países, a partir de una serie de pruebas se determinan las puntaciones y los ganadores. Este concurso se ha caracterizado por presentar cada año nuevos bancos de voces, los cuales tratan de asemejarse lo más posible a las condiciones reales de aplicación de los sistemas con vistas a comprobar su seguridad.

Actualmente existen bancos de voces para su uso de forma experimental en sistemas de identificación del locutor, KING-92, KING-SAM, YOHO, SWICHTBOARD, SPRIDE, TIMIT, AHUMADA (Campbell y Reynolds, 1997), son ejemplos de bases representativas de diversos tipos de habla, gran variedad de locutores y buenas condiciones de grabación. Pero estas bases además de no adecuarse a nuestras particularidades del habla, no nos son accesibles.

El problema afrontado en este trabajo consiste en la conformación de una base de datos de voz, que cumpla con los estándares internacionales y recoja las particularidades del habla cubana, para su utilización en los sistemas de reconocimiento de locutores por su voz.

Como aporte fundamental, este trabajo, propone la utilización de grabaciones de voces espontáneas a través de canales telefónicos, (con locutores masculinos y femeninos, con variedad de edades y varias sesiones de grabación de cada

locutor espaciadas en el tiempo), para la conformación de una base de datos cubana.

Objetivo:

Contribuir al desarrollo de los sistemas de reconocimiento del locutor a partir de la conformación de un banco de voces cubanas con la utilización de grabaciones espontáneas a través de canales telefónicos que pueda ser usado en la identificación de locutores.

Objetivos específicos:

1. Realizar un estudio sobre el estado del arte en la creación y organización de bancos de voces, para reconocimiento del locutor.

2. Selección y organización de una base de datos de voz espontánea útil en el reconocimiento de locutores.

3. Proponer como, a partir del banco de voces original, obtener otro con diferentes niveles de relación señal-ruido, que garanticen la aplicabilidad de la base de datos a diferentes escenarios de reconocimiento de locutores.

Hipótesis de Investigación

Se cuenta con una base de datos con voces cubanas que representa a ambos sexos de habla espontánea grabada a través de canales telefónicos de la red pública y con el etiquetado correspondiente.

Estructura del documento

Este documento está estructurado en Introducción, tres capítulos, conclusiones, recomendaciones y anexos.

El Capítulo 1: Importancia de las bases de datos de voz para el reconocimiento de locutores.

En este capítulo se hace una breve descripción de las dos grandes áreas que contemplan los sistemas de reconocimiento del locutor: la identificación y verificación. Significándose la importancia y necesidad de contar con bancos de voces lo más representativos y reales posibles, para el desarrollo y prueba de dichos sistemas. Se hace énfasis en la necesidad de contar, en el caso específico de nuestro país, con un banco de voces representativo del español que hablamos.

El Capítulo 2: Estudio de las características de las principales bases de datos existentes.

En este capítulo se describen los materiales y métodos utilizados para la elaboración de las bases para el reconocimiento y verificación de locutores. Se hace una breve descripción de las características específicas de varias bases de datos que se han utilizado como fuente para enriquecer los sistemas de reconocimiento del locutor

El Capítulo 3: Conformación de base de datos de voz cubana con independencia de texto.

Teniendo en cuenta lo estudiado de otras bases de datos, se realiza en este capítulo la conformación de una base de datos con un universo de 39 locutores, 22 hombres y 17 mujeres, con variedad de edades, con habla espontánea vía telefónica y tres sesiones de grabación por cada locutor espaciadas de dos semanas a tres meses, con un tiempo de duración de habla registrada no menor de un minuto. Se propone un método para elevar el desempeño de la base, a través de la incorporación a la misma de diferentes niveles de relación señal a ruido.

CAPÍTULO 1. IMPORTANCIA DE LAS BASES DE DATOS DE VOZ PARA RECONOCIMIENTO DE LOCUTORES.

Existen rasgos biométricos que permiten identificar a las personas, entre ellos se encuentran su huella dactilar, su mapa genético, fondo de la retina y la voz. Este último es el rasgo más variable y menos fiable de los cuatro, pero su disponibilidad es muy superior al resto, especialmente por vía telefónica, su captación es mucho menos invasiva, lo cual lo hace muy atractivo para utilizarlo como señal en diagnósticos médicos, sistemas de seguridad o sistemas de activación e interacción con medios de cómputo. Sin embargo la identificación mediante la voz debe ser analizada desde un punto de vista diferente, de forma análoga al reconocimiento facial o al análisis grafológico de la escritura, la variabilidad de la señal incorpora al proceso de reconocimiento, niveles adicionales de complejidad (Champod y Meuwly, 1998).

Gran cantidad de países en el mundo están enfrascados en el estudio e investigación de sistemas de reconocimiento de locutores. Desde hace algunos años se realizan concursos anuales dirigidos por el Instituto Nacional de Estándares y Tecnologías de Estados Unidos en los que participan los países que se encuentran a la vanguardia en el área de investigación de la voz como rasgo biométrico. En estos concursos se presentan los sistemas que se encuentran en desarrollo donde se evalúan diferentes parámetros que se tienen en cuenta para fortalecerlos.

Paralelo con el desarrollo de los sistemas de reconocimiento se trabaja en el diseño de nuevos bancos de voces. Estos han venido evolucionando desde pequeños bancos con pocos locutores, frases cortas leídas o cadenas de dígitos, hacia grandes bancos con varios minutos e incluso horas de grabaciones de habla espontánea a través de canales telefónicos.

El objetivo fundamental del desarrollo de los bancos de voces es que permitan realizar pruebas lo más reales posibles de los sistemas de reconocimiento de locutores por la voz.

1.1 LA GENERACIÓN DE LA VOZ, SU VALOR IDENTIFICATIVO

La voz es el efecto resultante de la actuación de tres componentes básicos: una fuerza propulsora, un elemento vibrante y un resonador. La fuerza es el flujo de aire procedente de los pulmones y que pasa a la tráquea; el elemento vibrante está formado por la faringe y las cuerdas vocales y el resonador comprende todas las cavidades supraglóticas: laringe, faringe, cavidad oral y cavidad nasal (Anexo I).

La voz se produce por la acción del flujo de aire que pasa por la glotis laríngea. Los músculos toráxicos y abdominales comprimen los pulmones y elevan la presión dentro del árbol respiratorio hasta alcanzar un nivel tal que provoca la apertura de la glotis. El flujo de aire que atraviesa la glotis se encuentra con las cuerdas vocales y las hace vibrar. A medida que la presión del flujo es mayor, la vibración también lo es y se traduce, en última instancia, en aumento de la intensidad sonora de la voz. La intensidad de la voz puede variar en un margen de hasta 100 dB si consideramos el abanico de posibilidades entre voz susurrada y un fuerte grito. En una conversación corriente se obtienen intensidades entre 40 y 50 dB.

La señal de voz porta varios niveles de información, primeramente el mensaje que se quiere transmitir a través de las palabras, pero además brinda información sobre el idioma, el estado emocional, el género y generalmente la identidad del hablante. Prueba de lo anterior es que los humanos somos capaces de identificar a familiares y personas conocidas, por su voz. Las características de la voz de un hablante vienen determinadas fundamentalmente por su fisiología, sus hábitos lingüísticos y por factores circunstanciales (Acero, 1993; Junqua y Haton, 1996; Campbell, 2003). Atendiendo a lo anterior los niveles identificativos de la voz se dividen en:

1 características de Bajo Nivel:
- **Segméntales:** Formantes, ancho de banda de los formantes, frecuencia fundamental.
- **Suprasegmentales:** Transición y ataque entre sonidos, coarticulación y concatenación.

2 características de Alto Nivel:
- **Nivel fonético:** entonación, acentuación, duración de los sonidos.
- **Nivel lingüístico:** ritmo, melodía, tiempo, jerga, léxico, reiteración de expresiones, variedad dialectal.

Los seres humanos reconocemos a los locutores combinando o fusionando los diferentes niveles identificativos.

La información de la identidad presente en la voz puede variar (Anexo II) debido a varios factores, como:

1 Factores Internos intrínsecos:
- **Permanentes:** sexo, edad, sesión, velocidad de articulación, tipo y cantidad de habla.

- **Transitorios:** estado emocional, patologías fonatorias.

2 Factores Internos forzados:

- **Efecto 'Lombard'** (voz en ambiente ruidoso)
- **Efecto 'cocktail- party'** (voz en voces concurrentes)

3 Factores Externos:

- **Canal (electro)acústico:** ruido acústico, reverberación, tipo de micrófono, distancia
- **Canal de comunicaciones:** ruido eléctrico, ancho de banda, margen dinámico, distorsión, codificación

El empleo de la voz como rasgo biométrico, da origen a los sistemas de reconocimiento de locutores, los cuales pueden ser divididos en Sistemas de Verificación de Locutores y Sistemas de Identificación de Locutores.

1.1.1 VERIFICACIÓN DE LOCUTORES

Este sistema comprueba la identidad del locutor que dice ser. Se necesita tener una referencia, plantillas o modelos del locutor (Bimbot, 2004, Furui, 1994). Siempre es necesario un umbral de decisión, siendo el error independiente del número de usuarios, dándose como resultado de la verificación la aceptación o rechazo del mismo.

Con la información de la voz captada microfónica o telefónicamente para la elaboración de las bases y la prueba, utilizando los métodos de cotejo existentes y conocidos, tanto dependientes como independientes del texto, se pueden establecer estas plantillas. Los sistemas de texto dependiente requieren que el usuario pronuncie una palabra o frase determinada por el sistema. Los sistemas de texto independiente están preparados para realizar el proceso de Verificación de locutores cualquiera que sea la palabra o frase pronunciada, se pueden distinguir dentro de cada uno de estos tipos de sistema aquellos de pronunciación

continua o los de palabra aislada. En estos últimos las palabras deberán estar separadas entre sí por pequeños instantes de silencio.

1.1.2 IDENTIFICACIÓN DE LOCUTORES.

Es el reconocimiento de una persona entre un conjunto de personas posibles (Campbell, 1997, Reynolds, 1995). La probabilidad de error crece con el número de personas a comparar. De cada locutor por los métodos de cotejo dependiente o independiente de texto, se elabora una referencia, plantilla o modelo. Entre la muestra de voz a prueba de cada locutor y los distintos modelos, se selecciona un máximo, dándose el locutor por identificado. En identificación se distingue entre un universo de población:

- **cerrado**: el locutor a prueba pertenece a uno de estos modelos.
- **abierto**: el locutor puede no pertenecer a ninguno de esos modelos.

Dentro de la identificación de locutores, algunas bases hacen frente a la problemática que conllevan las aplicaciones forenses relacionadas con el cotejo de voz. Se destaca la naturaleza no cooperativa del locutor implicado, grabaciones sin control en el proceso de adquisición, sistemas de reconocimiento con independencia de texto, grupos abiertos de locutores, condiciones del entorno muy variantes y ruidosas y en caso de error, el costo más que económico, es penal. A todos estos condicionamientos de la grabación forense se debe añadir que el entorno de grabación corresponde a una situación real.

1.2 FACTORES QUE AFECTAN EL DESEMPEÑO DE LOS SISTEMAS DE RECONOCIMIENTO DEL LOCUTOR.

En la evaluación de los sistemas de reconocimiento del locutor ha sido ampliamente reconocido que la cantidad y calidad de los datos usados, impacta directamente en el desempeño de los mismos, varios bancos de voces han sido desarrollados en años recientes con este fin (Campbell, 1999). Las evaluaciones preparadas dependen de numerosos factores en la selección y colección de los datos. A parte de los factores mencionados, en cuanto a las variaciones de la voz por las características propias del hablante, se unen los que dependen de particularidades tales como: las variaciones en el tipo de micrófono, el canal empleado en las grabaciones telefónicas o el desbalance provocado entre la fase de entrenamiento y la de prueba.

Previas evaluaciones del NIST (Przybocki y Martin, (1998, 2000, 2001); Doddington, G., et al, (2000)) han de mostrado que el desempeño de los sistemas de reconocimiento del locutor es grandemente enriquecido cuando el locutor utiliza el mismo tipo de micrófono en la fase de entrenamiento que en la de prueba. Por lo cual es altamente deseable que los bancos que se creen contengan gran cantidad de conversaciones donde se utilicen diferentes tipos de micrófonos. Ha sido demostrado además (Martin et al., 2005) que factores tales como: el tipo de micrófono telefónico, el tipo de canal (celular, inalámbrico o línea terrestre) y el lenguaje, son factores claves en el desempeño de dichos sistemas. En (Rossi, 1989) se clasifica la variabilidad de la voz, como factor que afecta el desempeño de los sistemas de reconocimiento del locutor, en: Variabilidad aleatoria, variabilidad de contexto, variabilidad intralocutor y variabilidad del lenguaje.

En estudios recientes utilizando el banco de voces CAVIS se han cuantificado los efectos de la calidad de la señal de voz, en el desempeño de dichos sistemas. En

(Nakasone, 2001, 2003) se ha demostrado que el mismo esta directamente relacionado, entre otros factores, con la duración de la señal y la relación señal a ruido de la grabación.

1.3 GRABACIÓN MULTISESIÓN

El registro o grabación de una base de datos tiene gran importancia describiendo la diferencia entre grabación monosesión y multisesión. Una base de datos monosesión de locutores consta de una única grabación bien sea vía microfónica o vía telefónica, donde todos los locutores han grabado en una sola vez todas las pruebas que se hayan creído oportunas, por parte del personal diseñador de la base de datos.

Es decir que evidentemente, el estado físico, psíquico o de otra índole que presente el locutor, estará en todas las grabaciones realizadas, debido a que sus condiciones no han variado nada o casi nada en el período de tiempo que pueda durar la sesión. En este caso los parámetros, que la voz de ese locutor presenta, son específicos de ese momento. Por lo que esto implica que no se pueda obtener más información.

Por ejemplo: si las sesiones de grabación están distanciadas en el tiempo, tratándose de pocos días o un mes, esta separación puede servir para que el locutor haya pasado por situaciones diferentes; así, un día podría estar anímicamente contento, en otra sesión podría estar estresado por problemas familiares, laborales o quizás por estar deprimido, es decir, en una base de datos multisesión, donde el tiempo que pasa de una grabación a la siguiente es corto, puede ser útil para encontrar características personales de la voz en cada locutor. También puede ser que efectivamente haya locutores cuyo estado psíquico sea muy similar en todas las sesiones, aunque si el número de éstas aumenta, la

posibilidad de encontrar al locutor en el mismo estado en todas las sesiones disminuye.

Supongamos que la distancia entre sesiones en el tiempo es ya más considerable, un año, dos, incluso muchísimo más tiempo; en este caso ya no serán los mismos parámetros los que variarán, sino que además de los anteriores habrá otros más, como por ejemplo la variación del tracto vocal.

En el caso de la multisesión, se pueden realizar grabaciones microfónicas y telefónicas distanciadas, es decir un determinado locutor puede grabar en los primeros días del mes la primera sesión microfónica y la primera sesión telefónica en los días finales de ese mismo mes. A pesar de pertenecer a la misma sesión, se podrían utilizar como sesiones diferentes, puesto que han transcurrido varios días de una toma a la otra.

1.4 LA NECESIDAD DE LA BASE DE DATOS CUBANA.

La utilización de bancos de voces estándares para el desarrollo y evaluación, ha probado ser muy importante en la promoción de los avances en las investigaciones relacionadas con el reconocimiento del habla y del locutor. El principal beneficio es que permite a los investigadores comparar los resultados de las diferentes técnicas en datos comunes, determinando cuales son las más promisorias. Adicionalmente, los bancos estándares pueden ser empleados para medir el desempeño de los sistemas que marcan el estado del arte actual en las áreas de reconocimiento del habla y del locutor.

En el mundo actual con el desarrollo de las tecnologías de la informática y las comunicaciones los bancos de voces han ido evolucionando desde pequeños en cuanto a cantidad de locutores y datos, hacia grandes bancos con cientos de

minutos de grabaciones de voz cada vez en condiciones más reales y espontáneas. Gran cantidad de países han invertido cuantiosos recursos en el área de investigación de la voz como rasgo biométrico, lo cual ha hecho posible la creación de una gran cantidad de bases de datos de voz en la mayoría de los idiomas más hablados en el mundo.

La necesidad de contar con una base de datos de voz del español que se habla en Cuba esta dada por dos razones fundamentales:

1 Estamos sometidos a un bloqueo económico por la potencia tecnológica más importante del mundo, lo cual nos impide el acceso a las principales bases de datos que se han desarrollado.

2 Aunque existen bancos de voces en idioma español, cada idioma es hablado con particularidades específicas en cada país, incluso dentro de un mismo país existen diferentes dialectos. Por ejemplo el inglés que se habla en Estados Unidos tiene 8 dialectos comprendido en las principales 8 regiones del país. En Cuba esta definido que el español se habla con características particulares en las tres regiones que esta dividida la isla (Occidental, Central y Oriental).

CONCLUSIONES PARCIALES.

En este capitulo se ha realizado una breve revisión de las principales características de la voz como rasgo biométrico, las técnicas y métodos a tener en cuenta en su captación y grabación como señal eléctrica, así como de la importancia de las bases de datos desarrolladas para aplicaciones tales como Reconocimiento del Locutor. Se resalta la importancia de contar con un banco de voces cubanas, que garantice el futuro desarrollo de sistemas e investigaciones propias en el área abordada.

CAPÍTULO 2. ESTUDIO DE LAS CARACTERÍSTICAS DE LAS PRINCIPALES BASES DE DATOS.

El uso de bases de datos de voz estándares para el desarrollo y evaluación, es uno de los factores principales en el progreso que han tenido los sistemas de reconocimiento automático del habla y del locutor. El principal beneficio radica en que los investigadores pueden comparar con datos comunes las diferentes técnicas y sistemas y por consiguiente determinar cuales son los más promisorios. Además permiten valorar los avances alcanzados y las áreas donde es preciso continuar investigando.

En este capítulo se realiza una valoración de las características y propósitos de las principales bases de datos de voz desarrolladas en diferentes países y de distintos idiomas. Los bancos descritos, fueron seleccionados a partir de la literatura pública disponible y de su aplicabilidad en los sistemas de reconocimiento del locutor. Los principales suministradores de estas bases son: el European Language Resources Association (ELRA) (http://www.icp.grenet.fr/ELRA/), el Linguistic Data Consortium (LDC) (http://www.ldc.upenn.edu/) y el Oregon Graduate Institute (OGI) (http://cslu.cse.ogi.edu/).

En particular nos centramos en cuatro aspectos fundamentales:

1. Cantidad y diversidad de hablantes.
2. Cantidad y separación en el tiempo de las muestras o grabaciones por hablante.

3. Tipo de habla (frase fija, cadena de dígitos, sentencias leídas o habla espontánea).
4. Medio de adquisición (canal telefónico, grabación microfónica de banda ancha, ambiente de estudio u oficina).

En la medida en que los bancos de voces reúnan una mayor cantidad y variabilidad de los aspectos anteriores, así será el grado de aplicación y utilidad.

El uso de un banco de voces para experimentación o evaluación de un sistema de reconocimiento del locutor requiere la definición de un procedimiento de evaluación, que especifique entre otras cosas: la partición del banco en la parte de entrenamiento y la parte de prueba, la prueba del sistema para condiciones específicas como por ejemplo, el desempeño cuando el entrenamiento y la prueba son hechos con grabaciones a través de diferentes micrófonos, requiere que el banco contenga suficientes grabaciones con dichas características. El procedimiento antes mencionado no es objetivo del presente trabajo, no obstante se puede profundizar en el tema en (Doddington, 1998 y NIST Speaker Recognition Evaluation Plans)

A partir del año 1987 con la creación de la base de datos KING por la ITT, bajo un contrato del gobierno de EU y diseñada principalmente para experimentos cerrados de identificación de hablantes, surgieron otras bases en las que se tuvieron en cuenta parámetros internacionales que las hacen aplicables a los sistemas de reconocimiento de locutores.

2.1 BASE DE DATOS KING-92

KING-92 es una base de datos derivada de la KING, se comenzó a desarrollar en el año 1992 por el LDC, e intervino la ITT. Tomaron muestras solamente de

hombres de edades similares y procedencia geográfica muy cercana. Se diferencia de la base original en los detalles que se especifican a continuación:

- Los datos originales se muestrearon a 10 Khz., siendo ahora a 8 Khz.
- Se ha producido un alineamiento mejor de los tramos de voz de la sesión microfónica y telefónica.
- Originariamente existía una trascripción ortográfica y fonética de los datos almacenados. Esto suponía gran cantidad de errores. Ahora sólo existe la ortográfica.

Esta base contiene grabaciones de 51 locutores hombres en dos versiones, una a través de canal telefónico y otra a través de un micrófono de alta calidad. Los locutores fueron divididos en dos grupos uno de 25 y otro de 26, haciéndose las grabaciones de cada grupo en dos locales diferentes. Por cada locutor y canal hay 10 ficheros correspondientes a sesiones de entre 10 y 60 segundos de duración cada una.

Tabla 1: Características Generales de la base de datos King-92.

# de locutores	51 (todos hombres)
# sesiones/locutor	10
Intervalo entre sesiones	Una semana a un mes
Tipo de habla	Descripciones de fotografías
Micrófono	De banda ancha y de teléfonos (electret)
Canales	Banda base y Red telefónica pública
Ambiente acústico	Controlado

2.2 BASE DE DATOS KING-SAM.

La KING-SAM se creó por el grupo del habla del NIST en 1993. Diseñada para soportar investigaciones en la detección automática de cambio de locutor o "monitoreo". Es decir, debe muestrear continuamente y automáticamente información de una voz conocida, por un canal y detectar cuándo ha cambiado. Esto tiene su extensión natural en la tecnología que emplea verificación de voz para autorizar transacciones financieras por teléfono.

Cada CD elaborado posee segmentos empalmados que contienen partes de los dos canales empleados en la KING. La mitad de los segmentos engloba diferentes sesiones de cada locutor y la otra mitad, sesiones de diferentes locutores. De este emparejamiento resultan 796 segmentos con una media de 40 segundos cada uno. Así se construyen alrededor de 10 horas de datos muestreados a 8 Khz. y con 16 bits por muestra. El punto de empalme de los segmentos es siempre en la ubicación de un silencio de corta duración.

Esta base de datos es una creación artificial, diseñada para permitir un monitoreado automático de locutores a partir de datos manejables con características bien conocidas y estructurados para su fácil corte y evaluación.

Tabla 2: Características Generales de la base de datos King-SAM

# de locutores	51 (todos hombres)
# sesiones/locutor	10
Intervalo entre sesiones	Una semana a un mes
Tipo de habla	Descripciones de fotografías

Micrófono	De banda ancha y de teléfonos (electret)
Canales	Banda base y Red telefónica pública
Ambiente acústico	Controlado

2.3 BASE DE DATOS YOHO

YOHO surge con el objetivo de que sirva como marco de comparación entre sistemas de verificación de locutor y alentar la competición entre grupos de investigación. En (Campbell, 1995) se sugiere incluso un protocolo de pruebas a seguir.

La base de datos fue adquirida en un entorno de oficina, en condiciones muy controladas y bajo ruido. Se utilizó un auricular telefónico de alta calidad para capturar la voz. Está compuesta por 138 locutores, de los cuales 108 son hombres y 30 mujeres. El idioma de la base de datos es inglés americano, siendo la mayoría de locutores de la zona de Nueva York. Sin embargo, también se incluyen algunos locutores no nativos. La base de datos se compone de datos de entrenamiento y de verificación. El entrenamiento se divide a su vez en 4 sesiones de 24 frases cada una. Los datos de verificación se adquirieron en 10 sesiones espaciadas una media de 3 días entre ellas y con 4 frases por sesión. En un escenario dependiente de texto, las frases son conocidas por el sistema, es decir, al usuario se le pide que las diga. La sintaxis utilizada en la base de datos es únicamente secuencias de tres pares de dígitos, a modo de contraseña de un candado de combinación.

Todas las sesiones se efectuaron con el mismo teléfono y las grabaciones no usan líneas comerciales o públicas. Los datos son muestreados a 8 Khz. con 12 bits de resolución.

Tabla 3: Características Generales de la base de datos YOHO.

# de locutores	138 (108 H/30 M)
# sesiones/locutor	4 entrenamiento, 10 verificación
Intervalo entre sesiones	Días-mes (3 días nominal)
Tipo de habla	Dígitos y frases
Micrófono	Telefónico de alta calidad
Canales	3.8KHz
Ambiente acústico	Oficina

2.4 BASE DE DATOS SWITCHBOARD

Es una extensa base de datos de conversaciones telefónicas. Su diseño engloba diversos tipos de habla, gran cantidad de locutores y buenas condiciones de grabación telefónica. SWITCHBOARD es una rica fuente para experimentar la verificación e identificación de locutores, ha constituido la fuente para la elaboración de los banco de los concursos anuales del NIST.

Elaborada por Texas Instruments, cuenta con dos bancos (Switchboard I y II) incluye más de 500 locutores de ambos sexos en el principal dialecto del inglés americano. Esto supone un volumen de 2430 conversaciones de un promedio de 6 minutos. Es decir, unas 240 horas de habla registrada y transcrita. Los datos han sido muestreados a 8 Khz. con 8 bits por muestra tras aplicar la codificación **"**μ**"** y con dos canales de audio por cada fichero de voz.

Se ha registrado sin la intervención humana, realizándose de forma totalmente automática. La automatización previene así contra la intrusión del experimentador y garantiza un grado de uniformidad a lo largo de toda la toma. Las relaciones transcritas indican que las conversaciones registradas son altamente naturales.

Los datos relativos al origen demográfico, así como las fechas de nacimiento y variedad intersesión de cada locutor, se registran en tablas. Un estudio de estos valores, sin especificar la identidad personal del locutor, es incluido en la base de datos. Entre estos datos se encuentra información relevante para estudiar la voz, el dialecto y otros aspectos del estilo del habla. También se incluyen las distribuciones estadísticas por edad, sexo, educación, residencia habitual y lugar de estancia durante sus años de formación. Es proporcionado igualmente el tiempo exacto y el código del área de origen de cada llamada telefónica efectuada, así como los distintos aparatos empleados.

SWITCHBOARD, debido a su gran tamaño y tipo de información telefónica, permite investigaciones sobre muchos tipos de experimentos, tanto abiertos como cerrados. El tener dos canales de audio por cada toma de voz, la hace viable para estudios de detección de cambio de locutor o “monitoreado”. Permite, asimismo, por el gran número de voces grabadas y la alta calidad de éstas, realizar estudios en profundidad de las características de la voz.

Esta base tiene ciertas limitaciones. La gran cantidad de datos (240 horas de voz o 12 gigabytes) es un obstáculo para muchos investigadores. Posee una limitada variabilidad de aparatos telefónicos por locutor, unido a los efectos de cancelación, debido a la alta calidad atípica del canal. De los teléfonos se conoce su número, de tal forma que, de cada grabación únicamente se tenga la evidencia de que

procede del aparato asociado a ese número, sin tener la seguridad de si el teléfono usado es el mismo de unas conversaciones a las siguientes.

Tabla 4: Características Generales de la base de datos Switchboard

# de locutores	543 y 657 (~50% H/50% M)
# sesiones/locutor	1-25 (5 minutos de conversación)
Intervalo entre sesiones	Días-Semanas
Tipo de habla	Conversacional
Micrófono	Telefónico variable
Canales	Red telefónica pública
Ambiente acústico	Casa/Oficina

2.5 BASE DE DATOS SPIDRE.

Es una base de datos derivada de SWITCHBOARD, con un tamaño de datos más reducido. Consta con más variabilidad de los aparatos telefónicos, siendo este el principal aporte respecto a la base de la cual procede. Si bien, al igual que en aquella, los aparatos no están documentados, sabiendo únicamente el número de teléfono.

A pesar de su pequeño tamaño (2 CD-ROM, con 1,2 gigabytes de datos), esta base permite efectuar experimentos abiertos y cerrados para sistemas de verificación de locutores, así como el monitoreado de éstos (detección de cambio

de locutor) a una menor escala que la SWITCHBOARD, pero no tan reducida como la KING.

La transcripción y alineamiento temporal de los ficheros permite la localización automática del habla de cada locutor. Los resultados obtenidos en verificación de locutores varían considerablemente según sean tomas de voz del mismo locutor con el mismo aparato telefónico, o bien, desde distintos aparatos.

Tabla 5: Características Generales de la base de datos SPIDRE.

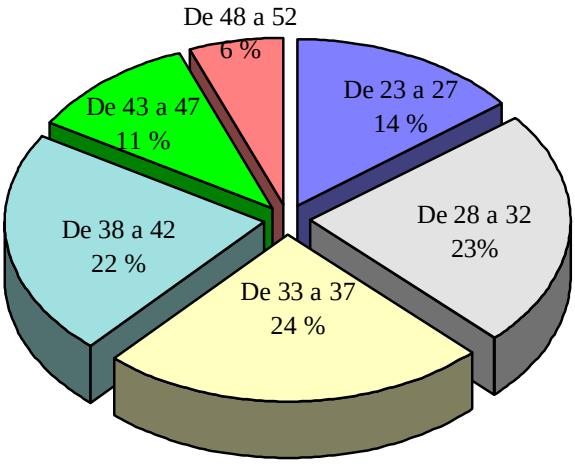
# de locutores	332
# sesiones/locutor	1-4 (5 minutos de conversación)
Intervalo entre sesiones	Días-Semanas
Tipo de habla	Conversacional
Micrófono	Telefónico variable
Canales	Red telefónica pública
Ambiente acústico	Casa/Oficina

2.6 BASE DE DATOS AHUMADA

AHUMADA es una de las Bases de Datos más extensas, en idioma Español, ha sido muy útil para evaluar y robustecer los sistemas de reconocimiento de locutores por sus voces. Ha participado en los concursos del NIST en los que ha brindado grandes aportes al desarrollo de investigaciones (Ortega, 2000).

Esta base principalmente está diseñada sobre grabaciones forenses con fines identificativos. Es muy amplia, cuenta con un universo de 103 locutores varones de diferentes edades de entre 23 y 52 años por lo que se realizó un estudio para conocer la distribución de la población utilizada.

La figura 1 representa la distribución de la población utilizada en la base de datos en cuanto a edades se refiere. Realizaron las distribuciones agrupando a los locutores en subdivisiones de 5 años, empezando por la edad del más joven.



*** Desviación standard = 7.**
*** Media = 35.63 años.**

Figura 1: Distribución por edades.

AHUMADA es una base multisesión en la que se realizaron tres tomas por locutor. Se grabaron por medios microfónicos y telefónicos y dentro de éstos últimos, una toma por línea interna y las otras dos por la línea pública telefónica. Los distintos aparatos telefónicos utilizados fueron numerados secuencialmente.

Se realizaron grabaciones de series de dígitos aislados, cadenas de dígitos, frases fonológicamente equilibradas de 8 a 10 palabras aproximadamente, textos propios extraídos de un libro y habla espontánea aproximadamente de un minuto y medio.

Tabla 6: Características Generales de la base de datos AHUMADA

# de locutores	103 H
# sesiones/locutor	3
Intervalo entre sesiones	Semanas
Tipo de habla	Cadena de dígitos, Texto leído y Espontáneo
Micrófono	Micrófonos de alta calidad y Telefónico variable
Canales	Red telefónica pública
Ambiente acústico	Casa/Oficina

2.7 BASES DE DATOS POLYCOST.

Este banco de voces fue desarrollado en el proyecto europeo COST-250 (Petrovska, D., et al, 1998), con el propósito de probar los sistemas de verificación de locutores. Incluye voz de hablantes, nativos y no nativos del idioma inglés, de 134 locutores (74 hombres y 60 mujeres), lo cual la hace útil además para la experimentación en tareas tales como reconocimiento del lenguaje y del acento. Fue grabada en cinco sesiones espaciadas semanalmente, a través de líneas

telefónicas digitales (ISDN) con auriculares variables, en las oficinas o casas de los participantes.

Tabla 7: Características Generales de la base de datos POLYCOST

# de locutores	134 (74H, 60M)
# sesiones/locutor	5
Intervalo entre sesiones	Semanas
Tipo de habla	Cadena de dígitos, Texto leído y Espontáneo
Micrófono	Telefónico variable
Canales	Red telefónica pública digital
Ambiente acústico	Casa/Oficina

2.8 BASE DE DATOS TIMIT/NTIMIT

TIMIT (Texas Instruments Massachusetts Institute of Technology) ha sido resultado de un conjunto de esfuerzos del proyecto de investigación, ciencias de la información, de oficinas de tecnologías (DARPA-ISTO), también esfuerzo de la unión entre el Instituto de Tecnología Massachussets (MIT), Instituto de investigaciones de Stanford (SRI) y Texas Instruments (TI).

La base de datos consta de 630 locutores que representan los 8 dialectos más importantes de inglés americano. De los 630 locutores, el 70% son hombres (438)

y el 30% restante son mujeres (192). Cada locutor lee un total de 10 frases fonéticamente balanceadas, divididas en entrenamiento y prueba. Para las pruebas se reservan 24 locutores, 2 hombres y una mujer de cada región dialectal.

Tabla 8: Características Generales de la base de datos TIMIT.

# de locutores	630 (438H/192M)
# sesiones/locutor	1
Intervalo entre sesiones	Ninguno
Tipo de habla	Texto leído
Micrófono	Micrófonos de alta calidad
Canales	
Ambiente acústico	Controlado

NTIMIT es la misma base TIMIT, pero las grabaciones furon pasadas a través de llamadas telefónicas locales o de larga distancia. Cada sentencia fue directamente acoplada a un micrófono telefónico de carbón, por lo cual la base NTIMIT se puede considerar como la misma TIMIT degradada debido a los transductores (micrófono), y a las condiciones de las líneas o canal telefónico.

Tabla 9: Características Generales de la base de datos NTIMIT.

# de locutores	630 (438H/192M)
# sesiones/locutor	1
Intervalo entre sesiones	Ninguno
Tipo de habla	Texto leído
Micrófono	Carbón (Telefónico)
Canales	Red telefónica pública
Ambiente acústico	Controlado

CONCLUSIONES PARCIALES

En este capítulo se realizó un estudio de los principales aspectos tenidos en cuenta en la elaboración de las diferentes bases de voces orientadas al reconocimiento del locutor. Se determinó que aspectos tales como: cantidad de locutores, distribución por edades y sexo, procedencia geográfica, espaciamiento en el tiempo y duración de las muestras obtenidas son claves en la creación de los bancos de voces.

CAPÍTULO 3. CONFORMACIÓN DE UNA BASE DE DATOS DE VOZ CUBANA CON INDEPENDENCIA DE TEXTO.

En este capítulo se realiza la conformación de una base de datos cubana con un universo de 39 locutores, de ambos sexos, con variedad de edades, utilizando habla espontánea vía telefónica, con tres sesiones de grabación por cada locutor espaciadas en el tiempo, de dos semanas a tres meses, con un tiempo de duración de habla registrada no menor de un minuto. Estas particularidades hacen que esta sea aplicable a los sistemas de reconocimiento y verificación de locutores por sus voces.

3.1 CONFORMACIÓN DE LA BASE DE DATOS CUBANA.

Se cuenta con conversaciones a las que se le realizó un trabajo previo de selección, edición y etiquetado para su uso en la base de datos, esto se ejecutó empleando el software SoundForge versión 8.0.

Se seleccionaron 39 locutores diferentes, cada uno de ellos con al menos un minuto y medio de conversación. Las grabaciones utilizadas se han obtenido de conversaciones espontáneas, desarrolladas sobre líneas telefónicas. Cada fichero contiene voz de dos interlocutores, es decir los dos lados de la conversación por lo que se editaron los segmentos de voz de un solo interlocutor formando un fichero con toda la información del locutor seleccionado. En esto se tuvieron en cuenta los

segmentos con mayor energía, contenido fonético y menores intervalos de silencio.

Para homogenizar las pruebas se ha limitado la duración de todos los segmentos de voz de los distintos locutores a 1 minuto, desechando el resto. La voz ha sido capturada a través de distintos terminales, a lo que hay que añadir la variabilidad de estados de ánimo y el alto grado de espontaneidad que presentan los locutores en las grabaciones.

3.1.1 DISEÑO DE LA BASE DE DATOS

Teniendo en cuenta el análisis realizado a las bases de datos plasmadas en el capítulo 2, se adecuaron a nuestras características los aspectos y parámetros más importantes, de forma tal que favorecieran la realización de la misma.

Esta base dentro de sus particularidades cuenta con:

- 39 locutores.
- Ambos sexos (22 hombres y 17 mujeres).
- Rango de edades de 18 a 59 años
- Tres sesiones de grabación.
- Duración de las llamadas de al menos un minuto.
- Grabaciones espontáneas.
- Por teléfono.
- Idioma Español característico de Cuba.

3.1.2 EL UNIVERSO DE LOCUTORES.

A diferencia de otras bases, como por ejemplo Ahumada (donde todos los locutores son hombres), esta base cuenta, como se puede apreciar en la tabla 10 y figuras 2 y 3, con un universo de locutores de ambos sexos, distribuidos según

las tres regiones principales del país (oriente, centro y occidente), donde el español hablado en Cuba reúne particularidades específicas en cuanto a: empleo de palabras y frases, entonación y ritmo del habla, entre otras.

Tabla 10: Distribución de locutores según región y sexo.

REGION	CANT. HOMBRES	CANT MUJERES	TOTAL
Occidente	4	2	6
Centro	13	9	22
Oriente	5	6	11
TOTAL	**22**	**17**	**39**

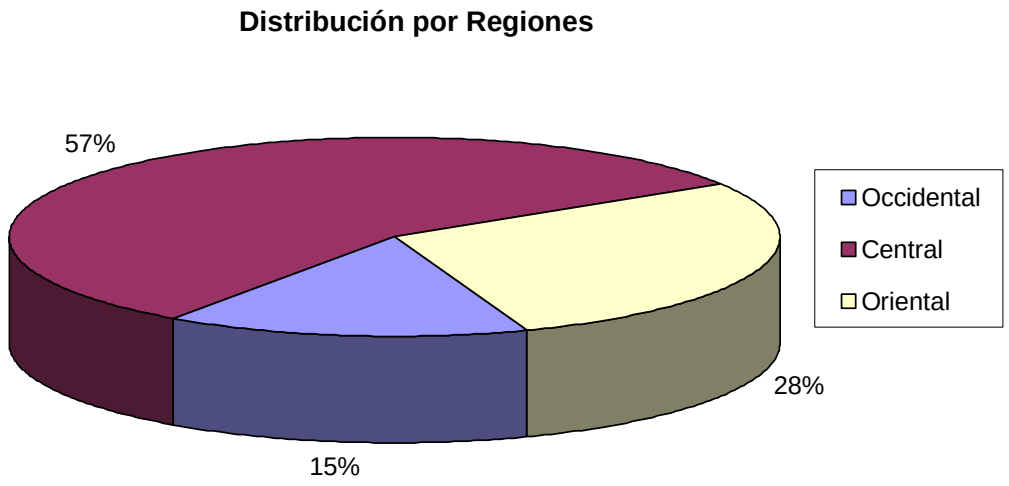


Figura 2: Distribución por Regiones.

Como se puede apreciar de la tabla 10 y figura 2, la región central es la más representada, lo cual se debe a que el mayor número de grabaciones con que se contó fue de la misma. Esto implica que es necesario para futuros trabajos elevar la representatividad de las dos restantes regiones, con vistas ha que la base permita realizar experimentos y pruebas relacionadas con el dialecto.

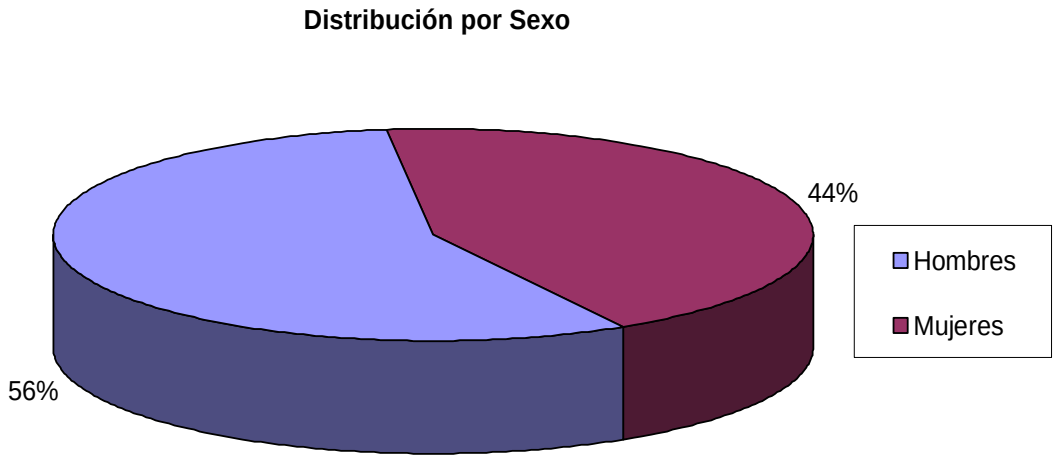


Figura 3: Distribución por Sexo

3.1.3 ESTUDIO POR EDADES:

En el universo de locutores de esta base se aprecia una distribución por edades desde los 18 años hasta los 59. Donde se agruparon a los locutores en subdivisiones de 10 años empezando por la edad del más joven. Con una media de 34, 28 años. La mayor representatividad en cuanto a las edades se encuentra entre los 18 a 39 años, esto se debe a que en este rango existe mayor

representatividad de la fuerza laboral, hacia la cual van dirigidas las principales aplicaciones de los sistemas.

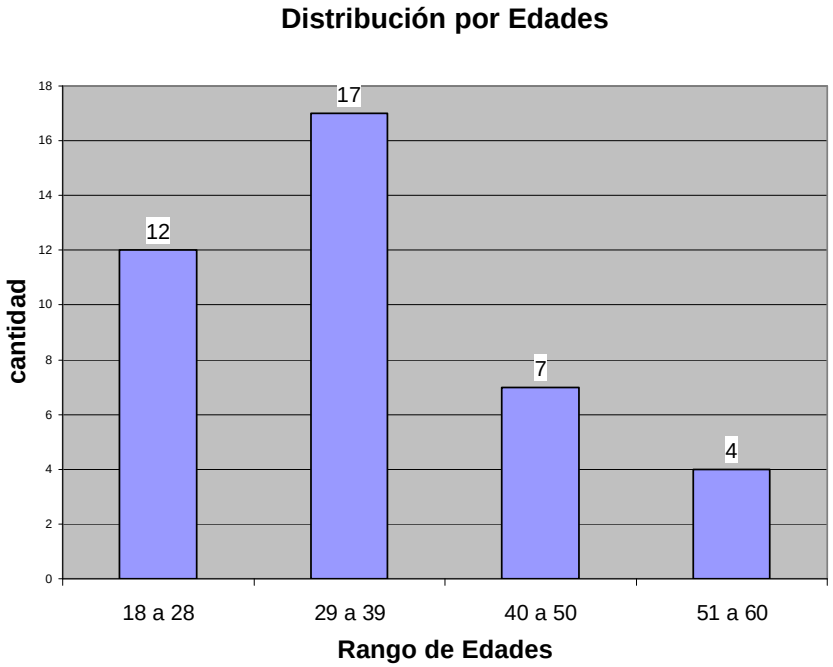


Figura 4: Distribución por Rango de Edades

3.1.4 CONTENIDO DE LA BASE.

Siguiendo las experiencias en la creación de AHUMADA, que es una base multisesión de habla espontánea, en la que se registraron sesiones de grabación microfónica y telefónicamente, de SWITCHBOARD, que es una base completamente grabada a través de canales telefónicos, y teniendo en cuenta el material con que contamos, se determinó conformar una base multisesión con voces espontáneas, grabadas por vía telefónica en los meses de mayo a septiembre del año 2007, lo que permitió conocer en mayor grado características personales de la voz de cada locutor, es decir, a medida que aumenta la cantidad de sesiones de grabación, menor es la posibilidad de encontrar al locutor con el mismo estado de ánimo.

Se seleccionaron tres sesiones de grabación, espaciadas en un tiempo de dos semanas a tres meses, dependiendo de esta separación se realizó una distribución de las distancias temporales entre las sesiones.

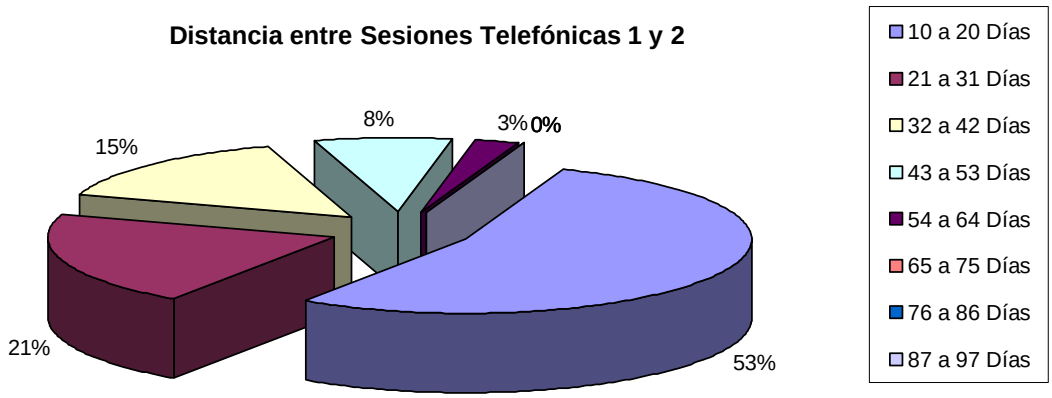


Figura 5: Distancia entre sesiones 1 y 2.

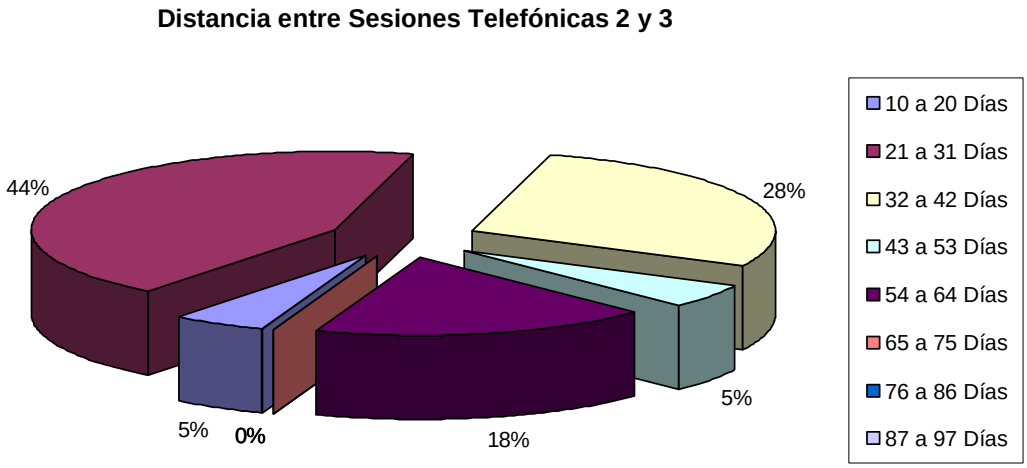


Figura 6: Distancia entre sesiones 2 y 3.

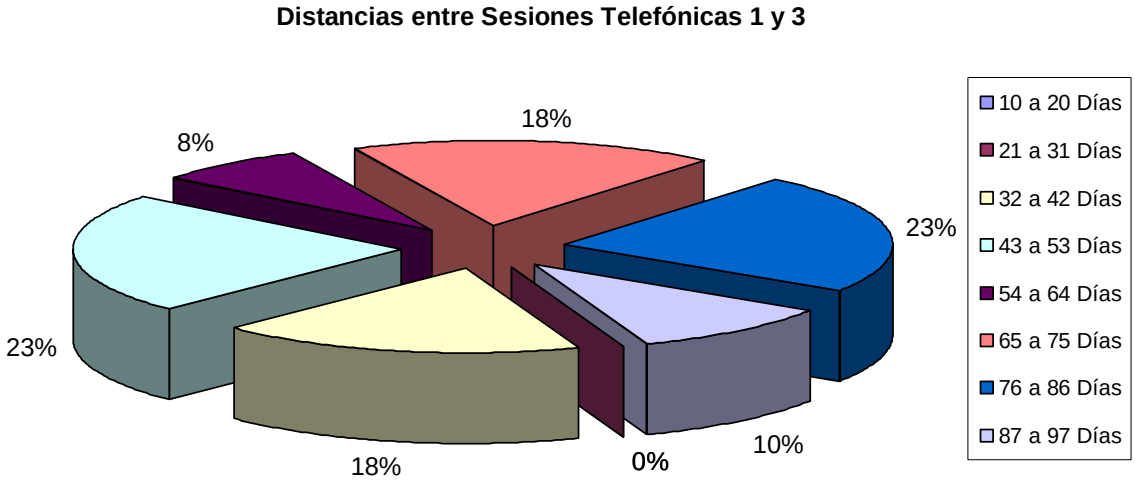


Figura 7: Distancias entre sesiones 1 y 3.

3.2 ETIQUETADO DE LA BASE DE DATOS.

3.2.1 NOMBRE DE LOS FICHEROS

Para organizar las grabaciones realizadas de cada locutor por ficheros, se realizó un estudio del procedimiento usado en la base de datos AHUMADA, en los que se tuvo en cuenta el formato de denominación utilizado.

Para almacenar los datos de los 39 locutores, sabiendo a la vista del nombre del fichero a qué locutor pertenece (del 01 al 39, de qué sesión se trata (1,2, 3), vía telefónica y en qué tarea nos encontramos, se ha adoptado realizarlo de la siguiente manera.

La nomenclatura consta de 6 caracteres, distribuidos como sigue:

A- Caracteres primero y segundo nos indican que es una grabación telefónica y de que sesión se trata.

"T1".- Grabación telefónica, 1ª sesión.

"T2".- Grabación telefónica, 2ª sesión.

"T3".- Grabación telefónica, 3ª sesión.

B- El tercer caractér especifica el sexo.

H- Hombre

M-Mujer

C- Caracteres cuarto y quinto determinan el locutor.

Del 01 al 39.

D- El sexto determina la tarea a realizar.

"A".- Habla Espontánea para todos los locutores

E- El nombre adoptado para la extensión que son las siglas WAV.

Un ejemplo de todo lo anterior nos puede aclarar la explicación hecha de la nomenclatura de los ficheros de voz.

Fichero T1M01A.WAV:

* Se trata de la grabación telefónica 1ra sesión". (Dos primeros caracteres T1)
* Sexo Mujer (tercer caractér M)
* Locutor 1 (cuarto y quinto caractér 01)
* Habla Espontánea. (Sexto caractér A)

3.2.2 ÁRBOL DE FICHEROS.

Existen variadas formas de construir y organizar los ficheros dentro del árbol de directorios. Se seleccionó ubicar dentro de un directorio denominado "BDVC" (Base de Datos de Voz Cubana) que cuelga de la unidad raíz todos los subdirectorios de locutores. Es decir, de "BDVC" penderían, las 39 carpetas que se corresponden con los locutores y que se denominan con la letra "L", de locutor, seguida del número a que corresponda. Por ejemplo: la carpeta "L04" corresponde al locutor número 4. Dentro de cada carpeta del locutor existen 3 subcarpetas que pertenecen a la sesión y vía de grabación. Estas subcarpetas son "T1", "T2" y "T3". Los nombres de los ficheros son únicos de cada sesión y locutor; nunca existirán dos nombres iguales dentro de toda la base de datos.

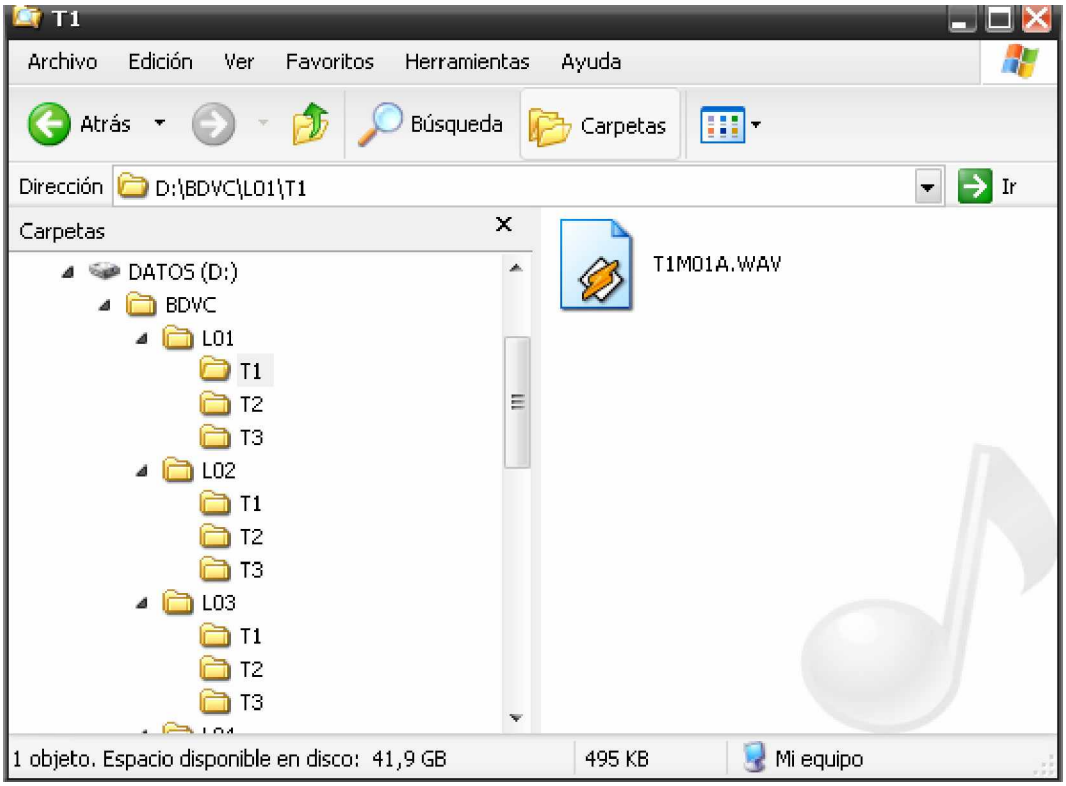


Figura 8: Árbol de Ficheros

3.2.3 CARACTERÍSTICAS TÉCNICAS

La voz fue digitalizada utilizando la tarjeta multimedia de una PC estándar en formato PCM, mono canal. En las llamadas se emplearon líneas públicas conmutadas de la red telefónica provincial y nacional, se efectuaron desde la vivienda o la oficina de los participantes, desconociéndose el tipo de auricular empleado. Las señales se digitalizaron a 8 Kbits con 8 bits de resolución y se guardaron en ficheros identificativos de cada locutor.

3.3 PROPUESTA PARA ELEVAR LA UTILIDAD DEL BANCO DE VOCES, APLICÁNDOLE DIFERENTES NIVELES DE RELACIÓN SEÑAL/RUIDO.

El diseño y desarrollo de sistemas de reconocimiento de locutores robustos al ruido, conlleva a la prueba del desempeño de los algoritmos con bancos de voces contaminados con diferentes niveles de ruido de fondo. Estos niveles de ruido deben ser medidos, basados fundamentalmente en la relación señal a ruido (SNR) presente. En aplicaciones prácticas nos encontramos con dos dificultades cuando tratamos con la señal de voz: primeramente debemos tomar en cuenta la naturaleza no estacionaria de la misma y en segundo lugar debemos resolver como estimar la SNR cuando contamos solamente con la señal más el ruido.

La definición estándar de la SNR viene dada por:

$$SNR = 10 \log \frac{\sigma_s^2}{\sigma_n^2} \qquad (1)$$

Donde:

σ_s^2 : es la potencia de la señal de voz.

σ_n^2 : es la potencia del ruido.

De acuerdo con lo anterior la SNR puede ser calculada si contamos con la potencia de la señal y con la potencia del ruido. En las aplicaciones prácticas generalmente se cuenta con la suma de la señal limpia mas el ruido, por lo cual se hace necesario estimar la SNR (Pollak, P. 2001, Vondrasek, 2005). La estimación de la SNR generalmente se lleva a cabo en el dominio de la potencia mediante:

$$\widehat{SNR} = 10 \log \frac{\widehat{\sigma}_s^2}{\widehat{\sigma}_n^2} = 10 \log \frac{\sigma_x^2 - \widehat{\sigma}_n^2}{\widehat{\sigma}_n^2}. \qquad (2)$$

Donde:

σ_x^2 : es la potencia de la señal más el ruido.

$\widehat{\sigma}_n^2$: es la potencia del ruido.

El criterio más generalizado es calcular la potencia del ruido en los intervalos de silencio y el de la señal en el resto.

Partiendo de lo expuesto anteriormente, se propone:

1. Realizar la estimación de la SNR presente en todas las grabaciones, de las diferentes sesiones que componen la base de voces, determinándose el promedio de las mismas.
2. A partir de las estimaciones anteriores, determinar los niveles necesarios de adición de ruido blanco gaussiano, con el objetivo de conformar réplicas del banco de voces originales con diferentes niveles de contaminación, que permitan probar los sistemas.

CONCLUSIONES PARCIALES.

La creación de bancos de voces sigue siendo un reto para las investigaciones que se llevan a cabo en el área de reconocimiento del locutor, se continúa trabajando en la creación de bancos de voces que cada vez constituyan una representación, lo más real posible, del universo de locutores a enfrentar por los diferentes sistemas en las distintas áreas de aplicación. En este capítulo se ha descrito la

creación de un banco de voces cubanas que, aunque pequeño, contiene habla de las tres regiones principales en las que se divide el país, por lo cual constituye una herramienta válida para la prueba de las investigaciones que se están desarrollando.

CONCLUSIONES Y RECOMENDACIONES

Conclusiones

Se realizó un estudio sobre la creación de varias bases de datos, determinándose las principales características y parámetros aplicables en la conformación de la base cubana. Se conformó una base de datos cubana contentiva del español que se habla en las tres regiones del país, cuyas características fundamentales son:

- 39 locutores (22 hombres y 17 mujeres).
- Multisesión (3 sesiones).
- Habla espontánea.
- Grabada a través de canales telefónicos.
- Al menos 1 minuto de grabación.

Recomendaciones

A partir del banco de voces conformado queda abierta la posibilidad de trabajar en su mejoramiento.

Como trabajo inmediato se recomienda:

1 Aumentar la cantidad de locutores con vista a fortalecer la base, para obtener mayor representatividad por regiones.

2 Aumentar la representatividad en las edades comprendidas entres los 40 y 60 años, por darse en este rango de tiempo los mayores cambios fisiológicos que conllevan modificaciones en el aparato productor de la voz.

3 Aumentar la duración de las grabaciones con vista a poder probar la utilización de rasgos de alto nivel, como enriquecimiento de los sistemas tradicionales de reconocimiento de locutores.

4 Contaminar la base original con diferentes niveles de relación S/N que permitirá utilizarla en condiciones que pueden encontrarse en aplicaciones prácticas.

REFERENCIAS BIBLIOGRÁFICAS

Acero, A., 1993. "Acoustical and Environmental Robustness in Automatic Speech Recognition". Kluwer Academic Publishers, Dordrecht.

Bimbot, F. et al (2004). "A Tutorial on Text-Independent Speaker Verification", EURASIP Journal on Applied Signal Processing 2004:4, 430–451

Campbell, J. (1995). "Testing with The YOHO CD-ROM Voice Verification Corpus," ICASSP. Detroit, May 1995, p. 341-344. http://www.biometrics.org/REPORTS/ICASSP95.html

Campbell, J.P (1997). "Speaker Recognition: A Tutorial", Proceedings of the ieee, vol. 85, no. 9, september 1997.

Campbell. J. P, Reynolds, D. A, (1999). "Corpora for the Evaluation of Speaker Recognition Systems" ICASSP '99, Phoenix, Arizona, pp. 2247-2250

Campbell. J. P, (2003). "Advances in Speaker Recognition: Getting to Know You". Biometric Consortium Conference. Arlington, VA

Champod, C., Meuwly, D., (1998). "The inference of identity in forensic speaker recognition". En: ESCA Workshop on Speaker Recognition and its Commercial and Forensic Applications, RLA2C, Avignon, pp. 125±134.

Doddington, G. (1998). "Speaker Recognition Evaluation and Methodology: An Overview and Perspective," Workshop on Speaker Recognition and its Commercial and Forensic Applications (RLA2C), Avignon, France, April 20-23, 1998, p. 60-66.

Doddington, G., et al., (2000) “The NIST Speaker Recognition Evaluation – Overview, Methodology, Systems, Results, Perspective”, Speech Communication 31 (2000), pp. 225-254

European Lang Resources Assoc. http://www.icp.grenet.fr/ELRA/

Furui, S. (1994). "An Overview of Speaker Recognition Tecnology ". ESCA Workshop on Automatic Speaker Recognition, Identification and Verification. Martigny, Switzerland. April 5-7, 1994.

Linguistic Data Consortium. http://www.ldc.upenn.edu/

Junqua, J.-C., Haton, J.-P., 1996. "Robustness in Automatic Speech Recognition". Fundamentals and Applications. Kluwer Academic Publishers, Dordrecht.

Martin, A. and Przybocki, M., (2000) “The NIST 1999 Speaker Recognition Evaluation – An Overview”, Digital Signal Processing 10, Num. 1-3, January/April/July 2000, pp. 1-18

Martin, A. and Przybocki, M., (2001). “The NIST Speaker Recognition Evaluations: 1996-2001”, Proc. 2001: A Speaker Odyssey, The Speaker Recognition Workshop, Crete, Greece, June 18-22, 2001, pp. 39-43

Martin, A. et al, (2004). “Conversational Telephone Speech Corpus Collection for the NIST Speaker Recognition Evaluation 2004“ Proc. Odyssey 2004, The

Speaker and Language Recognition Workshop, Toledo, Spain, May 31-June 3, 2004

Nakasone, H. and Beck, S. D., (2001). "Forensic Automatic Speaker Recognition", 2001: A Speaker Odyssey, The Speaker Recognition Workshop, Crete, Greece, June 18-22, 2001.

Nakasone, H., "Automated Speaker Recognition in Real World Condition: Controlling the Uncontrollable", Eurospeech 2003, Geneva, Switzerland, Nov. 1-4, 2003.

NIST Speaker Recognition Evaluation Plans. http://www.nist.gov/speech/test.htm

Oregon Graduate Institute. http://cslu.cse.ogi.edu/

Ortega, J. G, "AHUMADA: A large speech corpus in Spanish for speaker characterization and identification". Speech Communication 31 (2000) 255±264 www.elsevier.nl/locate/specom.

Petrovska, D., et al. (1998) "POLYCOST: A Telephone-Speech Database for Speaker Recognition," RLA2C, Avignon, France, April 20-23, 1998, p. 211-214. http://www.speech.kth.se/~melin/papers/rla2c_ply.ps

Przybocki, M. and Martin, A., (1998) “NIST Speaker Recognition Evaluation – 1997”, Proc. RLA2C, Avignon, France, April 1998, pp. 120-123

Reynolds D. A. and Rose R. C., (1995) ”Robust text-independent speaker identification using Gaussian mixture speaker models,” IEEE Trans. Speech and Audio Processing, vol. SAP-3, no. 1, pp. 72-83, January 1995.

Rossi, M. (1989) "De la quiddité des variables". Actes du seminaire Variabilité et spécificité du locuteur, Marseille Luminy. pp 11-31.

ANEXOS

Anexo I LA PRODUCCIÓN DE LA VOZ.

La voz es una onda de presión acústica. El proceso de producción natural de la voz pasa por varias etapas que incluyen: la concepción del mensaje en el cerebro, la transmisión de lo impulsos nerviosos necesarios a cada uno de los órganos y músculos involucrados y la ejecución en los mismos de los movimientos correspondientes. En esta última radica el basamento físico de generación de la voz. En la generación de la voz intervienen varias estructuras del cuerpo humano como son la laringe, que contiene a las cuerdas vocales, la faringe, las cavidades oral y nasal y una serie de elementos articulatorios como son los labios, los dientes, los alvéolos, el paladar, el velo del paladar, y la lengua.(Juang 1993) Además, los pulmones con su musculatura aledaña que son los encargados de suministrar el aire a la presión adecuada para excitar lo que se conoce como aparato fonador humano, el aparato se muestra en el esquema de la figura 8

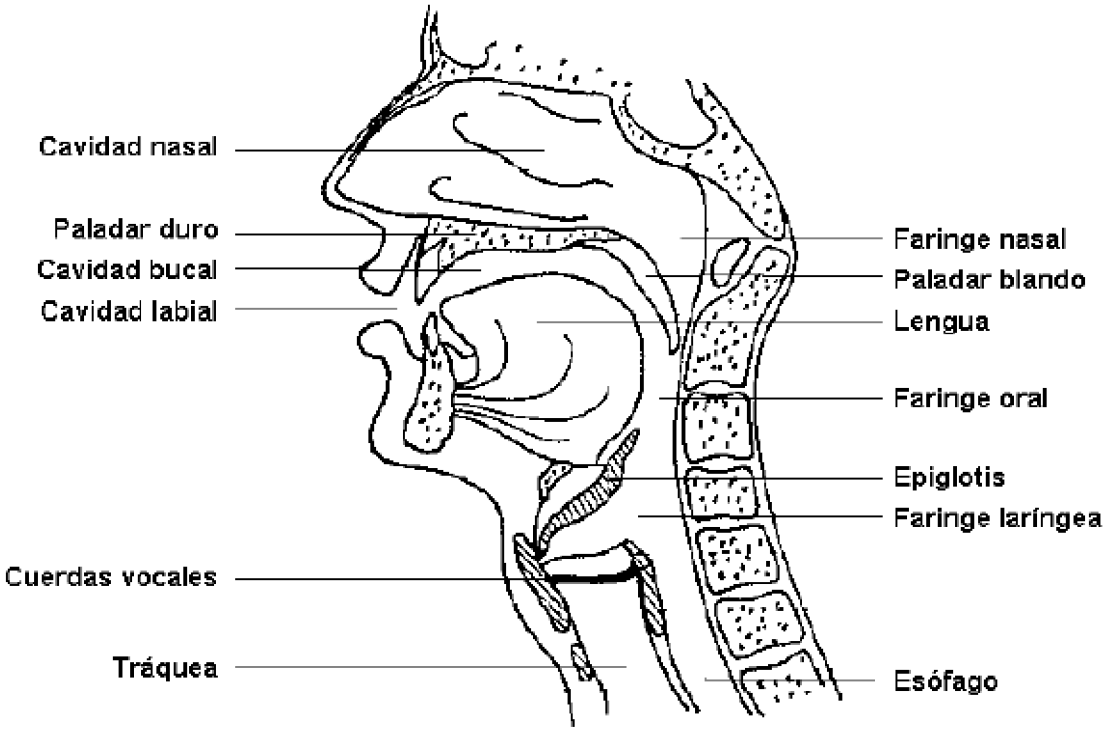


Figura 8 El aparato fonador humano (Tomado de: http://ispl.korea.ac.kr/~wikim/research/speech.html)

El funcionamiento aproximado del sistema fonador humano puede explicarse de manera muy sencilla como se ilustra en la figura 1.2, los pulmones actúan como una fuente de energía; un flujo de aire al ser comprimido por el diafragma pasa a través de las cuerdas vocales, que son en realidad dos membranas dentro de la laringe cuya abertura se conoce como glotis. Este proceso puede ocurrir en el momento en que la glotis esté completamente abierta, caso en el cual no se produce sonido o en el momento en que la glotis comienza a cerrarse, caso en que el aire experimenta una turbulencia produciéndose un ruido de origen aerodinámico conocido como aspiración o en el momento en que la glotis está aún más cerrada, que es cuando las cuerdas vocales comienzan a vibrar y se produce un sonido periódico cuya frecuencia depende del tamaño y masa de las cuerdas vocales, la tensión que se aplique y la velocidad del flujo de aire proveniente de los pulmones. Así, a mayor tamaño de las cuerdas vocales, menor frecuencia; a

mayor tensión aumenta la frecuencia y al aumentar la intensidad de emisión se tiende a aumentar el tono de la voz. Por su parte, la faringe, las cavidades orales y nasales y la cavidad labial llevan a cabo un proceso de filtrado, actuando como resonadores acústicos que enfatizan determinadas bandas de frecuencia, dando lugar a lo que se conoce como formantes. Es así que estas cavidades actúan modificando el espectro del sonido mientras que los elementos articulatorios llevan a cabo una modificación a nivel temporal relacionada con la emisión de los mismos, es decir, con el lugar del tracto vocal donde se producen y con los fenómenos transitorios que lo acompañan (Proakis 1987).

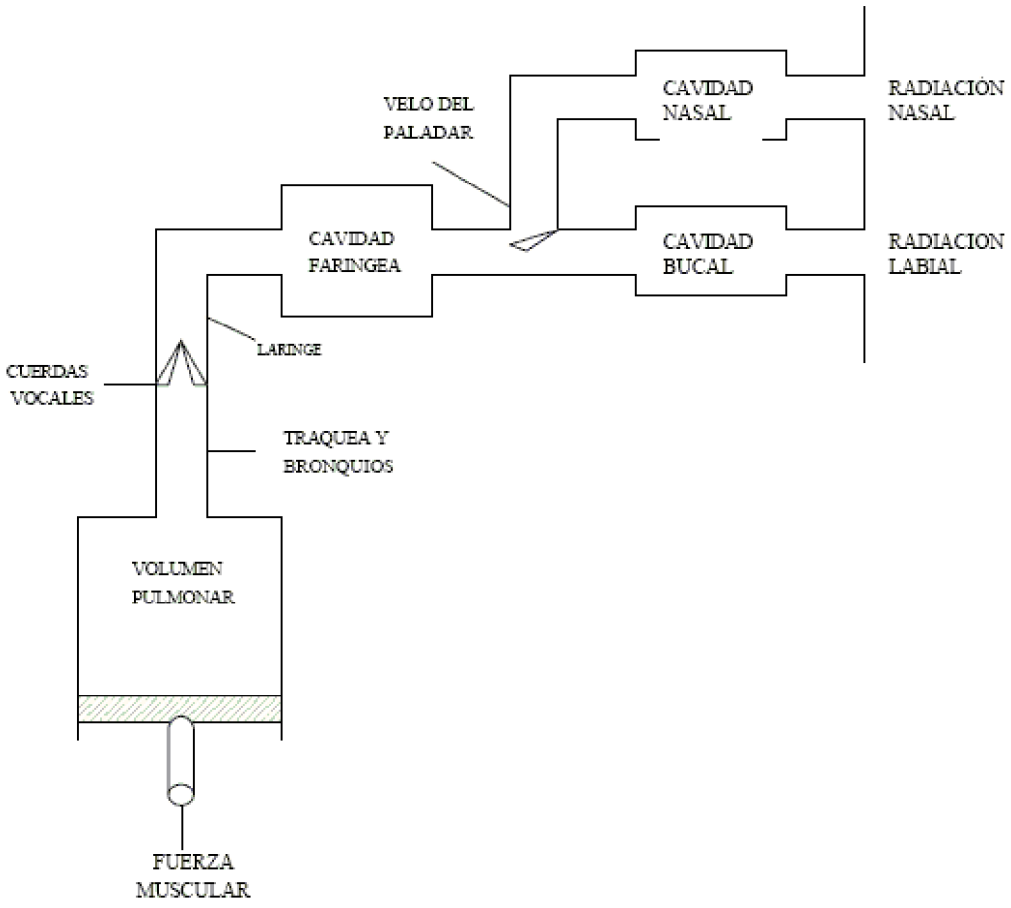


Figura 9: Representación Esquemática del mecanismo fisiológico de producción de voz. (Tomado de: http://ispl.korea.ac.kr/~wikim/research/speech.html)

El proceso conocido como síntesis de la voz se basa en el estudio del aparato fonador humano, y en la implementación de modelos basados en el mismo, mediante hardware o software, que permiten la producción de voz de manera artificial. Es válido aclarar que la voz sintética no es voz pregrabada, sino voz generada a partir del estudio y la simulación del aparato fonador humano. Existen básicamente dos características usadas para describir la calidad de un sistema de síntesis de la voz, estas son la naturalidad y la inteligibilidad. La naturalidad de un sintetizador de la voz se refiere a la similitud entre los sonidos sintetizados con relación a los sonidos producidos por una persona. La inteligibilidad describe la comprensibilidad de los sonidos. El sintetizador de la voz ideal se caracteriza por ser natural e inteligible y cada una de las tecnologías y algoritmos de síntesis de la voz tratan de maximizar estas características.

Anexo II **LA VOZ NORMAL**

ALTERACIONES

Existen distintas disfunciones que determinan que una voz no sea normal. La voz normal es la producida por el locutor sin ningún tipo de acondicionamiento interno ni externo, esta sufre modificaciones según se trate de la mañana, tarde o noche, también se pueden encontrar problemas vocálicos con los cambios de temperatura, humedad, pero ante situaciones de reposo del locutor puede mejorar.

Por ejemplo en el caso de las mujeres se añaden problemas en la dicción según su ciclo menstrual (antes, durante o después), así como con el embarazo, menopausia y uso de contraceptivos orales. Estos últimos afectan especialmente a las altas frecuencias.

La tensión nerviosa es un factor que no requiere tratamiento médico para su solución. Puede aparecer por la inexperiencia a hablar en público o ante una audiencia numerosa. También puede darse al extrañarse el entorno donde se desarrollan sus hábitos de vida. Este factor permite detectar también personas con

problemas de tartamudeo, el cual desaparece a medida que aumentan las condiciones de autoconfianza en la situación en la que está inmerso.

Los aspectos psicológicos sí requieren tratamiento médico. Uno de vital importancia en la voz es el estrés. Se define como un conjunto de perturbaciones fisiológicas y metabólicas producidas por el organismo como consecuencia de distintos o variados agentes agresores en un cierto período de tiempo. El estrés sobreviene cuando el organismo falla en su mecanismo de adaptación a estímulos de origen interno o externo. No es por si mismo una enfermedad, pero al despertar o perturbar los mecanismos de defensa psicológicos puede jugar un papel importante en el desarrollo de determinadas enfermedades. El estrés ocasiona problemas disfónicos que distorsionan la voz normal. Dentro de la psicología, se destacan además distintas patologías psiquiátricas que también afectan a la voz. La gravedad de estas enfermedades depende básicamente del grado de impedimento de contacto con la realidad y la alienación de la persona. Una de las más típicas es la neurosis. La histeria es una neurosis caracterizada por la expresión somática exagerada de conflictos emocionales inconscientes. Entre sus síntomas funcionales duraderos más comunes están las afonías o pérdida de la voz de conversación con tendencia al cuchicheo. La ansiedad o miedo, independientemente de su origen, igualmente puede afectar a la voz normal. Las neurosis obsesivas algunas veces tienden a reflejar algunas de sus características en el lenguaje o en la voz. La intensidad puede ser baja o a veces la articulación ser pobre. Suelen tener una inclinación al silencio. La depresión es uno de los problemas emocionales más comunes.

Existen dos tipos de trastornos de la voz que requieren en la mayoría de los casos intervención quirúrgica. En el grupo de los orgánicos se encuentran los carcinomas y papilomas, alteraciones de la mucosa, trastornos neuromusculares (paresia, atrofia de cuerdas vocales, temblor, distonía y parkinsonismo). El grupo de los funcionales lo integran los debidos al abuso o mal uso de algún órgano. Aquí se encuadran los nódulos vocales, las hinchazones simétricas bilaterales de las cuerdas vocales y las lesiones circunscritas en la unión de los tercios anterior y medio de las cuerdas vocales. Todas estas anomalías se consideran fruto de una

biomecánica laríngea anormal. Esta clasificación en orgánicos y funcionales no es tan clara en la realidad. La presencia de un granuloma puede deberse a abuso vocal (funcional), reflujo gastroesofágico (orgánico) o a una combinación de ambos. Entre las molestias y síntomas vocales más comunes ocasionados por estos trastornos, está la disfonía o voz anormal. Dentro de ésta, tenemos la diplofonía (doble tono), disresonancia (pérdida o cambio en la resonancia vocal), quiebre de la voz (cambio anormal del tono de la voz), odinofagia u odifonía (voz con dolor o incomodidad) y reducción de la tesitura (reducción del rango o margen dinámico de la voz).

Se resalta también la influencia que puede tener en la voz de los locutores el que sean drogodependientes o estén sometidos a una medicación frecuente por enfermedad distinta a las ya especificadas. La inmediata solución estriba en el cese de las actividades que originan el problema.