UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FISICAS Y MATEMATICAS
DEPARTAMENTO DE INGENIERIA ELECTRICA

EVALUACION AUTOMATICA DE PRONUNCIACION DE FRASES PARA HABLANTES NO NATIVOS

TESIS PARA OPTAR AL GRADO DE MAGISTER EN CIENCIAS DE LA INGENIERIA, MENCION INGENIERIA ELECTRICA

MEMORIA PARA OPTAR AL TITULO DE INGENIERO CIVIL ELECTRICISTA

LEOPOLDO FELIPE ANDRES BENAVIDES BERRIOS

PROFESOR GUÍA:
Sr. NESTOR BECERRA YOMA

MIEMBROS DE LA COMISION:
Sr. JUAN VELASQUEZ SILVA
Sr. CARLOS MOLINA SANCHEZ

SANTIAGO DE CHILE

2011

Resumen de la tesis para optar al título de:
Magíster en Ciencias de la Ingeniería, Mención Ingeniería Eléctrica.
Ingeniero Civil Electricista.
Nombre Alumno: Leopoldo Felipe Andrés Benavides Berríos.
2011.
Profesor Guía: Sr. Néstor Becerra Yoma.

"Evaluación automática de pronunciación de frases para hablantes no nativos"

En esta tesis se aborda el problema de la evaluación automática de pronunciación de frases para hablantes no nativos en el marco de los sistemas de aprendizaje de idiomas asistido por computadora (CALL *Computed Aided Language Learning*). Dentro de los sistemas CALL, existe un área enfocada a mejorar uno de los aspectos fundamentales del habla, la pronunciación. El entrenamiento de pronunciación asistido por computadoras (CAPT *Computer Aided Pronunciation Training*) invita a los estudiantes a mejorar la calidad del lenguaje hablado mediante la repetición y evaluación de su pronunciación, funciona como un profesor virtual y busca ser una herramienta complementaria a las clases tradicionales. En los últimos años ha habido grandes avances en esta área, en particular con palabras aisladas donde se alcanzan altos niveles de correlación entre la evaluación entregada por el sistema de reconocimiento de voz (ASR, *Automatic Speech Recognition*) y los evaluadores expertos. Esta tesis centra su atención en la evaluación de pronunciación en frases, ya que en la literatura generalmente se considera toda la frase como una única unidad a evaluar, sin tener presente por ejemplo que la correcta pronunciación puede variar entre las palabras de una frase o que la evaluación de una palabra aislada es diferente si está inmersa en una oración.

Este trabajo propone una nueva alternativa para la evaluación de pronunciación de frases donde se evalúa cada una de las palabras que componen las frases para luego obtener una evaluación objetiva para toda la frase. Además no se requieren estudios previos de errores comunes de pronunciación del lenguaje materno, para ello se generan modelos competidores para frases de manera no supervisada, se prueban variantes fonéticas para mejorar la exactitud de la evaluación y se proponen diferentes criterios subjetivos para ampliar la evaluación de frases, unos más exigentes que otros. Como resultado se obtuvieron correlaciones en el rango 0.4 a 0.52 entre la evaluación entregada por ASR y los evaluadores expertos. El mejor resultado se obtuvo al ocupar el promedio de las calificaciones obtenidas por palabras para generar la calificación por frase, 0,52 de correlación con 5 niveles de evaluación, en este caso la incorporación de las variantes de pronunciación significó un aumento de 10,8% con MCS (*Multiple Classifier System*) MVR (*Majority Vote Rule*) al compararla con la ausencia de esta alternativa.

Los resultados alcanzados ubican esta nueva alternativa para la evaluación de pronunciación de frases dentro del rango de correlación encontrados en la literatura, lo cual lo hace una alternativa viable para seguir investigando. Las correlaciones alcanzadas para frases tanto en este trabajo como en la literatura aun no son muy altas, por lo cual este sistema seguirá siendo una herramienta complementaria a las clases tradicionales y no busca reemplazar la evaluación entregada por un experto.

Agradecimiento

En primer lugar quiero agradecer a mi familia por todo el apoyo y comprensión brindada durante toda mi educación. A mi madre que siempre me entrego mucho amor y las palabras adecuadas cuando las cosas no iban bien, a mi padre por su apoyo incondicional en todo desafío o aventura que emprendí y a mi tía por ser una segunda madre. Soy un hombre afortunado, gracias a Dios por tan grande bendición.

Agradezco también a todos mis amigos y compañeros, ellos también han contribuido a ser la persona que hoy ustedes conocen y de lo cual me siento orgulloso, en particular a Javier Urbina, gran amigo y compañero de armas en los libros y cuadernos, grandes batallas nos preceden desde el colegio, luego en plan común y finalmente en el departamento de ingeniería eléctrica.

Finalmente agradezco a un pequeño grupo de amigos que me han acompañado en estos últimos años: Claudio, Jorge, Juan Pablo, Carlos, Chaitenino, Isaías, Pablo R y Víctor, el quipo del Laboratorio de Procesamiento y Transmisión de Voz, LPTV, para mí amigos Leales Pacientes Trabajadores y Virtuosos. También agradezco al profesor Néstor Becerra, líder del LPTV y mi profesor guía en este trabajo por darme la oportunidad de incorporarme a este gran grupo y orientarme en el desarrollo de la tesis.

Tabla de contenido

1. Introducción 10

1.1. Evaluación de pronunciación de Frases y aplicación al aprendizaje de un segundo idioma. 10

1.2. Motivación. 11

1.3. Hipótesis de investigación. 12

1.4. Estructura de la Tesis. 13

2. Reconocimiento automático de voz aplicado a la enseñanza del segundo idioma 15

2.1. Interfaces Hombre-Máquina. 15

2.2. Enseñanza segundo idioma asistido por computadora (CALL). 17

2.2.1. Sistemas CALL. 17

2.2.2. Evaluación de la Pronunciación en CALL (CAPT). 20

2.3. Reconocimiento Automático de Voz (ASR). 21

2.3.1. Formulación Matemática del problema de Reconocimiento de Voz. 23

2.3.2. Parametrización acústica. 25

2.3.3. Modelación acústica con HMM. 29

2.3.4. Modelo de lenguaje. 34

2.3.5. Algoritmo de Viterbi. 35

2.4. Evaluación de Pronunciación con ASR. 39

2.4.1. Categorías de evaluación de pronunciación. 40

2.4.2. Evaluación automática de pronunciación basada en extracción de características. 41

2.4.3. Introducción de errores de lenguaje materno. 44

2.4.4. Medidas de desempeño. 44

2.4.6. Sistemas de múltiples clasificadores (MCS). 51

2.4.7. Aporte. 54

3. Estimación automática de pronunciación de frases con ASR 56

3.1. Criterios subjetivos en frases. 57

3.2. Calificación objetiva de frases como una combinación de calificaciones objetivas para palabras. 59

3.3. Correspondencia entre la evaluación objetiva y subjetiva de frases. 60

3.4. ASR basado en métricas objetivas con vocabulario competitivo y clases basadas en modelo de lenguaje. 61

3.5. Generación automática competidoras en las clases. 64

3.6. Generación automática de vocabulario competitivo. 65

3.7. Generación automática de variantes fonéticas en español para palabras *target*. 66

3.8. Clases basadas en modelos de lenguaje en ASR. 69

3.9. Extracción de características para palabras basada en la lista *N-best.* 71

3.9.1. POS (*N-best position*). 71

3.9.2. REC (*Recognition flag*) . 72

3.9.3. WDCM (*Word density confidence measure*). 72

3.9.4.	LogWD (*Logarithmic word density confidence measure*).	73
3.10.	Estimación de la calificación objetiva de pronunciación de palabras.	73
3.10.1.	Regla del Promedio.	75
3.10.2.	MVR *(Majority vote rule*).	76
4.	**Implementación, Discusión y resultados**	**77**
4.1.	Implementación.	77
4.1.1.	Base de Datos.	77
4.1.2.	Generación Modelos competidores.	80
4.1.3.	Estructura de experimentos.	87
4.2.	Discusión y resultados.	89
5.	**Conclusiones**	**106**
5.1.	Conclusiones y análisis finales.	106
5.2.	Trabajo propuestos a futuro.	108
6.	**Glosario**	**109**
7.	**Referencias**	**113**
8.	**Anexo, “On modeling criteria for ASR based pronunciation quality evaluation of sentences"**	**121**

Índice de Tablas

Tabla 1: Características o parámetros típicos de un ASR 22
Tabla 2: Algoritmo de búsqueda de Viterbi 38
Tabla 3: Traducción fonemas Inglés a Español 67
Tabla 4: Traducción fonemas de Español al Inglés 68
Tabla 5: Ejemplo de evaluaciones de frases. 79
Tabla 6: Ejemplo de archivo modelo_1.tbl. 82
Tabla 7: Ejemplo de archivo modelo_clases_demo.txt. 83

Índice de Figuras

Figura 1: Diagrama que describe el funcionamiento de un ASR 22
Figura 2 : Diferencias en el dominio temporal (izquierda) y espectral (derecha) para la frase "*the airport is so big*" pronunciada por dos locutores diferentes. 26
Figura 3: Estructura de un HMM con topología de izquierda-derecha 30
Figura 4: Secuencia de HMMs que forman la frase "*The airport is so big*". 32
Figura 5: Representación gráfica del algoritmo de Viterbi 36
Figura 6: Diagrama de bloques del proceso de entrenamiento de las curvas a priori. 43
Figura 7: Diagrama que muestra la generación del score objetivo utilizando mapeo de Bayes. 43
Figura 8: Funciones de densidad de probabilidad condicionales para cierto fenómeno. Distr 1 corresponde a $p(x|\omega_1)$ mientras que Distr 2 a $p(x|\omega_2)$. 48

Figura 9: Fusión MCS en *abstract level*. 52
Figura 10: Fusión MCS en *feature leve.* 53
Figura 11: Diagrama de Bloques que muestra la generación del score objetivo para frases. 63
Figura 12: Descomposición en las diferentes variantes foneticas de pronunciación. 67
Figura 13: Sistema 2LL web diseñado por LPTV 78
Figura 14: Estructura de Experimentos. 88
Figura 15: Correlación entre la calificación objetiva y subjetiva v/s MNCW 90
Figura 16: Correlación entre la calificación objetiva y subjetiva v/s $[D_{\min}^{CL}, D_{\max}^{CL}]$ 90
Figura 17: Correlación entre la calificación objetiva y subjetiva v/s $PV_{m,l}$ 93
Figura 18: Correlación entre la calificación objetiva y subjetiva v/s $Class_{m,l}$ 93
Figura 19: Correlación entre la calificación objetiva y subjetiva v/s $D_{\min}^{PV}$ 94
Figura 20: Correlación entre la calificación objetiva y subjetiva v/s MNCW 95
Figura 21: Correlación entre la calificación objetiva y subjetiva v/s $[D_{\min}^{CL}, D_{\max}^{CL}]$ 97
Figura 22: Correlación entre la calificación objetiva y subjetiva v/s $PV_{m,l}$ 97
Figura 23: Correlación entre la calificación objetiva y subjetiva v/s $Class_{m,l}$ 98
Figura 24: Correlación entre la calificación objetiva y subjetiva v/s $D_{\min}^{PV}$ 99
Figura 25: Correlación entre la calificación objetiva y subjetiva v/s MNCW 100
Figura 26: Correlación entre la calificación objetiva y subjetiva v/s $[D_{\min}^{CL}, D_{\max}^{CL}]$ 101
Figura 27: Correlación entre la calificación objetiva y subjetiva v/s $PV_{m,l}$ 101
Figura 28: Correlación entre la calificación objetiva y subjetiva v/s $Class_{m,l}$ 102
Figura 29: Correlación entre la calificación objetiva y subjetiva v/s $D_{\min}^{PV}$ 104

Figura 30: Correlación subjetiva-objetivoa con MCS promedio. Cada criterio subjetivo es modelado con ObjMetrComb1, ObjMetrComb2 y ObjMetrComb3 ocupando la mejor configuracion obtenida para cada criterio subjetivo. 104

Figura 31: Correlación subjetiva-objetivoa con MCS MVR. Cada criterio subjetivo es modelado con ObjMetrComb1, ObjMetrComb2 y ObjMetrComb3 ocupando la mejor configuracion obtenida para cada criterio subjetivo. 104

CAPITULO I

1.Introducción

1.1.Evaluación de pronunciación de Frases y aplicación al aprendizaje de un segundo idioma.

En la actualidad, las tecnologías computacionales han incursionado prácticamente en todas las áreas y actividades que se realizan cotidianamente. El aprendizaje de un segundo idioma no es una excepción [Franco *et al*, 1997; Cincarek *et al.*,2008; Molina *et al* ,2009]. La utilización de tecnologías computacionales entrega un mayor grado de interactividad al estudiante, le permite avanzar a su propio ritmo y participar activamente.

Conforme avanza la tecnología también lo hacen las exigencias del público y del mercado. Si en un comienzo ejercicios de lectura y evaluación de un texto eran suficientes como complemento para el aprendizaje de un segundo idioma, ya no lo son. Actualmente se busca desarrollar sistemas basados en reconocimiento de voz que puedan complementar las actividades de comprensión, lectura, pronunciación y escucha. Los sistemas de aprendizaje de idiomas asistido por computadora (CALL *Computed Aided Language Learning*) pueden plantear problemas, analizar la respuesta y automáticamente entregar al estudiante un *feedback* en forma de calificación y/o presentar los errores que se han cometido.

Dentro de los sistemas CALL, existe un área enfocada a mejorar uno de los aspectos fundamentales del habla, la pronunciación. El entrenamiento de pronunciación asistido por computadoras (CAPT *Computer Aided Pronunciation Training*) invita a los estudiantes a mejorar la calidad del lenguaje hablado mediante la repetición y evaluación de su pronunciación, funciona como un profesor virtual. En los últimos años ha habido grandes avances en esta área, pero se mantienen ciertos problemas e imprecisiones que impiden su masificación. En esta tesis se aborda la evaluación de pronunciación en frases ya que se considera que es un tema muy importante en CAPT, no ha sido suficientemente abordado y presenta una infinidad de posibilidades.

1.2. Motivación.

La evaluación de la calidad de pronunciación ha sido ampliamente abordada en la literatura. En palabras aisladas se han obtenido correlaciones superiores a 0.7 [Cincarek *et al.*,2008; Molina *et al* ,2009]. Sorprendentemente se ha encontrado que la evaluación de frases en general se realiza como un todo y no se presta mayor atención a las palabras que la forman salvo contadas excepciones, en estos casos la correlación obtenida para frases oscila entre 0,4 y 0,7 [Franco H. *et al*., 1998; Neumeyer L. *et al*, 1999; Moustroufas N. and Digalakis V.,2006]. La evaluación de pronunciación de frases presenta dificultades extras que ameritan tratar este tema, estas dificultades son: a) La correcta pronunciación puede variar entre las palabras de la frase. b) La evaluación de una palabra aislada es diferente si está inmersa en una oración. c) Una o más palabras mal pronunciadas pueden significar que no se entienda la

oración. Estos problemas en una o más palabras pueden ser enmascarados por el resto de la frase entregando una evaluación de pronunciación errada.

Para intentar dar solución a esta problemática y obtener una evaluación de pronunciación más adecuada, se busca una calificación de la frase basada en la calificación para cada una de sus palabras, de esta manera es posible ampliar la evaluación de frases a diferentes criterios subjetivos, unos más exigentes que otros. Finalmente se busca obtener una alta correlación entre la evaluación subjetiva y la objetiva entregada por el reconocedor.

1.3. Hipótesis de investigación.

Esta tesis propone una nueva alternativa para la evaluación de pronunciación de frases basada en la extracción de características de las palabras que las componen, esto se reflejará en obtener una correlación del orden o superior a la que se puede encontrar en la literatura para frases.

Esta nueva alternativa para la evaluación de pronunciación de frases busca ser una herramienta complementaria a las clases tradicionales, para lo cual es necesario que la evaluación sea similar a la entregada por un calificador experto.

Los objetivos específicos de este trabajo son los siguientes:

- Generar modelos competitivos para frases de manera no supervisada.

- Generalizar el método para generar un modelo competitivo con distancia entre palabras para frases.

- Incorporar alternativas de pronunciación dentro de las frases competidoras y probar si mejoran la exactitud de la evaluación de calidad de pronunciación.

- Modelar los Criterios subjetivos utilizados para obtener una buena correlación entre las notas subjetivas y objetivas entregadas por el reconocedor.

- En base a la evaluación por palabras obtener una evaluación de pronunciación para frase objetiva cercano a la evaluación subjetiva.

1.4.Estructura de la Tesis.

El capítulo 2 introduce al lector en los temas de reconocimiento de voz, los sistemas de evaluación de pronunciación y enseñanza de segundo idioma asistido por computadoras. Tiene como objetivo entregar una base teórica que permita comprender los métodos propuestos y análisis realizados.

En el capítulo 3 se describe el método propuesto para la evaluación de pronunciación de frases. El desarrollo comprende la generación del modelo competitivo no supervisado, la incorporación de variantes fonéticas y la obtención de las notas objetivas obtenidas para las frases con el reconocedor.

En el capítulo 4 se explica la implementación del trabajo propuesto y se muestran las diferentes configuraciones y los experimentos realizados para probar los métodos propuestos en el capítulo 3. Se presentan y se discuten las mejores configuraciones para los distintos criterios subjetivos.

Finalmente, en el capítulo 5 se presentan las conclusiones y análisis finales de este trabajo, también se proponen trabajos futuros al respecto.

Como contribución adicional y fruto de este trabajo fue enviado el *paper* "*On modeling criteria for ASR based pronunciation quality evaluation of sentences*" a SPECOM (*Speech Communication*). Publicación de EURASIP (*European Association for Signal Processing*) e ISCA (*International Speech Communication Association*), publicación que posee un factor de impacto 1.19. Se adjunta el *paper* en los anexos.

CAPITULO II

2. Reconocimiento automático de voz aplicado a la enseñanza del segundo idioma

En este capítulo se interiorizará al lector en la tecnología de reconocimiento de voz, la evaluación de pronunciación y su aplicación a la enseñanza de un segundo idioma. Se busca generar una base teórica suficiente para comprender los análisis y técnicas presentes en esta tesis. Las siguientes secciones presentan el tema desde una perspectiva general y a medida que se avanza en este capítulo se va profundizando en los temas más relevantes para este trabajo.

2.1. Interfaces Hombre-Máquina.

Una de las formas más naturales de comunicación del ser humano es el lenguaje hablado. Debido a esto no es de extrañar que desde hace mucho se ha logrado establecer una comunicación hablada con las maquinas que utilizamos, esto se puede extender a entregar una orden para iniciar un proceso productivo en una fabrica o establecer un dialogo conversacional para aprender un segundo idioma.

En particular los diálogos conversacionales presentan importantes desafíos para la introducción de tecnologías del habla. Los fenómenos que puedan suceder en una interacción hombre – máquina son muy difíciles de conocer debido a lo impredecible del comportamiento humano, en este contexto, la tecnología en ocasiones no es suficiente para conseguir el resultado deseado. Los sistemas de reconocimiento automático de voz (ASR *Automatic Speech Recognition*) han tenido que ir adaptándose a las diferentes aplicaciones en las cuales se ocupan, aquí se incluyen temas como la robustez en el traspaso de la información hasta la percepción de la aplicación por parte de los usuarios.

La búsqueda de una comunicación amigable es el camino a seguir de la interacción hombre - máquina y la evolución a través del tiempo así lo demuestra. En sus inicios interfaces con teclados y botones hasta la incorporación de elementos multimedia en la actualidad buscan que el usuario entienda y disfrute del uso de la interfaz.

Desarrollos en este campo han permitido a los usuarios interactuar con sistemas de transferencia de información más complejos, como en aplicaciones telefónicas donde el usuario navega a través de plataformas IVR (*Interactive Voice Response*) [Seneff *et al*., 1998] permitiendo que a través del teléfono y mediante un dialogo conversacional se pueda tener acceso a diversas fuentes de información. En la actualidad internet es un campo fértil para aplicaciones que utilicen el reconocimiento de voz que aun no somos capaces de imaginar.

2.2. Enseñanza segundo idioma asistido por computadora (CALL).

El idioma es un sistema complejo de comunicación que se analiza en varios niveles. Destacan: fonología, sintaxis, morfología y semántica entre otros. El aprendizaje de un segundo idioma se define como el proceso de aprender un lenguaje a cualquier nivel cuando ya se ha adquirido uno anteriormente, el lenguaje materno.

2.2.1. Sistemas CALL.

El aprendizaje de idiomas asistido por computadora (CALL *Computed Aided Language Learning*) consiste en utilizar computadores y otras herramientas multimedia como internet para presentarle al estudiante los contenidos de una manera interactiva facilitando y motivando el aprendizaje e instrucción de alguna lengua extranjera.

Enseñar y potenciar todas las habilidades necesarias para aprender una lengua, es una tarea difícil de lograr para un profesor en una sala de clases con muchos alumnos y con un tiempo limitado. En particular podemos suponer que un profesor no es capaz de evaluar adecuadamente la pronunciación de cada uno de sus alumnos en un tiempo prudente en un curso de más de 30 alumnos. CALL se presenta como una herramienta complementaria a la metodología tradicional, la enseñanza presencial, en ningún caso intenta reemplazar el trabajo efectuado por un profesor.

Otra ventaja de los sistemas CALL es que incentiva a los estudiantes al auto aprendizaje. Debido a la libertad que experimentan, ellos pueden elegir y practicar una actividad todas las veces que lo crean necesario. Incluso el profesor puede realizar actividades personalizadas, según las necesidades de sus estudiantes, lo cual aumenta su interés y motivación. Si a todo esto se le agrega la capacidad tecnológica existente en la actualidad como video, animaciones y la integración de reconocimiento de voz los sistemas CALL pueden ser mucho más motivadores, atractivos y útiles a la hora de aprender un idioma.

El aprendizaje de lenguas extranjeras asistidas por computadoras se remonta a los años 60s, como una relación simbiótica entre la tecnología y la pedagogía. El desarrollo de los sistemas CALL se divide en tres fases: estructural; comunicativa; integrativa [Warschauer, 2008].

La fase estructural se ubica entre los 60s y finales de los 70s, consta de actividades en forma de texto. El estudiante leía y entregaba una respuesta, el sistema la evaluaba y entregaba un *feedback* al estudiante. La fase comunicativa se extiende entre los 80s y los 90s donde se enfatiza la interacción y la enseñanza ayudadas fuertemente por la aparición y masificación de los computadores personales. En esta época aparecen aplicaciones como juegos que entregan un contexto a los estudiantes para usar el lenguaje. Finalmente la fase integrativa se extiende hasta nuestros días y ocupa las tecnologías multimedia, redes computacionales como internet y tecnologías de procesamiento de señales y reconocimiento automático de voz. Todos estos elementos buscan generar un escenario lo mas verídico posible donde aprender y practicar el idioma.

La posibilidad de recibir respuestas habladas por parte de aplicaciones de instrucción de lenguaje basadas en computador, permite complementar las actividades de lectura y audición, con actividades de producción de lenguaje y entregar respuestas tal como lo haría un instructor humano, como calificar la pronunciación realizada o identificar los errores cometidos [Franco *et al*, 1997].

Para implementar verdaderos diálogos entre el estudiante y el computador que simulen una situación real se ocupa la tecnología de reconocimiento automático de voz (ASR *Automatic Speech Recognition*). A grandes rasgos existen dos enfoques: respuesta cerrada y respuesta abierta. En respuesta cerrada el sistema tiene un número finito de posibles respuestas, las cuales se despliegan en pantalla y el estudiante sabe exactamente lo que debe decir. Por otro lado, en los sistemas de respuesta abierta se formula una pregunta al estudiante y este debe generar una respuesta acorde a la pregunta, estos sistemas son mucho más complejos en cuanto a requerimientos técnicos.

Las mayores expectativas y exigencias en la actualidad hacen que los sistemas necesiten computadores con mayor capacidad de procesamiento, pensando en la masificación del aprendizaje de segundo idioma enfocado a colegios de escasos recursos se aborda la estrategia distribuida [Molina *et al*, 2008]. El sistema distribuido consiste en un servidor central con una gran capacidad de procesamiento al cual se conectaran vía LAN o internet los clientes, estudiantes, esto disminuye en gran medida los costos ya que los clientes pueden contar con computadores mucho más básicos ya que todo el procesamiento que involucra un gran uso de recursos lo efectuará el servidor

central que puede estar ubicado en el mismo colegio o en otra entidad educativa como una universidad.

2.2.2. Evaluación de la Pronunciación en CALL (CAPT).

La tecnología ASR ha sido aplicada también a la enseñanza de pronunciación. La pronunciación y su calificación corresponden a uno de los problemas principales que muchos científicos han intentado resolver en el marco de los sistemas CALL y el reconocimiento de voz. El entrenamiento de pronunciación asistido por computadoras (CAPT *Computer Aided Pronunciation Training*) es una gran posibilidad para la masificación de las tecnologías de voz en los sistemas CALL.

La mejor manera de aprender un idioma es estar expuesto a su influencia el mayor tiempo posible. Como se dijo antes, para un profesor en una clase tradicional le es muy difícil prestar atención a todos sus alumnos. Además requeriría de mucho tiempo para practicar la pronunciación con cada uno de ellos. Considerando que el *feedback* del profesor es imprescindible para el aprendizaje de la pronunciación. Aquí se puede ver la importancia de CAPT como una herramienta complementaria a las clases tradicionales.

Se pueden encontrar las siguientes ventajas sobre el método tradicional de educación: el estudiante puede practicar la pronunciación con un computador en cualquier parte, no solo en la sala de clases; el estudiante puede practicar la misma lección las veces que desee, hasta quedar conforme con su

pronunciación; los estudiantes no son expuestos a situaciones estresantes y/o embarazosas frente a sus compañeros de clase por no tener una buena pronunciación; el profesor puede preparar lecciones personalizadas para sus estudiante poniendo énfasis en sus áreas más débiles; además, es una buena herramienta autodidacta para todas las personas que deseen aprender un segundo idioma y no cuentan con un profesor calificado por diferentes razones.

Aunque los beneficios que presenta la enseñanza asistida por computadoras son muchos, no ha estado exenta de polémicas. La principal crítica hecha por los profesores es que no confían en el *feedback* que entrega el sistema, ellos ponen en duda la capacidad del sistema de emitir un juicio preciso capaz de emular la opinión de un profesor.

2.3. Reconocimiento Automático de Voz (ASR).

Un ASR (*Automatic Speech Recognition*) básicamente convierte una señal acústica obtenida con un micrófono o mediante otra fuente en una secuencia de palabras. Las palabras reconocidas pueden ser ocupadas en diferentes aplicaciones o pueden servir como entradas a un proceso lingüístico más avanzado.

Las diferentes aplicaciones para las cuales se utiliza el ASR dan origen a diferentes características o parámetros tales como: tipo de habla; estilo de habla; enrolamiento y vocabulario entre otras, en la Tabla 1 [Ravest P., 2009] se pueden ver parámetros típicos para caracterizar la capacidad de un ASR.

Tabla 1: Características o parámetros típicos de un ASR.

Parámetros	Descripción
Tipo de habla	Palabras aisladas a dialogo continuo
Estilo de habla	Dialogo leído a dialogo espontaneo
Enrolamiento	Dependiente de locutor a independiente de locutor
Vocabulario	Pequeño (< 20 palabras) a grande (> 20.000 palabras)
Modelo de Lenguaje	Estado finito a sensible al contexto
SNR	Alto (>30dB) a bajo (<10dB)
Transductor	Micrófono con cancelación de ruido a teléfono

El proceso comienza cuando la señal de voz digitalizada es transformada en un conjunto útil de características a una tasa fija. Estas son utilizadas para buscar la secuencia más probable utilizando restricciones impuestas por los modelos acústicos y de lenguaje. En las siguientes secciones se abordara con profundidad estos temas, por el momento la Figura 1 muestra los principales componente de un ASR.

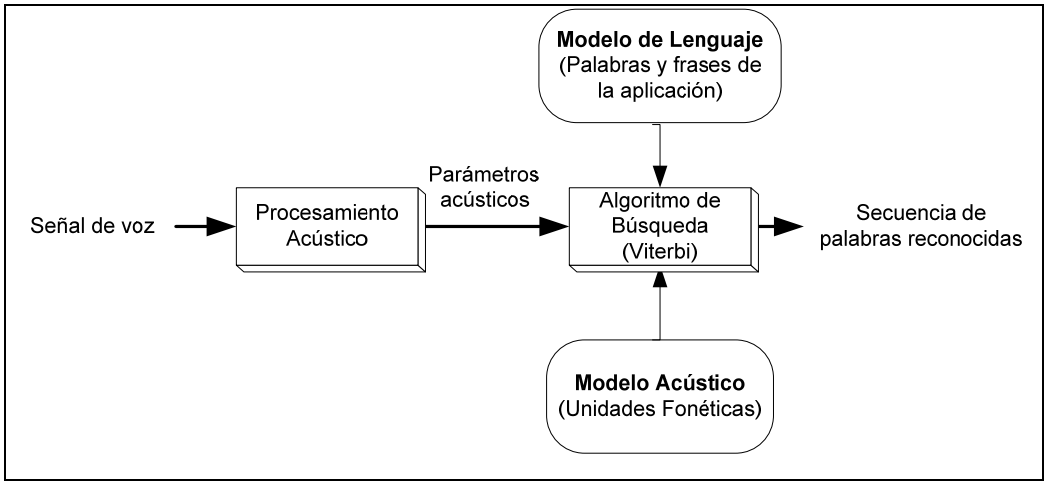

Figura 1: Diagrama que describe el funcionamiento de un ASR

2.3.1. Formulación Matemática del problema de Reconocimiento de Voz.

Puede formularse matemáticamente la tarea de reconocimiento de voz utilizando un enfoque estadístico [Jelinek F., 1998], lo cual permite una descomposición del problema en unidades más simples de manejar.

A partir de una señal acústica, se obtiene una secuencia de vectores de parámetros que llamaremos vector de observación $O = \{o_1, o_2, \ldots, o_T\}$. En base a estos parámetros el ASR decidirá qué palabras fueron pronunciadas. Por otro lado existe un vocabulario fijo que dependerá de la aplicación específica en que se usara el ASR, este vocabulario está compuesto por una secuencia de palabras $W = \{w_1, w_2, \ldots, w_T\}$.

Si $P(W/O)$ es la probabilidad que la secuencia de palabras W haya sido pronunciada dado que se observó la secuencia de vectores de observación O, y asumiendo que la ocurrencia de una secuencia de vectores de observación O, es igualmente probable, $P(O)$ es igual para todo O [Jelinek F., 1998], el problema de maximización se transforma en encontrar la mejor secuencia de palabras $\widehat{W}$ que maximice la probabilidad $P(W/O)$.

$$\widehat{W} = argmáx_W P(W/O) \quad (1)$$

Ocupando el Teorema de Bayes se puede reescribir el término $P(W/O)$ de la ecuación (1) de la siguiente forma:

$$P(W/O) = \frac{P(W)P(O/W)}{P(O)} \tag{2}$$

Donde $P(W)$ es la probabilidad de que la secuencia de palabras W sea pronunciada, $P(O/W)$ es la probabilidad de que dada una secuencia de palabras W haya generado una secuencia de vectores de parámetros O, y $P(O)$ es la probabilidad promedio de que O sea observado. Como la maximización en (1) considera una única secuencia de observación O, utilizando (2) se puede reescribir como:

$$\widehat{W} = argmáx_W P(W)P(O/W) \tag{3}$$

La expresión $P(W)$ del argumento de maximización de la ecuación (3) representa la ocurrencia de clases (palabras), que es conocido como el modelo de lenguaje del sistema. El segundo término $P(O/W)$, se conoce como el modelo acústico del reconocedor de voz. Se describirán detalladamente estos modelos en secciones siguientes.

Para obtener las probabilidades se usa la etapa de entrenamiento. Aquí se estiman los parámetros que determinan las probabilidades de manera de maximizar las verosimilitudes del conjunto de señales.

2.3.2. Parametrización acústica.

Para realizar una parametrización de una señal de voz adecuadamente es necesario considerar dos características intrínsecas: la señal de voz es un proceso estocástico no-estacionario; existen variaciones temporales entre señales que tienen la misma información fonética.

Existen diferentes tipos de variabilidad temporal en las señales de voz. Variabilidad intra-locutor corresponde a la información acústico fonética que se extrae de la señal de voz variando las elocuciones de un mismo locutor. La variabilidad inter-locutor son las variaciones entre elocuciones de diferentes locutores. La variabilidad temporal en la voz, la dependencia de la fuente donde se genera la señal y las variaciones del ambiente son elementos que dificultan el reconocimiento de la voz. En la Figura 2 se muestra una señal de voz en el dominio del tiempo (figuras a) y c)) y en el dominio espectral (figuras b) y d)) de una misma frase, *the airport is so big*, para dos locutores distintos. El tiempo se representa en el eje horizontal en número de muestras (de 0 a 50000). En los espectrogramas el eje vertical corresponde a la frecuencia, de 0 a 8000 Hz, y la energía se representa en escala de grises (de blanco, menos intenso, a negro, más intenso). Finalmente en el dominio del tiempo el eje vertical denota la amplitud de la señal cuyo rango se sitúa entre -3000 y 3000. Aquí se pueden apreciar

estas diferencias que deben tomarse en cuenta a la hora de caracterizar las señales acústicas.

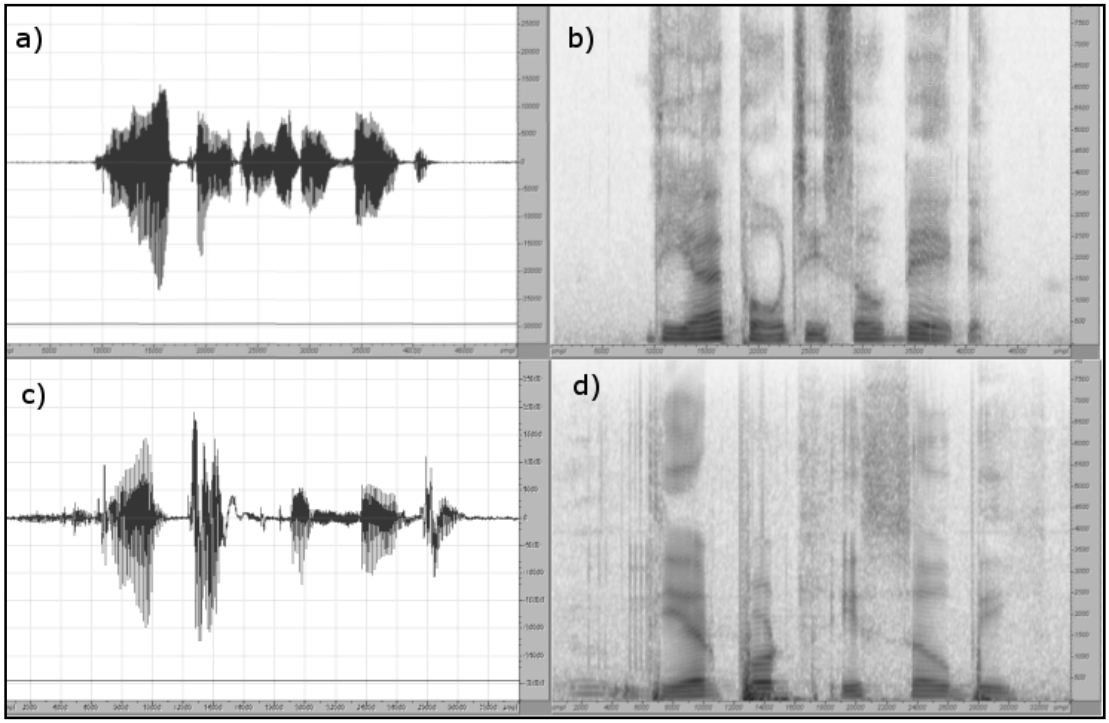

Figura 2 : Diferencias en el dominio temporal (izquierda) y espectral (derecha) para la frase "*the airport is so big*" pronunciada por dos locutores diferentes.

Para facilitar la extracción de parámetros, y considerar la variabilidad antes de extraerlos, se realiza un pre-procesamiento de la señal. En esta etapa se busca realzar la información de voz que contiene la señal y dejar todas las señales a analizar en condiciones similares para la extracción de características, algunas de las técnicas utilizadas son: eliminación de silencios; detección de inicio y fin; compensación de ruido aditivo y/o convolucional.

El pre-procesamiento de la señal se inicia con la conversión análogo-digital de la señal de voz, tarea realizada por interfaces telefónicas o por tarjetas de sonido, hardware de captura. A esta señal digitalizada se le aplica un filtro de inicio-fin [Lamel *et al*., 1981]. Básicamente elimina de la señal acústica los períodos de silencio antes del primer pulso de voz y después del último.

La siguiente etapa corresponde al proceso de segmentación, consiste en dividir la señal en segmentos, llamados *frames* o ventanas, que pueden ser considerados estadísticamente estacionarios. Para la segmentación típicamente se utilizan intervalos de 10 a 30 [ms] y un traslape de hasta 50% entre *frames* consecutivos. El enventanado de *Hamming* [Picone *et al.*, 1993] se utiliza para evitar las distorsiones en el análisis espectral que puede ocurrir en los límites de cada *frame*.

Corresponde realizar un análisis espectral por cada *frame*, primero la señal es procesada por la transformada discreta de Fourier (DFT *Discrete Fourier Transform*), luego se utiliza un banco de filtros debido a que la percepción auditiva no es capaz de distinguir frecuencias individuales, si no franjas de frecuencias. Además la percepción acústica no es lineal en el espectro de frecuencias de una señal de voz (entre un rango aproximado de 300[Hz] y 3400[Hz]), esto hace necesario utilizar una escala adecuada que concentre los filtros donde la capacidad de discriminación del oído sea mayor. Una de las escalas más utilizadas es la escala Mel. En la ecuación (4) se describe la transformación asociada a esta escala, para una frecuencia f :

$$MEL(f) = 2595 \times \log_{10}\left(1 + \frac{f}{700}\right) \quad (4)$$

Los filtros utilizados son un conjunto de funciones triangulares con ganancia unitaria para la frecuencia central, con superposición de 50% y un ancho de banda constante en escala Mel. Con la aplicación del banco de filtros se termina el pre-procesamiento de la señal.

Se acostumbra trabajar con los coeficientes cepstrales de la señal para la extracción de parámetros en reconocimiento de voz. El análisis de una señal de voz en el dominio cepstral o *cepstrum* contribuye a enfatizar la estructura de los formantes del tracto vocal, incluso con ruido. Los parámetros basados en el *cepstrum* se han convertido en uno de los métodos más usados para las técnicas de clasificación de patrones acústicos y ya se ha transformado en un estándar dentro del campo de reconocimiento de voz [Forsyth,1995].

Los coeficientes cepstrales en escala Mel (MFCC *Mel Frequency Cepstral Coefficient*) se calculan a partir de la energía contenida en cada filtro y mediante una transformada coseno discreta (DCT *Discrete Cosine Transform*). De esta manera se obtiene un vector de parámetros MFCC para cada *frame* de la señal, en otras palabras, la caracterización de una señal de voz resulta ser una secuencia de vectores de observación en el dominio MFCC.

2.3.3. Modelación acústica con HMM.

Un modelo de Markov consiste en una secuencia finita de estados conectados entre sí por probabilidades de transición. En el contexto del procesamiento de voz cada unidad temporal corresponde a un *frame,* debe evaluar la posibilidad de mantenerse en el estado actual o pasar al siguiente estado, esta decisión está determinada por las funciones de distribución de probabilidades de cada estado.

Los modelos ocultos de Markov (HMM *Hidden Markov Models*) han sido ampliamente utilizados en los sistemas de reconocimiento de voz y en especial los modelos de Markov de primer orden donde el estado actual de una señal depende solamente del estado anterior [Rabiner,1989;Jelinek,1997]. La salida de una secuencia de estados no es la secuencia misma, esta permanece oculta en el proceso, sin embargo se conoce que esa secuencia produjo un conjunto de parámetros acústicos de la señal.

En aplicaciones relacionadas con el procesamiento de la voz, se considera una topología izquierda-derecha donde solo se permiten transiciones al siguiente o al mismo estado, los saltos entre estados también están prohibidos. En la Figura 3 se puede observar un ejemplo de una estructura de este tipo.

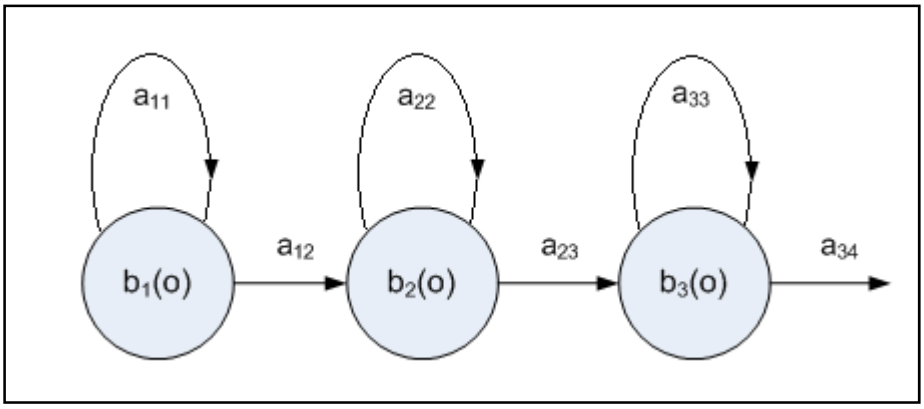

Figura 3: Estructura de un HMM con topología de izquierda-derecha.

Un HMM queda definido por las probabilidades de transición de estados, la función de distribución de probabilidades (f.d.p), y las probabilidades iníciales [Rabiner, 1989]. Las probabilidades de transición para un HMM con M estados debe cumplir con la siguiente restricción:

$$\sum_{j=1}^{M} a_{ij} = 1 \qquad \forall i = 1,\ldots,M \tag{5}$$

Donde a_{ij} es la probabilidad de estar en el estado j dado que el anterior estado fue i .

La distribución de probabilidad de que una observación haya sido generada por el estado j se presenta en la ecuación (6). En el caso de

reconocimiento de voz es usual utilizar poblaciones para modelar las f.d.p de estados. En este caso suponemos una población de G distribuciones normales independientes, cada una con un peso de probabilidad asignado, y restringido por la ecuación (7).

$$b_j\left(O_t\right)=\sum_{g=1}^{G}\left\{p_g\cdot\prod_{n=1}^{N}\left[\left(2\pi\right)^{-0.5}\cdot\left(Var_{j,g,n}\right)^{-0.5}\cdot e^{\left(-\frac{1}{2}\cdot\frac{O_{t,n}^{o}-E_{j,g,n}}{Var_{j,g,n}}\right)}\right]\right\} \tag{6}$$

$$\sum_{g=1}^{G}p_g=1 \tag{7}$$

Donde: j,g,n son los índices para el estado, la componente Gaussiana y el coeficiente del vector de observación, respectivamente; p_g corresponde al peso de probabilidad de la población g-esima; $O_t=\left[O_{t,1}^{o},O_{t,2}^{o},\ldots,O_{t,N}^{o}\right]$ es el vector de observación de la señal acústica de dimensión N en el instante t; $E_{j,g,n}$ y $Var_{j,g,n}$ son la media y varianza para un determinado modelo en el estado j, componente Gaussiana g y coeficiente cepstral n. Es importante mencionar que la matriz de covarianza de las Gaussianas es supuesta diagonal, es por esta razón que en el párrafo se hace mención a la varianza.

Los HMMs representan unidades fonéticas. En reconocimiento de voz las unidades que se utilizan son los tri-fonemas, estos se componen de una unidad fonética central mas dos segmentos de fonemas que preceden y suceden a la unidad central [Schwartz *et al.*, 1985].

Una palabra está formada por una secuencia de tri-fonemas, esto significa que cada palabra está formada por una secuencia de HMMs. Generalizando este concepto podemos decir que una frase también es una secuencia de HMMs. La Figura 4 representa esta idea con la frase "*The airport is so big*"

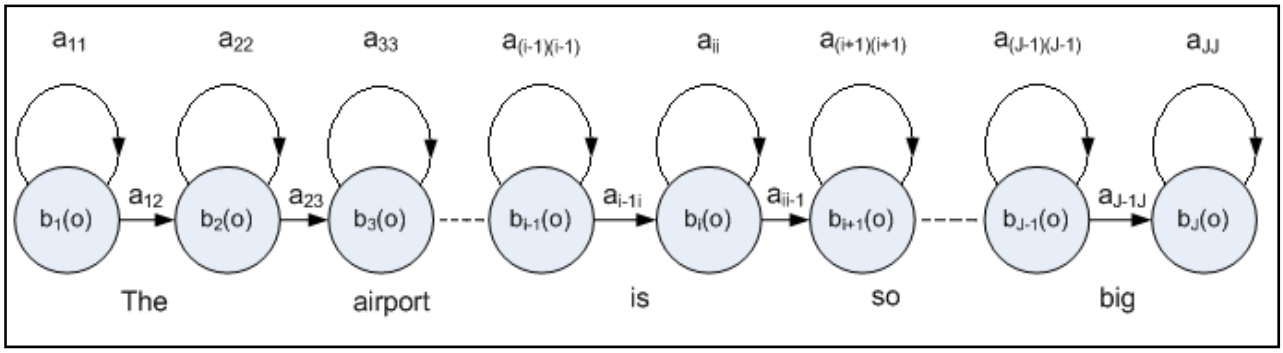

Figura 4: Secuencia de HMMs que forman la frase "*The airport is so big*".

Retomando la ecuación (3), la probabilidad que un vector de parámetros acústicos O haya sido generado por el HMM de la secuencia de palabras W queda dada por [Jelinek,1997;Rabiner,1989]:

$$P(O/W)=\sum_{S\in\Lambda}P(O,S/W)=\sum_{S\in\Lambda}P(S/W)\cdot P(O/S) \tag{8}$$

Donde $S=\left[S_1,S_2,\ldots,S_T\right]$ representa cualquier secuencia de estado dentro del conjunto Λ; el conjunto Λ son todas las posibles secuencias de estados que son capaces de generar la secuencia de vectores de parámetros acústicos O. Al descomponer la ecuación (8) según la descripción de los HMM se obtiene la siguiente ecuación:

$$P(O/W)=\left(A(O,S_1)\cdot\prod_t A(S_t,S_{t+1})\right)\cdot\left(\prod_t b_{S_t}(O_t)\right) \tag{9}$$

Reemplazando la ecuación (9) en (3), se obtiene:

$$\begin{aligned}\widehat{W} &= argmáx_{W,S}P(W)P(O/W)\\ &= argmáx_{W,S}P(W)\cdot\left(A(O,S_1)\cdot\prod_t A(S_t,S_{t+1})\right)\cdot\left(\prod_t b_{S_t}^{W}(O_t)\right)\end{aligned} \tag{10}$$

Con este resultado queda establecida la maximización para la primera parte de la ecuación (3). Queda por modelar la información proveniente del modelo de lenguaje.

2.3.4. Modelo de lenguaje.

El Modelo de lenguaje representado en la ecuación (3) por el término $P(W)$ entrega información a priori en la tarea de reconocimiento de voz. Para estimar el modelo de lenguaje existen distintos métodos que van desde un algoritmo de reglas gramaticales, hasta ser netamente una representación estadística del lenguaje.

Los modelos estadísticos de tipo M-grama son los más utilizados. Estos consideran que la probabilidad de ocurrencia de una palabra dentro de una sucesión de ellas está condicionada a la probabilidad de M-1 palabras anteriores. En la ecuación (11) se muestra lo que se obtiene de un modelo M-grama.

$$P(W_1, W_2, \ldots, W_N) = \prod_{i=1}^{N} P(W_i \mid W_{i-M+1}, \ldots, W_{i-2}, W_{i-1}) \qquad (11)$$

El criterio para la estimación de los parámetros que determinan el modelo de lenguaje es el estimador de máxima verosimilitud. En él se maximiza la probabilidad de observar las secuencias de algún conjunto de entrenamiento.

Los modelos estocásticos presentan un problema, no considera la probabilidad para las secuencias de palabras que no se encuentran en el conjunto de entrenamiento. Según la definición, estas probabilidades quedan en cero para aquellos casos en que no existe ocurrencia. El problema de generación de modelo de lenguaje es tratado con diversas técnicas. Por ejemplo, existe el modelo de lenguaje por palabras [Becchetti and Prina,1999; Laurila *et al.*, 1998]. Para medir la dificultad de los modelos de lenguaje existe la formula de perplejidad. Esta mide cuán bien el modelo representa a las secuencias del conjunto de *test*.

El algoritmo de Viterbi es un método para encontrar la secuencia de estados óptima que genera un vector de paramentos acústicos. Una secuencia de estados S determina inmediatamente una secuencia de HMMs, estos a su vez determinan la secuencia de palabras reconocidas W .

2.3.5. Algoritmo de Viterbi.

Para encontrar la secuencia de estados óptima que genera el vector de parámetros acústicos se necesita evaluar todas las posibles secuencias de estado para cada instante de tiempo en la señal de voz. Como se puede prever, esto es impracticable en un sistema computacional. Para realizar esta tarea de una manera más eficiente y minimizar la carga computacional, existe el algoritmo de Viterbi. La Figura 5 muestra la gráfica que genera el algoritmo de Viterbi operando sobre un modelo HMM de 8 estados con topología izquierda – derecha sin salto de estado.

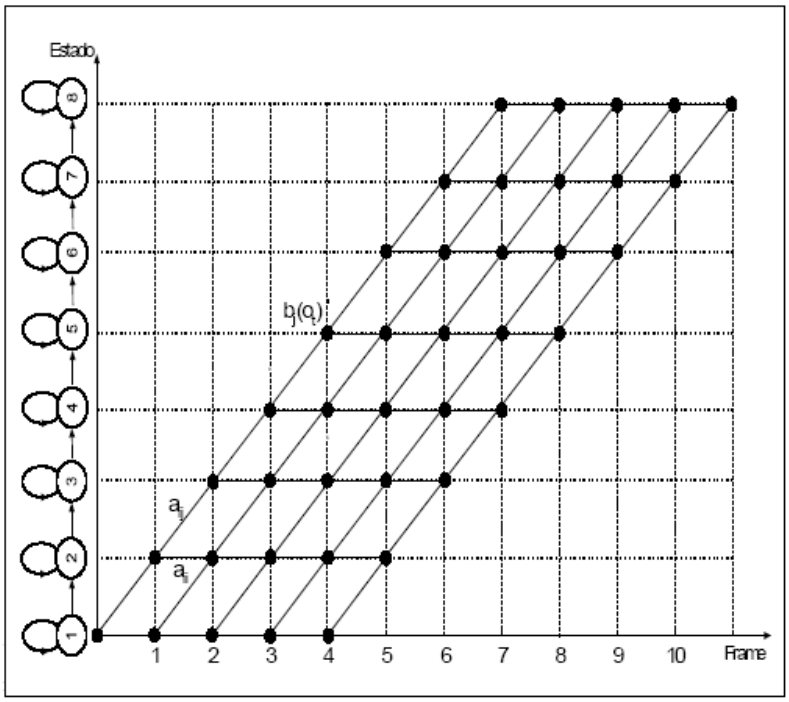

Figura 5: Representación gráfica del algoritmo de Viterbi.

El Algoritmo de Viterbi va optimizando a nivel local las secuencias de estado, Así reduce el campo de búsqueda ya que va descartando secuencias y genera un algoritmo viable desde el punto de vista computacional, en forma inductiva, se resuelve el problema de optimización global [Jelinek, 1997].

Definamos a $\delta_t(i)$ como la probabilidad de observar la secuencia de parámetros acústicos O hasta el tiempo t junto con la secuencia de estados más verosímil hasta t y que además, el estado s en t sea i. Es decir:

$$\delta_t(i) = máx_{s_1,s_2,\ldots,s_{t-1}} P(s_1,s_2,\ldots,s_{t-1},s_t = i,o_1,o_2,\ldots,o_t \mid \lambda_W) \quad (12)$$

Suponiendo la recursividad del algoritmo esto se traduce en:

$$\delta_t(i) = máx_{s_1,s_2,\ldots,s_{t-1}} P(o_t,s_t = i \mid s_1,\ldots,s_{t-1},o_1,\ldots,o_{t-1},\lambda_W) \cdot P(s_1,\ldots,s_{t-1},o_1,\ldots,o_{t-1},\lambda_W)$$

$$\delta_t(i) = b_i(o_t)\, máx_{s_1,s_2,\ldots,s_{t-1}} P(s_1,s_2,\ldots,s_{t-1},o_1,o_2,\ldots,o_{t-1},\lambda_W) \quad (13)$$

$$\delta_t(i) = b_i(o_t)\, máx_{s\in\Gamma}\left(a(s,i)\cdot\delta_{t-1}(s)\right)$$

Para llegar al cálculo de $\delta_t(i)$ se debe evaluar todos los posibles caminos para llegar a $s_t = i$. Estos posibles caminos están agrupados en el espacio Γ, por lo tanto Γ es un conjunto de secuencias de t estados, es decir $\Gamma \in \Re^t$. λ_W es el modelo de la secuencia de palabra W hasta el instante t. El término $a(s,i)$ determina la probabilidad de transición del último estado en la secuencia s al estado dado en t que es i. Asumiendo la recursividad del algoritmo, si buscamos la secuencia de estados más verosímil para llegar a s_t, la

secuencia anterior debe ser $\delta_{t-1}(s)$ donde s pertenece al conjunto Γ. Con esto, en forma recursiva, se llega a que $\delta_t(i)$ es la secuencia más probable de estados para llegar al tiempo t con el estado i [Jelinek F. ,1997].

Luego, para obtener la información del estado en el cual se está en el tiempo t se define la función $\Psi_t(i)$, que a medida que avanza el algoritmo guardará la información del estado óptimo. La Tabla 2 muestra el algoritmo de búsqueda de Viterbi, utilizado para obtener la secuencia de estados óptima y la verosimilitud máxima asociada a esta.

Tabla 2: Algoritmo de búsqueda de Viterbi.

1.-Inicialización

$$\delta_1(i) = \Pi_i \cdot b_1(o_1) \qquad i \in \Gamma$$

$$\Psi_1(i) = 0$$

2.-Recursión

$$\delta_t(i) = b_i(o_t)\, máx_{S\in\Gamma}\left(a(s,i)\cdot\delta_{t-1}(s)\right) \qquad i \in \Gamma \quad 2 \le \mathrm{t} \le \mathrm{k}$$

$$\Psi_t(i) = \arg máx_{S\in\Gamma}\left(a(s,i)\cdot\delta_{t-1}(s)\right)$$

3.-Terminación

Se determina la probabilidad de la secuencia de estados más verosímil y el último estado de dicha secuencia.

$$P_{\max} = máx_{i\in\Gamma}\left(\delta_k\left(i\right)\right)$$

$$\hat{s} = argmáx_{i\in\Gamma}\left(\delta_k\left(i\right)\right)$$

4.-Alineamiento

Finalmente se reconstruye la secuencia de estados más verosímil. Recordar que la función fue creada especialmente para esta etapa.

$$s_t = \Psi_{t+1}\left(s_{t+1}\right) t = 1,\ldots,k-1$$

2.4.Evaluación de Pronunciación con ASR.

La pronunciación y su posterior calificación corresponde a uno de los problemas principales que se ha intentado resolver, en el marco de la aplicación al aprendizaje de un segundo idioma y del reconocimiento de voz en general [Molina *et al* ,2009; Neumeyer L. *et al*, 1999; Moustroufas N. and Digalakis V.,2006]. En esta sección se aborda las diferentes categorías de evaluación vistos en la literatura, se pondrá especial énfasis en las evaluación de frases que es el tema central de esta tesis, también se trataran los temas de extracción de características del ASR, teoría de decisión de Bayes y múltiples clasificadores entre otros, temas que el lector debe conocer antes de introducirse en esta tesis.

2.4.1. Categorías de evaluación de pronunciación.

Debido a la importancia que tiene la evaluación de pronunciación para los sistemas CALL, se ha llevado a cabo investigación para fonemas [Kim Y. *et al*., sin año ;Witt and Young, 2000], palabras [Cincarek *et al.*,2008; Molina *et al* ,2009], frases [Franco H. *et al*., 1998; Neumeyer L. *et al*, 1999; Moustroufas N. and Digalakis V.,2006] y a nivel de locutor [Cucchiarini C. *et al*.,1997; Bernstein,1990; Minematsu,2004]. Prevalecen los enfoques basados en la tecnología de reconocimiento de voz. Características que describen la evaluación de pronunciación son extraídas desde la salida del reconocedor de voz, en la sección 2.4.2 se entrara más en detalle.

Cabe destacar que existen diferencias importantes entre la producción de una palabra aislada o una frase que deben ser consideradas a la hora su respectiva evaluación. El ser humano pronuncia las palabras de forma continua, y debido a la inercia de los órganos articulatorios, que no pueden moverse instantáneamente, se producen efectos coarticulatorios que deben ser tomados en cuenta. Por otro lado las variaciones producidas por la prosodia, hace que una palabra al principio de una frase sea diferente que cuando se dice en medio o al final, o que sea diferente dependiendo que es lo que le precede o le sigue.

La evaluación de la calidad de pronunciación ha sido ampliamente abordada en la literatura. En palabras aisladas se han obtenido correlaciones superiores a 0.7[Molina *et al* ,2009], medida de desempeño utilizada en la evaluación de pronunciación, es abordada en la sección 2.4.4. Por otra parte la correlación obtenida para frases oscila entre 0,4 y 0,7 dependiendo de múltiples

factores. Sorprendentemente se ha encontrado que la evaluación de frases en general se realiza como un todo y no se presta mayor atención a las palabras (unidades) que la forman salvo contadas excepciones. La evaluación de pronunciación de frases presenta dificultades extras que ameritan tratar este tema, estas dificultades son: a) La correcta pronunciación puede variar entre las palabras de la frase. b) La evaluación de una palabra aislada es diferente si está inmersa en una oración. c) Una o más palabras mal pronunciadas pueden significar que no se entienda la oración. Estos problemas en una o más palabras pueden ser enmascarados por el resto de la frase entregando una evaluación de pronunciación errada.

2.4.2. Evaluación automática de pronunciación basada en extracción de características.

Uno de los primeros acercamientos en la literatura a entregar una evaluación numérica a la pronunciación se encuentra en el trabajo de [Franco, 1997] y corresponde a la verosimilitud entregada por el algoritmo forzado de Viterbi. Este algoritmo obtiene el logaritmo de la probabilidad de una secuencia de observaciones en un modelo basado en HMM, en la sección 2.3 se entra más en detalle. Como mejora se normaliza la verosimilitud por duración de cada estado. Aunque mejores resultados se han obtenido con el método de la probabilidad a posteriori [Franco, 1997], con este método se han podido conseguir calificaciones mucho mas similares a las entregadas por calificadores humanos expertos (profesores de lingüística).

Una variante del método tradicional de evaluación de pronunciación mediante el algoritmo de Viterbi es simplemente no forzar la búsqueda a solo una palabra, si no realizar la búsqueda sobre palabras competidoras [Anguita J. et al, 2005; Hamid and Rashwan, 2004; Moustroufas and Digalakis, 2007; Molina *et al* ,2009]. En este caso el ASR debe entregar la palabra que más se asemeja a la pronunciada por el usuario. La gran dificultad está en generar las palabras competidoras para cada evaluación y graduar las escalas para las calificaciones. En [Molina *et al* ,2009] se propuso un método automático para la generación de estas palabras competidoras.

El ASR entrega la posibilidad de extraer un gran número de características o medidas de confiabilidad. El problema de la evaluación de pronunciación consiste en como mapear las medidas objetivas que entrega el ASR a medidas subjetivas que imitan la evaluación que entregaría un evaluador humano. Cada característica puede considerarse como una nota entregada por un clasificador. Algunas de las características o medidas de confiabilidad utilizadas en la literatura son: WDCM (*Word Density Confidence Measure*, [Kwan *et al.*, 2002]); POS (*Position in the N-best*, [Molina *et al.* ,2009]); PC (*Average phone confidence*) y NPC (*Time-normalized* PC) [Deshmukh, 2008] entre otras. Con las características se estiman las f.d.p *a priori* con una base de entrenamiento de señales y su evaluación subjetiva. En la Figura 6 se muestra un diagrama de bloques que ilustra este proceso.

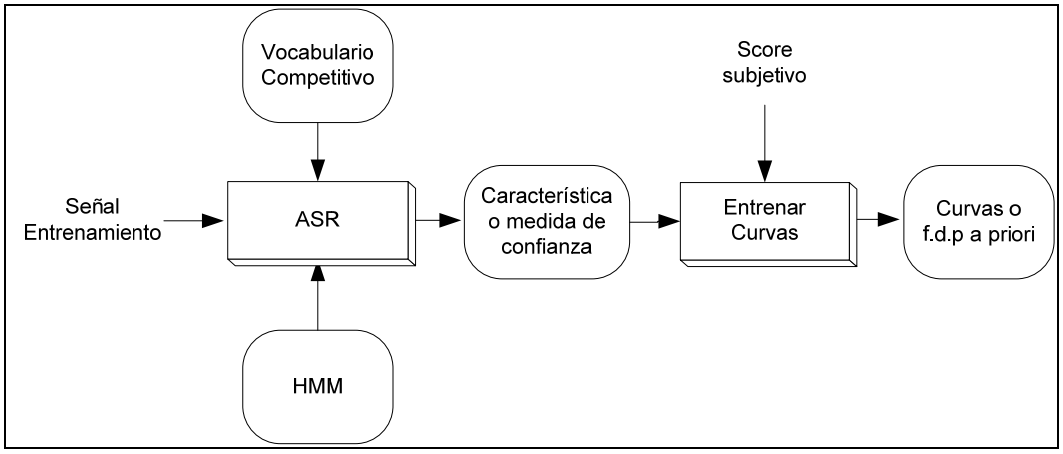

Figura 6: Diagrama de bloques del proceso de entrenamiento de las curvas a priori.

Una vez calculada las curvas a priori con la regla de Bayes es posible estimar una calificación objetiva, llamada en adelante *score* objetivo entregado por el ASR, a partir de una medida de confiabilidad. Este proceso se puede ver en la Figura 7.

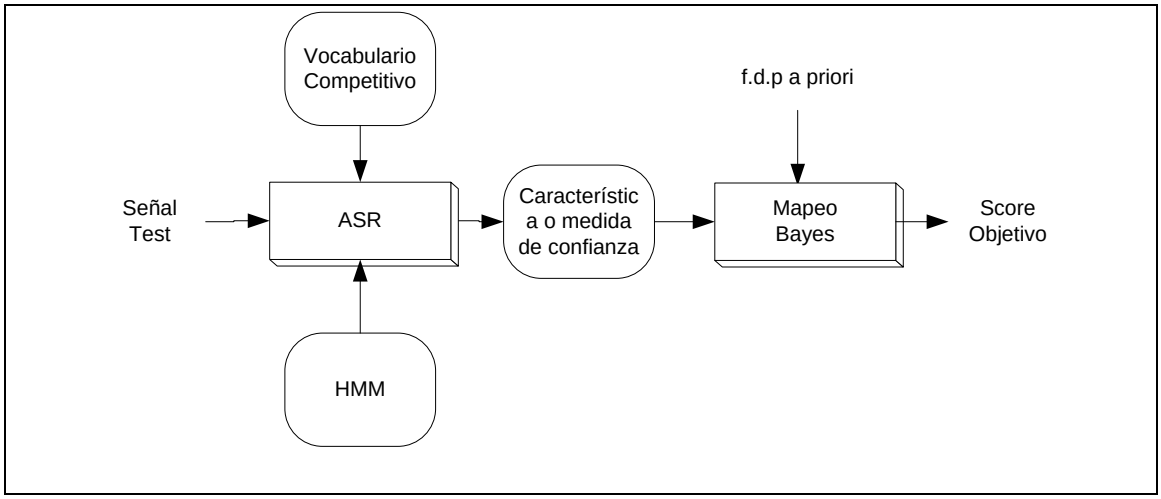

Figura 7: Diagrama de la generación del score objetivo utilizando mapeo de Bayes.

2.4.3. Introducción de errores de lenguaje materno.

Como una fuente importante de información para discriminar pronunciaciones correcta o incorrecta se utiliza la información acústica a priori de individuos parlantes no nativos. La naturaleza de estos errores depende del idioma nativo del individuo y pueden ser utilizados para mejorar los modelos utilizados por el sistema CAPT. La modelación de estos errores de pronunciación requiere la grabación de bases de datos sofisticadas que incorporen los diferentes tipos de errores para cada lenguaje en particular. [Bonaventura P. et al., - ;Cincarek *et al.*,2008].

Una manera alternativa de introducir errores del lenguaje materno sin un estudio tan acabado es intentar generalizar reglas para algunas pronunciaciones deficientes, esto puede resultar ser más inexacto pero una mejor opción si se busca un balance entre los beneficios esperados y los recursos que se disponen, un ejemplo de esta técnica se pude ver en [Molina *et al*., 2009] donde se incorporan variantes fonéticas que intentan incorporar información genérica del lenguaje materno sin implementar un estudio detallado de los errores de pronunciación.

2.4.4. Medidas de desempeño.

El término que más se repite en la literatura para evaluar la pronunciación de una palabra, frase o locutor es la correlación [Molina *et al* ,2009; Moustroufas N. and Digalakis V.,2006; Cucchiarini C. *et al*.,1997]. El

coeficiente de correlación es un índice que mide la relación o dependencia que existe entre dos variables, también sirve para valorar el grado de ajuste de los puntos a una línea recta.

El coeficiente de correlación de Pearson es, quizá, el más utilizado para estudiar el grado de relación lineal existente entre dos variables. Se suele representar por r :

$$\mathrm{r} = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y} \tag{14}$$

Donde σ_{xy} es la covarianza de (x, y), σ_x y σ_y son las desviaciones estándar de las variables x e y respectivamente.

El coeficiente de correlación de Pearson toma valores entre -1 y 1: un valor de 1 indica una relación lineal perfecta positiva; un valor de -1 indica una relación lineal perfectamente negativa; un valor de 0 indica una relación lineal nula, pero esto no necesariamente implica que las variables sean independientes, pueden existir relaciones no lineales entre las dos variables.

Para el caso de la evaluación de pronunciación las variables utilizadas son la evaluación subjetiva realizada por un evaluador calificado y la evaluación entregada por el sistema de evaluación de pronunciación. Lo que se busca es

estudiar la relación entre estas variables. Mientras más cerca este la correlación de alcanzar el valor 1 significa que las evaluaciones del sistema y de los expertos son más parecidas.

En la literatura dependiendo de los niveles de clasificación, notas, que van desde dos niveles (bueno o malo) hasta graduaciones de 10 niveles (donde 1 es una mala pronunciación y 10 es la pronunciación nativa de un idioma) y el tipo de evaluación, ya sean palabras, frases o locutor los rangos de correlaciones varían. En particular para frase el rango de correlación que se observa en las publicaciones oscila entre 0.4 y 0.7.

2.4.5. Teoría de decisión de Bayes.

Parte importante del trabajo está relacionada con la teoría de decisión de Bayes, por esta razón se le dedica esta sección. La teoría de decisión de Bayes es un acercamiento estadístico fundamental en el problema de reconocimiento de patrones [Duda, 1973]. La suposición fundamental es que todo problema de decisión se pude resolver en términos probabilísticos, y que todos los valores probabilísticos relevantes son conocidos.

Supongamos que se desea implementar un clasificador para separar dos tipos de estados diferentes y solo existe la posibilidad de pertenecer a uno de los dos estados. Definiremos el estado 1 como $s = s_1$ y el estado 2 como $s = s_2$. Se considera s como una variable aleatoria.

Las probabilidades *a priori* de que la característica pertenezca a un estado es lo primero que se calcula. Estas probabilidades dejan ver el conocimiento previo que se tiene de cuan probable es la pertenencia a algún estado sin siquiera haber realizado la observación. La probabilidad *a priori* de que la característica pertenezca al estado 1 es $P(s_1)$ y al estado 2 es $P(s_2)$, estas probabilidades son estrictamente positivas y suman 1. Como la única información que se permite utilizar es el valor de las probabilidades *a priori* ya que se está obligado a tomar una decisión sobre el estado al que pertenecerá cierta característica sin poder observarla, es razonable utilizar la siguiente regla de decisión:

$$\begin{aligned} \textit{caracteristica} \quad &\in \quad s_1 \qquad si\ P(s_1) > P(s_2) \\ \textit{caracteristica} \quad &\in \quad s_2 \qquad\qquad \sim \end{aligned} \tag{15}$$

Que este procedimiento funcione bien va a depender del valor de las probabilidades *a priori*. Si $P(s_1)$ es mucho mayor que $P(s_2)$, la decisión de elegir s_1 será correcta en la mayoría de los casos. Si $P(s_1) = P(s_2)$, la probabilidad de escoger correctamente es de un 50%, sin importar si se elige s_1 o s_2.

En general, se puede tomar decisiones con más información. Cuando se conocen observaciones de los patrones obtenidas previamente, y si sabe a cuál estado pertenecen, es natural expresarlos en términos probabilísticos. Denominando a estos patrones como x, se consideran como variables aleatorias continuas, cuya distribución depende del estado. Sea $p(x|s_j), j=1,2$ la función de densidad de probabilidad condicional respecto al estado s_j para x, en otras palabras, es la densidad de probabilidad condicional para x dado el estado de procedencia s_1. Un ejemplo de funciones de densidad de probabilidad condicionales para un cierto fenómeno se puede apreciar en la Figura 8.

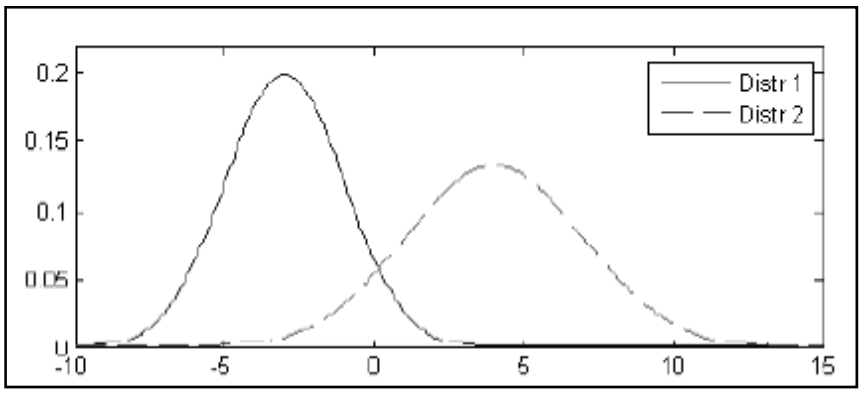

Figura 8: Funciones de densidad de probabilidad condicionales para cierto fenómeno Distr 1 corresponde a $p(x|\omega_1)$ mientras que Distr 2 a $p(x|\omega_2)$.

En esta etapa se supone que se conocen las probabilidades *a priori* $P(s_j)$ y las densidades condicionales $p(x|s_j)$, y se asume que se conoce el valor de una observación del fenómeno en cuestión, obteniendo como resultado

x . La influencia de esta medida puede dar una noción del estado al que pertenece dicha observación a través de la regla de Bayes.

$$p(s_j \mid x) = \frac{p(x \mid s_j) \cdot P(s_j)}{p(x)} \tag{16}$$

$$donde\ p(x) = \sum_{j=1}^{2} p(x \mid s_j) \cdot P(s_j)$$

La regla de Bayes muestra como observaciones de x , cambian la probabilidad *a priori* $P(s_j)$ a la probabilidad *a posteriori* $P(s_j \mid x)$. Si se tiene una observación x para la cual $P(s_1 \mid x)$ es mayor que $P(s_2 \mid x)$, uno se inclinaría por decidirse por el estado de procedencia s_1 ya que esta decisión minimiza la probabilidad de error. Por lo tanto se tiene la siguiente Regla de Decisión de Bayes que minimiza la probabilidad de error:

$$decidir\ \ s_1\ \ si\ \ p(s_1 \mid x) > p(s_2 \mid x) \tag{17}$$

$$decidir\ \ s_2 \qquad \sim$$

Aquí se enfatiza el rol de las probabilidades *a posteriori*. (17) se puede expresar en términos de las probabilidades *a priori* y de la densidad de probabilidad condicional. Al tomar la decisión el termino $p(x)$ es irrelevante ya que básicamente es un factor de escala que asegura que $P(s_1 \mid x) + P(s_2 \mid x) = 1$. Eliminando este factor de escala, se obtiene la siguiente regla de decisión, totalmente equivalente a (17).

$$\textit{decidir} \quad s_1 \quad \textit{si} \quad p(x \mid s_1) \cdot P(s_1) > p(x \mid s_2) \cdot P(s_2) \tag{18}$$

$$\textit{decidir} \quad s_2 \qquad\qquad \sim$$

Hay que tener en cuenta algunos casos particulares. Por ejemplo, si para algún x, $p(x \mid s_1) = p(x \mid s_2)$, la observación particular no ofrece ninguna información del estado del cual proviene. En estos casos, la decisión recae totalmente en las probabilidades *a priori*. Por otro lado, si $p(s_1) = p(s_2)$, luego los estados son equiprobables *a priori*. En este caso, la decisión se basa completamente en $p(x \mid s_j)$, la verosimilitud de s_j respecto a x. En general, ambos factores son importantes al momento de tomar una decisión, y la decisión de Bayes los combina para lograr una mínima probabilidad de error.

2.4.6. Sistemas de múltiples clasificadores (MCS).

Usar más de una medida de confiabilidad o característica puede ser interpretado como un problema en el contexto de los sistemas de múltiples clasificadores (MCS *Multiple Classifier System*). En MCS el objetivo es no depender de un único clasificador en la toma de decisiones. En lugar de ello todos los clasificadores se utilizan en la toma de decisiones mediante una combinación de sus opiniones, en otras palabras se busca alcanzar un conceso entre los diferentes clasificadores. El problema de reconocimiento de patrones utilizando MCS ha sido ampliamente discutido y utilizado en diferentes áreas, destacan el reconocimiento de patrones en escritura [Xu, 1992; Kitttler, 1998] y en datos multimedia como voz e imagen [Fumera and Roli, 2005; Kitttler, 1998].

Las topologías de combinación de clasificadores se pueden agrupar en tres categorías [Ranawana & Palade, 2006]: serie o cascada; paralelo; y jerárquico o hibrido. En la fusión en serie, los resultados obtenidos por cada clasificador son la entrada al siguiente, hasta que la decisión final es obtenida por el último clasificador de la cadena. En la fusión en paralelo, todos los clasificadores operan en paralelo sobre los datos de entrada, y los resultados de todos son fusionados con algún método con el fin de obtener una decisión en consenso. En la combinación jerárquica se utilizan los métodos de cascada y paralelo.

Los métodos de fusión en paralelo son los más utilizados en MCS y en particular los que usaremos en esta tesis. Los MCS en paralelo son combinados

generalmente en dos niveles [Mak *et al.*, 2003]: nivel abstracto (*abstract level)* y nivel de características (*feature level)*. En *abstract level*, cada clasificador realiza una decisión y posteriormente son combinadas. En el esquema *feature level* la fusión se realiza con la característica o *feature* de cada clasificador.

En la

Figura **9** y Figura 10 se utilizan conceptos vistos en las secciones anteriores para ilustrar estas ideas en el marco de la evaluación de pronunciación, C_j es el clasificador j, donde $1 \le j \le J$ y J es el número total de clasificadores, WF_j es la característica o *feature* correspondiente a cada clasificador j.

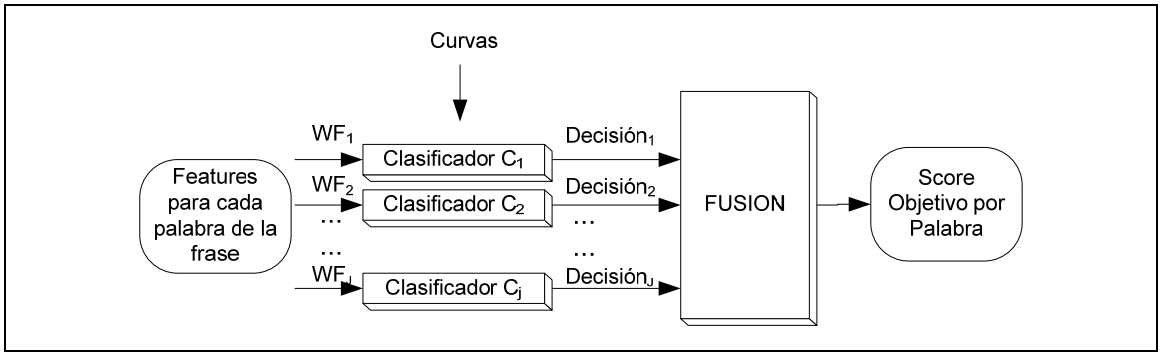

Figura 9: Fusión MCS en *abstract level*.

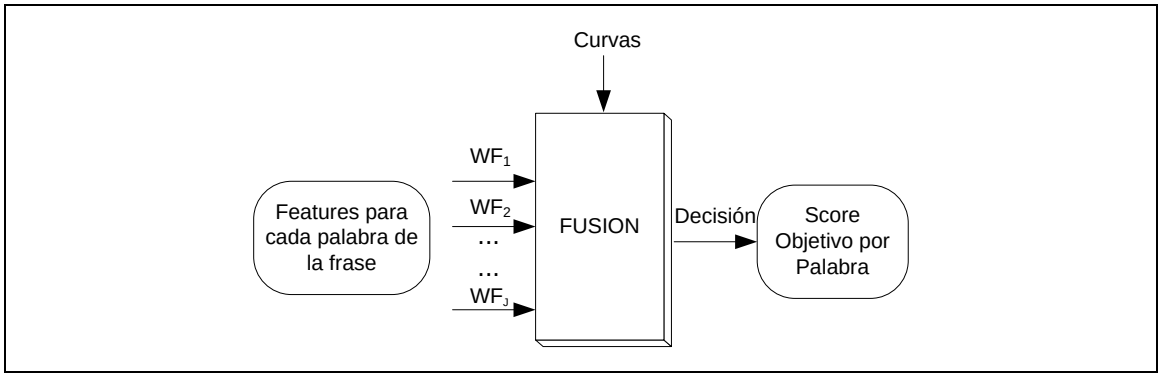

Figura 10: Fusión MCS en *feature leve.*

Teóricamente la vía óptima para la fusión de clasificadores es mediante la regla de Bayes [Duda and Hart,1973;Kittler *et al.*,1998]

$$\begin{aligned} D(O) &= argmax_{C_m} P(C_m / \overrightarrow{WF}_j(O)) \\ &= argmax_{C_m} \left\{ \frac{P(\overrightarrow{WF}_j(O) / C_m) P(C_m)}{P(\overrightarrow{WF}_j(O))} \right\} \end{aligned} \tag{19}$$

Donde $D(O)$ es la decisión final que corresponde a la señal O, $\overrightarrow{WF}_j(O) = \left[WF_1(O), WF_2(O), \ldots, WF_J(O) \right]$ es el set de J características utilizadas, C_m corresponde a la clase m, donde $1 \le m \le M$ y M es el número total de clases.

Teóricamente, el error de clasificación es minimizado por (19). Sin embargo, $P(\overline{WF}_j(O)/C_m)$ es una distribución de probabilidad multivariable y para su estimación es necesario una gran cantidad de datos para obtener un resultado confiable [Kittler *et al*., 1998]. Este problema puede ser simplificado si es calculado en términos de distribuciones de probabilidad por cada clasificador individual. Las técnicas más clásicas para la simplificación bayesiana son [Kittler *et al*., 1998; Kuncheva, 2001 y 2002]: regla del producto; máximo; mínimo; promedio; y votación (MVR Majority Vote Rule). Promedio y MVR son las más usadas [Kittler & Alkoot, 2003; Fumera & Roli, 2005].

2.4.7. Aporte.

Los aportes de esta tesis se enmarcan en una nueva propuesta para la evaluación de pronunciación de frases. En la literatura no se ha tratado la evaluación de frases considerando las unidades que la conforman, las evaluaciones de pronunciación encontradas son de palabras aisladas, locutor, párrafos o frases. La forma de evaluar las frases es considerándolas como un todo y a lo más se marcan algunos errores dentro de ellas. Esta propuesta considera evaluar cada una de las palabras que forman las frases y con ello poder obtener una evaluación objetiva para toda la frase.

Se presenta como tema novedoso incorporar diferentes criterios subjetivos para la evaluación de frases dependiendo de los diferentes niveles que se pretenda evaluar.

Se propone un método para generar un modelo competitivo automático para frases, este modelo se construye con las distancia entre palabras [Molina *et al*.,2008], distancia calculada a cada una de las palabras objetivo que forman la frase, palabras *target* en adelante, obteniendo palabras competidoras para cada palabra *target*. De la unión de las palabras competidoras se generan las frases competidoras. Además, en el modelo competitivo se incorporan variantes fonéticas de pronunciación e información de la lengua materna del estudiante sin necesidad de un estudio a priori de los errores más comunes para mejorar la exactitud de la evaluación de pronunciación. Estos aportes representan una nueva alternativa en el estado del arte de la evaluación de pronunciación.

CAPITULO III

3. Estimación automática de pronunciación de frases con ASR

La estimación de pronunciación de frases en el aprendizaje del segundo idioma es un problema mucho más complejo que la evaluación de la pronunciación de una palabra aislada. Cuando un evaluador experto califica la pronunciación de toda una frase narrada por un estudiante, este evaluador experto en el lenguaje que se intenta aprender (en este caso Inglés) puede aplicar diferentes criterios para evaluar la calidad de la pronunciación. Por ejemplo, una nota subjetiva asociada a una palabra aislada, $SubWordScore_W$, puede ser definida en base a la calidad de la producción acústica de los fonemas [Molina *et al*,2009]. En contraste, la nota subjetiva asociada a toda una frase, $SubSentenceScore_S$, puede depender de diferentes criterios, en este trabajo se presentan tres criterios que puede emplear un profesor: El mínimo $SubWordScore_W$ dentro de la frase; el promedio de los $SubWordScore_W$ dentro de la frase; y primera impresión. Estos tres criterios son modelados como tres posibles combinaciones de calificaciones objetivas de palabras, $ObWordScore_W$, para estimar una calificación objetiva para toda la frase, $ObSentenceScore_S$: el mínimo $ObWordScore_W$; el promedio $ObWordScore_W$; y, la moda de $ObWordScore_W$.

3.1. Criterios Subjetivos en Frases.

La evaluación subjetiva realizada por los evaluadores expertos se basa en la pronunciación fonética, la entonación y la expresión del significado correcto en la frase. De estos parámetros es la pronunciación fonética el más importante a la hora de asignar una calificación.

Considerando una frase objetivo m, frase *target* en adelante, $S_m = \left\{W_{m,1}, W_{m,2}, W_{m,3}, \ldots, W_{m,l}, \ldots, W_{m,L_m}\right\}$ compuesta por L_m palabras, donde $W_{m,l}$ denota la l^{th} palabra. Como se menciono antes, $SubSentenceScore_S$ puede ser el resultado de uno de los siguientes criterio aplicados por el evaluador experto:

Criterio Subjetivo 1 (SubCrit1): La evaluación subjetiva de pronunciación de la frase *target* S_m corresponde a la menor calificación subjetiva asociada a una de las palabras $W_{m,l}$ donde $1 \le l \le L_m$:

$$SubSentenceScore_{S_m} = \min_{1 \le l \le L_m} \left\{ SubWordScore_{W_{m,l}} \right\} \tag{20}$$

Criterio Subjetivo 2 (SubCrit2): La evaluación subjetiva de pronunciación de la frase *target* S_m corresponde al promedio percibido de las

calificaciones subjetivas asociada a las palabras $W_{m,l}$ de la frase, donde $1 \le l \le L_m$:

$$SubSentenceScore_{S_m} = promedio\ percibido_{1 \le l \le L_m} \left\{ SubWordScore_{W_{m,l}} \right\} \quad (21)$$

El "promedio percibido" es descrito como el promedio de la evaluación subjetiva de todas las palabras pronunciadas que componen la frase *target*. En este trabajo el promedio percibido de $SubSentenceScore_{S_m}$ es modelado como:

$$promedio\ percibido_{1 \le l \le L_m} \left\{ SubWordScore_{W_{m,l}} \right\} = \frac{1}{L_m} \sum_{l=1}^{L_m} SubWordScore_{W_{m,l}} \quad (22)$$

Criterio Subjetivo 3 (SubCrit3): La evaluación subjetiva de pronunciación de la frase *target* S_m corresponde a la primera impresión obtenida por el evaluador experto, este criterio es mucho más difícil de modelar en forma objetiva que los anteriores ya que el evaluador califica la frase solamente al escucharla una vez y a diferencia de los dos criterios este se aplica a la frase completa y no a cada palabra.

3.2. Calificación objetiva de frases como una combinación de calificaciones objetivas para palabras.

Si cada palabra $W_{m,l}$ en la frase S_m, donde $1 \le l \le L_m$, es asociada a una calificación objetiva $ObWordScore_W$, entonces la calificación objetiva para una frase $ObSentenceScore_{S_m}$ puede ser estimada empleando por ejemplo una de las siguientes combinaciones de métricas:

Combinación 1 de métricas objetivas (ObjMetrComb1): La evaluación objetiva de la pronunciación de la frase *target* S_m corresponde a la menor calificación objetiva asociada a una de las palabras $W_{m,l}$ de la frase, donde $1 \le l \le L_m$:

$$ObSentenceScore_{S_m} = \min_{1 \le l \le L_m} \left\{ ObWordScore_{W_{m,l}} \right\} \tag{23}$$

Combinación 2 de métricas objetivas (ObjMetrComb2): La evaluación objetiva de la pronunciación de la frase *target* S_m corresponde al promedio de las calificaciones objetivas asociadas a las palabras $W_{m,l}$ donde $1 \le l \le L_m$:

$$ObSentenceScore_{S_m} = \frac{1}{L_m}\sum_{l=1}^{L_m} ObWordScore_{W_{m,l}} \quad (24)$$

Combinación 3 de métricas objetivas (ObjMetrComb3): La evaluación objetiva de la pronunciación de la frase *target* S_m es igual a la moda estadística o a la más alta frecuencia de una calificación asociada a las palabras $W_{m,l}$ en la frase S_m, donde $1 \le l \le L_m$:

$$ObSentenceScore_{S_m} = \max_{1\le l\le L_m}\left\{frequency\left[ObWordScore_{W_{m,l}}\right]\right\} \quad (25)$$

Donde $frequency\left[ObWordScore_{W_{m,l}}\right]$ indica la repetición de las calificaciones de las palabras $ObWordScore_{W_{m,l}}$ en S_m.

3.3. Correspondencia entre la evaluación objetiva y subjetiva de frases.

En este punto la correspondencia entre el criterio de calificación subjetiva visto en la sección 3.1 y las combinaciones de calificaciones objetivas en la sección 3.2 son sencillos en los siguientes casos: SubCrit1 y ObjMetrComb1; y ; SubCrit2 y ObjMetrComb2. Sin embargo, SubCrit3, el cual

es definido como primera impresión, es más difícil de definir y se ha considerado en ObjMetrComb3 la moda como una posible forma de modelarlo.

En este trabajo la correspondencia entre la calificación del criterio subjetivo y el objetivo es estimado por medio de la correlación entre las calificaciones subjetivas proporcionadas por un calificador experto y las métricas objetivas entregadas por el ASR.

3.4. ASR basado en métricas objetivas con vocabulario competitivo y clases basadas en modelo de lenguaje.

En este trabajo la calificación objetiva de pronunciación asociado a la frase S_m, $SubSentenceScore_{S_m}$, es estimado como la combinación de las calificaciones objetivas de las palabras, $ObWordScore_{W_{m,l}}$, donde $W_{m,l}$ son las palabras que componen S_m, como se propone en la sección 3.2. Las calificaciones $ObWordScore_{W_{m,l}}$ son estimadas por un sistema de reconocimiento de voz continuo con clases basado en modelos de lenguaje. Dada una frase $S_m = \{W_{m,1}, W_{m,2}, W_{m,3}, \ldots, W_{m,l}, \ldots, W_{m,L_m}\}$, las clases competitivas serán $Class_{m,l} = \{Class_{m,1}, Class_{m,2}, Class_{m,3}, \ldots, Class_{m,l}, \ldots, Class_{m,L_m}\}$ donde cada clase $Class_{m,l}$ puede estar compuesta por la palabra *target* $W_{m,l}$, el vocabulario competitivo y una variante fonética de $W_{m,l}$. Tanto el vocabulario competitivo como las variantes fonéticas no requieren un análisis previo de los errores comunes y son generados de manera automática. El sistema de reconocimiento

de voz continuo es empleado para hacer competir la pronunciación correcta de la frase S_m con la pronunciación de las frases compuestas por palabras similares y variaciones fonéticas de las palabras *target*. La Figura 11 muestra el diagrama de bloques del esquema propuesto para la estimación de la calidad de pronunciación de frases basado en reconocimiento de voz continuo.

La lista *N-best* que resulta de la decodificación de Viterbi entrega un set de características para las palabras, $WF_{W_{m,l}} = [WF^1_{W_{m,l}}, WF^2_{W_{m,l}}, ..., WF^j_{W_{m,l}}, ..., WF^J_{W_{m,l}}]$ llamados *word features o confidence measure* en adelante, $WF^j_{W_{m,l}}$, donde $1 \le j \le J$ y J es el número de *word features* por palabra. Ejemplos de *word features* son la posición en la lista *N-best* donde la palabra *target* es contenida o WDCM (*word density confidence measure*) de los cuales se hablará más adelante. Cada *word feature* $WF^j_{W_{m,l}}$ entregado por la decodificación de Viterbi es mapeado a una calificación objetiva, $ObWordScore^j_{W_{m,l}}$ usando la regla de decisión de Bayes. A continuación, la calificación objetiva de pronunciación asociada a la palabra $W_{m,l}$, $ObWordScore_{W_{m,l}}$ es obtenida combinando $ObWordScore^j_{W_{m,l}}$ con $1 \le j \le J$ empleando técnicas de fusión de clasificadores. Finalmente, la calificación objetiva de pronunciación de la frase S_m es obtenida combinando $ObWordScore_{W_{m,l}}$, donde $1 \le l \le L_m$, empleando los criterios discutidos en la sección 3.2.

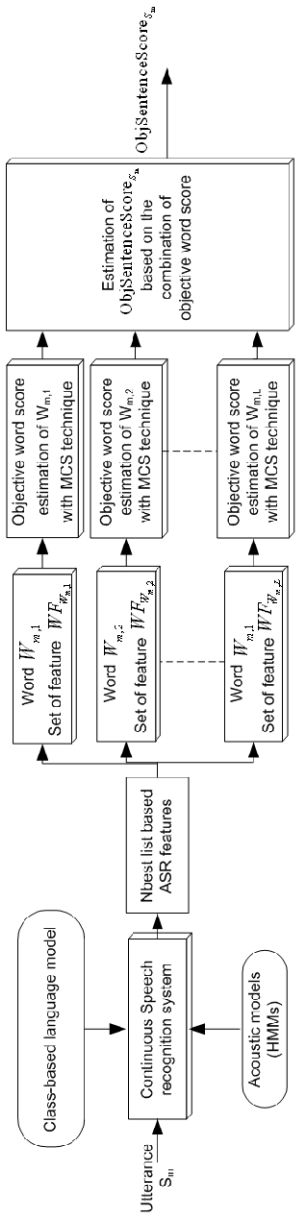

Figura 11: Diagrama de Bloques que muestra la generación del score objetivo para frases.

3.5. Generación automática de competidoras en las clases.

Cada palabra dentro de la frase *target* genera una clase que está compuesta por: a) palabra *target* $W_{m,l}$ con la correcta pronunciación; b) vocabulario competitivo similar a la palabra *target* con la correcta pronunciación; y c) variantes fonéticas de las palabras *target* de acuerdo al lenguaje materno. Como resultado, las clases $Class_{m,l}$ pueden ser representadas como:

$$Class_{m,l} = \left\{ W_{m,l}, CL_{m,l}, PV_{m,l} \right\} \tag{26}$$

Donde $CL_{m,l} = \left\{ CL^1_{m,l}, CL^2_{m,l}, CL^3_{m,l}, \ldots, CL^k_{m,l}, \ldots, CL^{K_{m,l}}_{m,l} \right\}$ es el vocabulario competitivo compuesto por las palabras $CL^k_{m,l}$, donde $1 \le k \le K_{m,l}$ y $K_{m,l}$ es el número de palabras en $CL_{m,l}$; y $PV_{m,l}$ es la variante fonética de la palabra *target* de acuerdo al lenguaje materno en la $Class_{m,l}$. No se requiere un análisis previo basado en los errores hechos por los estudiantes para alcanzar una integración eficiente del material didáctico en la tecnología ASR sin asistencia humana. Notar que la definición y generación de $Class_{m,l}$ intenta encontrar un *trade-off* entre la exactitud de la estimación de pronunciación y las limitaciones de la tecnología ASR: a mayor número de palabras competidoras y variantes fonéticas es más difícil el reconocimiento.

3.6. Generación automática de vocabulario competitivo.

El vocabulario competitivo $CL_{m,l}$, ayuda a forzar una competencia simultánea entre la pronunciación correcta y palabras fonéticamente similares lo cual es crucial para hacer exitosa la tecnología ASR en CAPT. Este trabajo emplea el mismo enfoque propuesto en [Molina *et al*, 2008]. Primero que todo, Se estima la distancia KL definida en [Sooful and Botha, 2002] entre la palabra *target* $W_{m,l}$ y palabras tomadas de un diccionario. El diccionario debe ser lo suficientemente completo y representativo del lenguaje *target* con el propósito de incluir una importante gama de distancias de palabras. Segundo, se almacena el vocabulario cuya distancia a la palabra *target* está en un intervalo definido por un mínimo, $D_{\min}^{CL}$, y un máximo, $D_{\max}^{CL}$. Entonces, el vocabulario en el intervalo $[D_{\min}^{CL}, D_{\max}^{CL}]$ es almacenado con respecto a la distancia a la palabra *target* y uniformemente seleccionados para reducir el número de palabras seleccionadas a MNCW (*Maximun Number of Competitive Words* – Número Máximo de Palabras Competidoras). $[D_{\min}^{CL}, D_{\max}^{CL}]$ es definido para considerar un *trade-off* entre la habilidad de discriminación resultante de la distancia entre el vocabulario competitivo de la palabra *target* y la precisión de la tecnología de reconocimiento de voz.

3.7. Generación automática de variantes fonéticas en español para palabras *target.*

Para mejorar la precisión en la calidad de evaluación de la pronunciación, se incorporan variantes fonéticas de la palabra *target*, $W_{m,l}$, de acuerdo al lenguaje materno del estudiante (en este caso español). $PV_{m,l} \subset \left\{PV^1_{m,l}, PV^2_{m,l}, PV^3_{m,l}, PV^4_{m,l},\right\}$, son incluidas en las clases competitivas $Class_{m,l}$. Esta estrategia intenta incorporar información del lenguaje materno del usuario, sin implementar un estudio detallado de los errores de pronunciación realizados por el estudiante. La generación de las variantes fonéticas se pueden ver en la Figura 12. La palabra *target* $W_{m,l}$ puede ser descompuesta de acuerdo a las reglas fonéticas del Inglés o del lenguaje materno del estudiante. En este caso la descomposición fonética en Inglés, entrega dos posibilidades: usando fonemas en Inglés; y reemplazando los fonemas en Inglés con el fonema más similar del lenguaje materno del estudiante, en este caso el Español (ver Tabla 3). En el caso de la descomposición de acuerdo a las reglas fonéticas del lenguaje materno del estudiante, también existen dos posibilidades: emplear fonemas en español; y, reemplazar los fonemas en español con las unidades fonéticas del Inglés más similares de acuerdo a la Tabla 4.

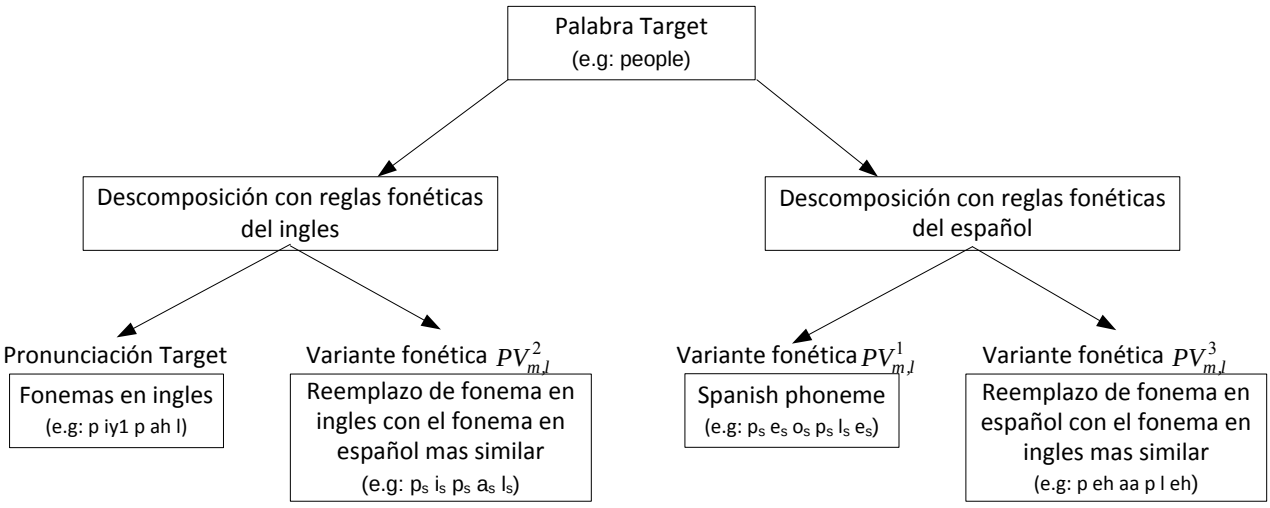

Figura 12: Descomposición en las diferentes variantes fonéticas de pronunciación.

Como resultado las variantes fonéticas que componen $PV_{m,l}$ en la $Class_{m,l}$ de acuerdo a 3.5 es definido como:

$PV^1_{m,l}$ - Descomposición de la palabra *target* $W_{m,l}$ de acuerdo a las reglas fonéticas del lenguaje materno del estudiante y fonemas en español.

$PV^2_{m,l}$ - Descomposición de la palabra *target* $W_{m,l}$ de acuerdo a las reglas fonéticas en Inglés reemplazando los fonemas en Inglés con los más similares del lenguaje materno del estudiante de acuerdo a la Tabla 3.

Tabla 3: Traducción fonemas Inglés a Español.

English	Spanish	English	Spanish
Ah, Ae,	a_s	Ow	o_s u_s
Aa, Ao	o_s	P	p_s

B, V	b_s	R	r_s
Ch, Sh	ch_s	S, Z, Th	s_s
D, Dh	d_s	Zh, Jh, Y	y_s
Eh, Er	e_s	T	t_s
F	f_s	Uh, Uw	u_s
G	g_s	Oy	$o_s\ i_s$
Ih, Iy	i_s	Aw	$a_s\ u_s$
Hh	j_s	W	$g_s\ u_s$
K	k_s	Ng	$n_s\ g_s$, n_s
L	l_s	Ay	$a_s\ i_s$
M	m_s	Ey	$e_s\ i_s$
N	n_s		

$PV^3_{m,l}$ - Descomposición de la palabra *target* $W_{m,l}$ de acuerdo a las reglas fonéticas del lenguaje materno del estudiante reemplazando los fonemas del lenguaje materno del estudiante con los fonemas del Inglés más similares de acuerdo a la Tabla 4.

Tabla 4: Traducción fonemas de Español al Inglés

Spanish	English	Spanish	English
a_s	Ah	m_s	M
b_s	B	n_s	N
ch_s	Ch	o_s	Aa
d_s	D, Dh	p_s	P
e_s	Eh	r_s	R
f_s	F	s_s	S
g_s	G	t_s	T
i_s	Ih	u_s	Uh
j_s	Hh	x_s	KS
k_s	K	y_s	Y
l_s	L		

$PV_{m,l}^{4}$ - Corresponde a la alternativa de pronunciación que agrupa las tres descomposiciones anteriores, incluye tanto las reglas fonéticas como los fonemas del idioma *target* y el idioma materno.

En rigor al ver el esquema de descomposición, existe la pronunciación *target* que corresponde a la descomposición de la palabra *target* $W_{m,l}$ de acuerdo a las reglas fonéticas y fonemas del Inglés, corresponde a la pronunciación correcta de la palabra *target* y por lo tanto no debe considerarse una alternativa de pronunciación.

Entonces, $PV_{m,l}^{i}$, donde $1 \le i \le 4$ es incluida en $PV_{m,l}$ si la distancia KL entre $PV_{m,l}^{i}$ y la palabra *target* $W_{m,l}$ es mayor o igual que $D_{\min}^{PV}$. El umbral $D_{\min}^{PV}$ define un *trade-off* entre la habilidad de discriminación resultante de la distancia entre la variante fonética y la palabra *target*, y la precisión de la tecnología de reconocimiento de voz. Es importante mencionar que la Tabla 3 y 4 fueron generadas por expertos en el idioma Inglés.

3.8. Clases basadas en modelos de lenguaje en ASR.

El reconocimiento continuo de voz es realizado utilizando clases basadas en modelos de lenguaje [Zhang J. *et al*, 2001]. Como se mencionó antes, las clases competidoras $Class_m$ son generadas a partir de las frases S_m. Entonces, los trigramas $Pr(Class_{m,p} | Class_{m,q}, Class_{m,\gamma})$ son definidos como sigue:

$$Pr(Class_{m,p} \mid Class_{m,q}, Class_{m,\gamma}) = \begin{cases} 1 & si\ q = p-1 \quad y \quad \gamma = q-1 \quad y \quad p \geq 3 \\ 0 & en\ otro\ caso \end{cases} \tag{27}$$

Donde $p = 2$, $Pr(Class_{m,p} \mid Class_{m,q}, Class_{m,\gamma})$ es reemplazado por el bigrama $Pr(Class_{m,p} \mid Class_{m,q})$:

$$Pr(Class_{m,p} \mid Class_{m,q}) = \begin{cases} 1 & si\ q = p-1 \quad y \quad p \geq 2 \\ 0 & en\ otro\ caso \end{cases} \tag{28}$$

Una manera sencilla para suavizar las frecuencias de trigramas es mediante la interpolación lineal de las frecuencias relativas de trigrama, bigrama y unigrama como se muestra en (29).

$$P(c_3 \mid c_1, c_2) = \lambda_3 f(c_3 \mid c_1, c_2) + \lambda_2 f(c_3 \mid c_2) + \lambda_1 f(c_3) \tag{29}$$

Donde $f(|)$ corresponde a la función de frecuencia relativa, y los pesos no negativos satisfacen $\lambda_3 + \lambda_2 + \lambda_1 = 1$ [HSU B., 2007]. En este trabajo el modelo de lenguaje no considera suavizado. De esta manera, debido a que la

probabilidad de trigramas ausentes es cero, la secuencia de palabras reconocidas es restringida a las secuencias que forman las frases bajo evaluación.

3.9. Extracción de características para palabras basada en la lista *N-best.*

Dada una palabra *target* $W_{m,l}$, el ASR con clases basadas en modelos de lenguaje permite extraer varias características, *features* o medidas de confiabilidad por palabra de manera eficiente. En este trabajo cuatro características entregadas por el ASR son empleadas: POS (*N-best position*); REC (*Recognition flag*) ; WDCM (*Word density confidence measure*) y LogWD (*Logarithmic word density confidence measure*).

3.9.1. POS (*N-best position*).

La posición en la lista *N-best* de la palabra $W_{m,l}$ en la frase *target* S_m, $POS_{m,l}$, que corresponde al índice de la hipótesis más probable donde $W_{m,l}$ es reconocida:

$$POS_{m,l} = \arg\max_{\gamma} \left\{ \left[Q\left(h_r\right) \right] \mid r \in E\left(W_{m,l}, H\right) \right\} \tag{30}$$

donde $Q\left(h_r\right) = P\left(h_r\right)^{\gamma} P\left(O / h_r\right)$; h_r es la r^{th} hipótesis en la lista *N-best* de Viterbi; $Q\left(h_r\right)$ es la probabilidad entregada por la búsqueda de Viterbi; $P\left(h_r\right)$

es la probabilidad del modelo de lenguaje de h_r; $P(O/h_r)$ es la probabilidad de observación de h_r; γ es el factor de escala del modelo acústico; $E(W_{m,l},H)$ corresponde a los índices de las hipótesis donde la palabra $W_{m,l}$ se encuentra; y finalmente, H denota todos los alineamientos de las hipótesis obtenidas de la decodificación de Viterbi.

3.9.2. REC (*Recognition flag*) .

Esta medida de confiabilidad binaria asociada a la palabra $W_{m,l}$ en la frase *target* S_m, denotada por $REC_{m,l}$ es definida como:

$$\mathrm{REC}_{\mathrm{m,l}}=\begin{cases}1 & \text{si} \quad \mathrm{W}_{\mathrm{m,l}} \subset \mathrm{h}_1 \\ 0 & \text{si} \quad \mathrm{W}_{\mathrm{m,l}} \not\subset \mathrm{h}_1\end{cases} \tag{31}$$

donde h_1 es la primera hipótesis en la lista *N-best* de Viterbi.

3.9.3. WDCM (*Word density confidence measure*).

Esta medida de confiabilidad de la palabra $W_{m,l}$ en la frase *target* S_m, $WDCM_{m,l}$, es definida como:

$$WDCM_{m,l} = \frac{\sum_{r \in E(W_{m,l},H)} Q(h_r)}{\sum_{l=1}^{N} Q(h_l)} \tag{32}$$

3.9.4. LogWD (*Logarithmic word density confidence measure*).

Logarithmic word density confidence measure de la palabra *target* $W_{m,l}$, $LogWDCM_{m,l}$, es definida como:

$$LogWDCM_{m,l} = \frac{\sum_{r \in E(W_{m,l},H)} \log\left(Q(h_r)\right)}{\sum_{l=1}^{N} \log\left(Q(h_l)\right)} \tag{33}$$

3.10. Estimación de la calificación objetiva de pronunciación de palabras.

Como en [Molina *et al*, 2008], la calificación objetiva de pronunciación de palabras, $ObjWordScore_{W_{m,l}}$, es estimada empleando MCS (Sistemas de múltiples clasificadores). Como se describió en la sección anterior, cuatro características o medidas de confiabilidad son evaluadas para cada palabra en la frase *target*: $POS_{m,l}$, $REC_{m,l}$, $WDCM_{m,l}$ y $LogWDCM_{m,l}$. El problema de la

evaluación de la calidad de pronunciación es modelada como un mapeo entre la medida de confiabilidad y la calificación $ObjWordScore_{W_{m,l}}$ que simula la opinión de un instructor humano. Supongamos que la nota subjetiva $SubjWordScore_{W_{m,l}}$ es cuantizada en M niveles, en este caso $M = 5$. Consecuentemente, cada medida de confiabilidad puede ser asumida como una nota entregada por un clasificador y cada nivel de nota subjetiva puede ser una clase. Considerando que O es la secuencia de vectores de observación correspondientes a la frase *target* S_m pronunciada por el estudiante. Usando la regla de Bayes, $ObjWordScore_{W_{m,l}}$, puede ser estimada como:

$$\begin{aligned} ObjWordScore_{W_{m,l}}(O) &= argmax_{C_m} P(C_m / \overrightarrow{WF}_{W_{m,l}}(O)) \\ &= argmax_{C_m} \left\{ \frac{P(\overrightarrow{WF}_{W_{m,l}}(O) / C_m) P(C_m)}{P(\overrightarrow{WF}_{W_{m,l}}(O))} \right\} \end{aligned} \tag{34}$$

donde $ObjWordScore_{W_{m,l}}$ es la decisión final para $W_{m,l}$ que corresponde a la señal O. Teóricamente el error de clasificación es minimizado por (34), $P(C_m)$ se asume uniformemente distribuida e igual a $\frac{1}{M}$. Por otro lado las f.d.p multivariables a priori $P(\overrightarrow{WF}_{W_{m,l}}(O) / C_m)$ y $P(\overrightarrow{WF}_{W_{m,l}}(O))$ requieren una cantidad inmanejable de datos de entrenamiento para ser estimados fidedignamente.[Kittler et al, 1998]. El problema es simplificado si la

maximización en (34) se expresa en términos de los cálculos realizados por los clasificadores individuales. Las técnicas clásicas para simplificar la regla de Bayes [Kittler et al, 1998; Kuncheva et al, 2001,2002] son: regla del producto; Regla del máximo; Regla del mínimo; regla del promedio; y; MVR (*Majority vote rule)*. En este trabajo se utiliza la Regla del Promedio y MVR.

3.10.1. Regla del Promedio.

El promedio es definido como:

$$ObjWordScore_{W_{m,l}}(O) = argmax_{C_m}\left\{\frac{1}{J}\sum_{j=1}^{J} P(C_m / \overline{WF}_{W_{m,l}}(O))\right\} \qquad (35)$$

$$= argmax_{C_m}\left\{\frac{1}{J}\sum_{j=1}^{J} \frac{P(\overline{WF}_{W_{m,l}}(O) / C_m)P(C_m)}{P(\overline{WF}_{W_{m,l}}(O))}\right\}$$

.

Como se menciono antes, $1 \le m \le M$ y M es el número total de posibles niveles de calidad de pronunciación y J es el número total de características o medidas de confiabilidad.

3.10.2. MVR *(Majority vote rule*).

Es un esquema que combina las salidas individuales de los clasificadores. Dado un *set* de clasificadores con decisiones individuales $ObjWordScore_{W_{m,l}}(O) = \left[ObjWordScore^{1}_{W_{m,l}}(O), ObjWordScore^{2}_{W_{m,l}}(O), \ldots, ObjWordScore^{J}_{W_{m,l}}(O) \right]$, $ObjWordScore^{j}_{W_{m,l}}(O)$ es la decisión entregada por el clasificador correspondiente a la característica *j* y la decisión final $ObjWordScore_{W_{m,l}}(O)$ será la calificación que reciba el mayor número de votos como consenso.

CAPITULO IV

4. Implementación, Discusión y resultados

4.1. Implementación.

Para llevar a cabo el trabajo propuesto en esta tesis fue necesario generar una base de datos adecuada para verificar la propuesta de esta tesis, ocupar una serie de programas desarrollado por el equipo del laboratorio y crear otros para generar los modelos necesarios, e idear una estructura de experimentos para probar los diferentes conceptos.

4.1.1. Base de Datos.

A diferencia de otros trabajos presentes en la literatura de evaluación de pronunciación de frases, aquí se plantea trabajar con las unidades que forman las frases: las palabras. Debido a esto se necesita una base de datos donde se le asigne una calificación a cada palabra. Este trabajo fue realizado por lingüistas expertos de la Universidad de Chile, quienes calificaron un total de 423 frases. Las frases fueron extraídas del sistema 2LL web diseñado por LPTV, ver Figura 13, fue revisado por un experto en el idioma Inglés y su fonética con el fin de lograr un set de datos balanceado fonéticamente.

Figura 13: Sistema 2LL web diseñado por LPTV

La base de datos está compuesta por las siguientes frases: “It was nice to see my relatives”; “I was so happy to see my parents”; “I missed my kiwi family”; “Lots of people play soccer”; “People live a healthy life here”; “New Zealand is a country in Oceania”; “She gently shows me my seat”; “I have never had this experience”; “To meet different people”; y , “The airport is so big”.

Cada frase fue pronunciada por 43 locutores, los cuales mostraron diferentes niveles de dominio del Inglés, las grabaciones se realizaron con

micrófonos de escritorio de bajo costo y en ambientes poco controlados y sin ningún cuidado especial o condiciones de laboratorio, entiéndase por computadores ubicados en las casas de los usuarios o en salas de clase.

El primer paso corresponde a seleccionar las frases desde el reconocedor del LPTV, esta selección corresponde exclusivamente a aspectos técnicos que no tiene nada que ver con la pronunciación del locutor, por ejemplo fueron eliminadas todas las señales saturadas, los audios cortados y las señales que eran ininteligibles debido al ruido ambiente, siete frases fueron descartadas por presentar los problemas mencionados aquí, con lo cual la base de datos queda conformada por 423 frases.

Las señales seleccionadas fueron transcritas en una planilla y empaquetadas en set de 50 señales y enviadas a los lingüistas, cada señal fue revisada por dos lingüistas, en caso de haber mucha discrepancia en la calificación un tercer lingüista contribuía al consenso. Cinco diferentes niveles de pronunciación fueron evaluados por los lingüistas, donde calificación 5 corresponde a una pronunciación correcta y una calificación 1 denota la peor pronunciación. Como resultado se obtuvieron 423 frases evaluadas palabra a palabra para obtener los criterios del promedio y el mínimo. La calificación del criterio de primera impresión fue obtenida al escuchar por primera vez y sin repetición la frase completa. En la Tabla 5 se ve un ejemplo de las evaluaciones entregadas por los lingüistas.

Tabla 5: Ejemplo de evaluaciones de frases.

Nombre	Palabras						First Imp	Min	Prom
T266.wav	The	airport	is	so	big				
	5	2	5	5	3		2	2	4
T267.wav	I	have	never	had	this	experience			
	5	4	4	3	4	3	3	3	3.8

Una vez obtenidas las señales fueron divididas en dos set uno de entrenamiento orientado a entrenar las curvas y otro set de *test* utilizado para evaluar el sistema mediante la correlación entre la nota objetiva entregada por el reconocedor y la subjetiva dada por los lingüistas.

4.1.2. Generación Modelos competidores.

Esta tesis tiene como génesis el trabajo realizado por [Molina *et al*,2008] para palabras aisladas y busca expandirlo a frases. Para ello se deben crear los modelos para el reconocedor propuestos en la sección 3.5, con la estructura $Class_{m,l} = \{W_{m,l}, CL_{m,l}, PV_{m,l}\}$.

4.1.2.1. Cálculo de la distancia.

La elección de las palabras competidoras se basa en la distancia K-L que existe entre la palabra *target* y un set de palabras escogido. El set de palabras ocupado en este trabajo es el diccionario de pronunciación CMU (*Carnegie Mellon University*), este diccionario cuenta con a lo menos 6.000

palabras y su transcripción, además fue ocupado en el trabajo de palabras aisladas en [Molina *et al*,*2008*].

Las frases *target* son separadas en palabras y se calcula la distancia entre las palabras *target* y el diccionario ocupando el programa "distancia palabras" el cual entrega un archivo de texto por cada palabra *target* con la distancia calculada al resto de las palabras del set.

4.1.2.2. Búsqueda de las palabras competidoras.

Como se explica en la sección 3.6, para escoger las palabras competidoras se escoge un rango de distancias $[D_{\min}^{CL}, D_{\max}^{CL}]$ y un número máximo de competidoras MNCW. El programa "gen modelo concurrente" genera las palabras competidoras para cada una de las palabras *target*. Este programa ordena las palabras por distancia, selecciona todas las palabras que están dentro del rango y selecciona x-espaciadamente las palabras dependiendo el número MNCW. Como resultado entrega un archivo de texto por cada palabra que contiene la palabra *target* y sus competidoras obtenidas con la configuración dada.

4.1.2.3. Generar los modelos para las frases.

Esta etapa consiste en agrupar las palabras *target* y sus competidoras para generar los modelos para las frases. Además con esta metodología se

pueden incorporan las variantes fonéticas $PV_{m,l}$ expuestas en la sección 3.7 de ser necesario junto con el umbral $D_{\min}^{PV}$ mostrado en el mismo capítulo.

Como resultado el programa "gen frases" entrega:

- Un archivo de texto "modelo_x.tbl" para cada frase que puede contener: la frase *target*; las frases competidoras originadas al agrupar las respectivas palabras competidoras correspondientes a las palabras *target*; la alternativa de pronunciación seleccionada con el umbral correspondiente. A continuación en la Tabla 6 se puede ver un ejemplo de este tipo de archivos: "modelo_1.tbl" donde está la frase *target* "*it was nice to see my relatives*" junto a 5 frases competidoras generadas al agrupar las distintas palabras competidoras para cada palabra *target*, la última frase del ejemplo incorpora la variante de pronunciación $PV_{m,l}^{3}$ con un umbral de pronunciación "3", las palabras que tengan una distancia K-L a la palabra *target* menor que "3" son reemplazadas por una palabra competidora correspondiente.

Tabla 6: Ejemplo de archivo modelo_1.tbl.

T1.wav it was nice to see my relatives
T1.wav tripling hands arrest predictably scandinavian bhopal coattails
T1.wav condition renamed claimants indicted handicapped denver reports
T1.wav wide drams inventory demonstration kobe argue hostility
T1.wav expansions jones kincaid spell calculation drifted du
T1.wav scales superb deteriorating makes visitor committees joseph
T1.wav scales was3 nice3 to3 see3 my3 joseph

- El Archivo "modelo_clases_demo.txt" donde cada clase corresponde a la palabra *target* y sus competidoras, se compone de dos columnas: la primera contiene la palabra que se desea clasificar y en la segunda se escribe la clase a la cual pertenece la palabra. A modo de ejemplo en la Tabla 7 se muestran las tres primeras clases que corresponden a las palabras *it, was y nice*.

Tabla 7: Ejemplo de archivo modelo_clases_demo.txt.

it c1	was c2	nice c3
tripling c1	hands c2	arrest c3
condition c1	renamed c2	claimants c3
wide c1	drams c2	inventory c3
expansions c1	jones c2	kincaid c3
scales c1	superb c2	deteriorating c3
	was3 c2	nice3 c3

4.1.2.4. Generación Modelo de Lenguaje.

Una vez generados los modelos, modelos macro y el archivo de clases se puede crear el modelo de lenguaje. El programa "modelo de lenguaje" entrega el archivo "modelo_lenguaje.mtp" que contiene el modelo de lenguaje.

4.1.2.5. Modelo acústico fonético.

El reconocedor utilizado en las pruebas considera un modelo acústico entrenado utilizando HTK y considerando como unidad básica el trifonema. Cada trifonema fue modelado con una topología de izquierda a derecha (*left-to-*

right) de tres estados sin transiciones de salto de estados, con ocho densidades Gaussianas. En el sistema, la frase reconocida corresponde a la primera hipótesis (la más probable) dentro de la lista de las N-mejores obtenida por la decodificación de Viterbi. Se consideró un máximo de 10 hipótesis en la lista de Viterbi.

Los modelos acústicos del Inglés fueron entrenados con el *Corpus* CSR-I WSJ0 [Garofalo *et al*., 1993]. Las señales de voz de esta base de datos fueron grabadas con micrófonos de alta calidad y con una tasa de muestreo de 16 [kHz]. Las 20.055 frases que la componen fueron utilizadas para el entrenamiento de los modelos acústicos. Para el entrenamiento de los modelos acústicos del español se utilizó la base de entrenamiento LATINO 40 [LDC, 1995]. Las señales de esta base de datos están compuestas por habla continúa de 40 locutores nativos de Latinoamérica, cada locutor lee 125 frases de un periódico en español.

Al entrenar el HMM si el número de repeticiones de un trifonema en la base de entrenamiento es menor a un umbral entonces el trifonema es reemplazado por el monofonema correspondiente. En este caso el mínimo número de repeticiones en la base de entrenamiento permitido para trifonemas es 5, este valor se obtuvo de los resultados mostrados en [Molina *et al.* ,1998].

4.1.2.6. Binarización del modelo acústico.

El modelo acústico entrenado contiene modelos para todos los trifonemas presentes en la base de entrenamiento, pero sólo algunos trifonemas serán necesarios para el proceso de reconocimiento por lo cual es posible generar un HMM reducido con el cual se acelera la etapa correspondiente a la búsqueda de hipótesis.

El programa "lectura HMM" extrae del HMM sólo los modelos necesarios generando un HMM reducido el cual además se guarda como un archivo binario, este programa también binariza el archivo de trifonemas guardándolo con el nombre de "dic_tri.bin".

4.1.2.7. Entrenamiento de curvas de Bayes.

En esta etapa el reconocedor cuenta con todo lo necesario para funcionar. Ahora se deben entrenar las curvas de Bayes que permitirán la discriminación entre las diferentes calificaciones. Para esto se toma la base de datos de entrenamientos que consta de 212 señales y se pasa por el reconocedor. El reconocedor entrega un archivo de texto llamado "archivo de detalles" que contiene información de las señales. De este archivo se rescatan los valores de las medidas de confiabilidad, para esto se ocupa el programa "características" que genera archivos independientes por cada característica, en este caso POS (*N-best position*); REC (*Recognition flag*) ; WDCM (*Word density confidence measure*) y LogWD (*Logarithmic word density confidence measure*).

Para calcular las probabilidades se ocupa matlab, se crearon rutinas para automatizar el proceso, la función "todos_histo_mibase.m" agrupa todas las rutinas y entrega como resultado un archivo de texto por medida de confiabilidad con las probabilidades de cada calificación en determinados rangos de valores de las medidas de confiabilidad utilizando la regla de Bayes.

Finalmente los archivos con las probabilidades son binarizados con el programa "Genera BBCM", a estos archivos les llamaremos curvas de Bayes y son reemplazados por los archivos existentes en el reconocedor.

4.1.2.8. Test.

Con las curvas entrenadas y con los archivos binarizados reemplazados se ocupa la base de test en el reconocedor. De este proceso se vuelve a ocupar el archivo de texto "Archivo de detalles", esta vez nos interesan las calificaciones que le otorgo el reconocedor a las diferentes señales.

Se toman las calificaciones y con una planilla de cálculo se calcula la correlación que existe entre las calificaciones objetivas entregadas por el reconocedor y las calificaciones subjetivas dadas por los lingüistas.

4.1.3. Estructura de experimentos.

En este trabajo, se busca encontrar configuraciones que modelen de buena forma los tres criterios subjetivos expuestos en la sección 3.1. Debido a la gran cantidad de variables a ajustar para los diferentes criterios (Nº de competidoras, umbrales, variantes fonéticas de pronunciación) resulta complejo hacer todas las combinaciones posibles, por esto se realizo una estructura de experimentos que progresa en serie y va modificando solo una variable en cada bloque, este esquema se observa en la Figura 14. En el primer bloque se mantiene constante el umbral de distancias $[D_{\min}^{CL}, D_{\max}^{CL}]$ tomado de [Molina *et al,2008*] y se varia el número de competidoras, el experimento que obtenga una mayor correlación definirá para el segundo bloque el número de competidoras y se variara el umbral de distancias $[D_{\min}^{CL}, D_{\max}^{CL}]$, luego el experimento con mayor correlación pasara a la siguiente etapa y así sucesivamente.

Esta estructura de experimento se repita para cada uno de los criterios subjetivos buscando la configuración que entregue una mejor correlación para dicho criterio.

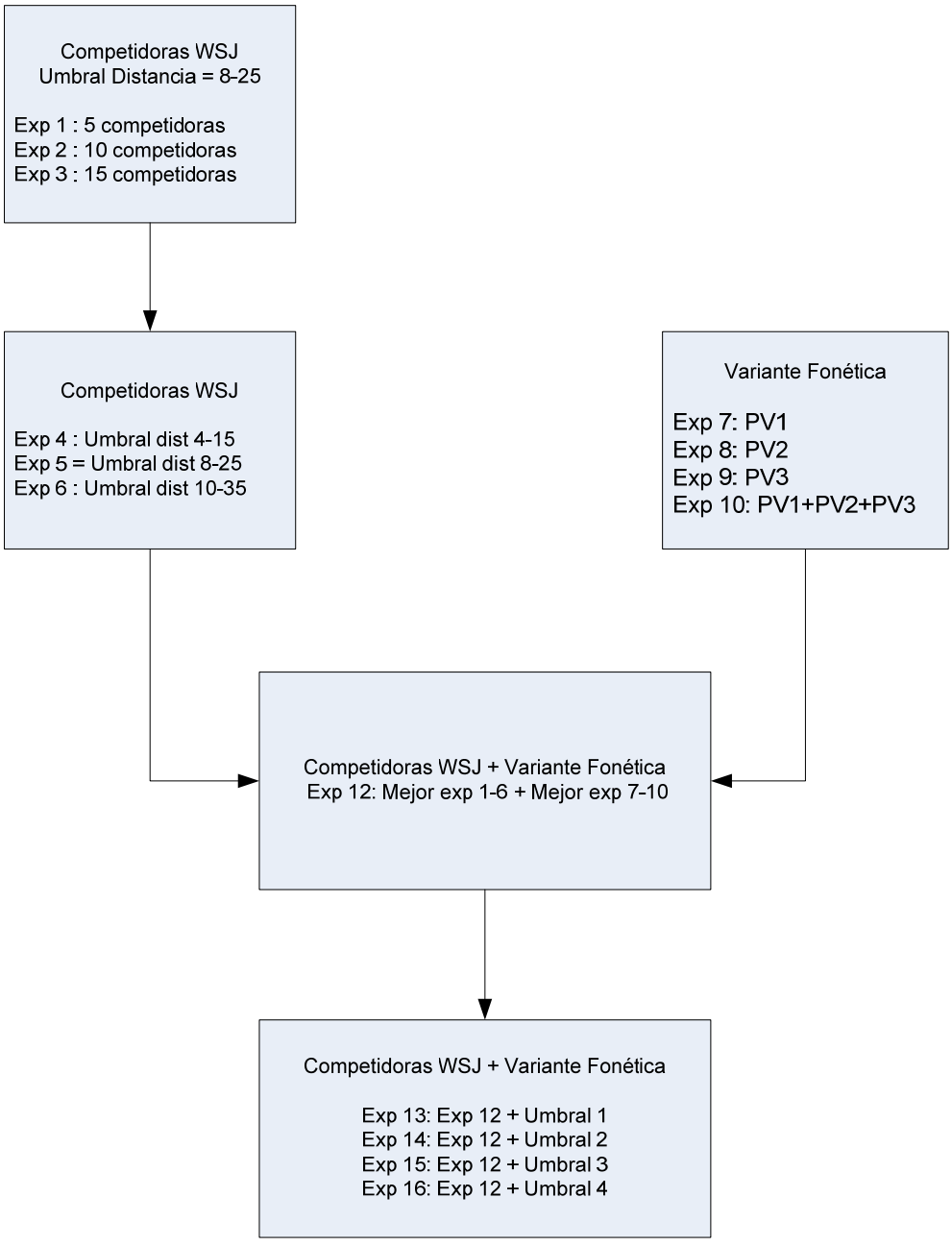

Figura 14: Estructura de Experimentos

4.2. Discusión y resultados.

A continuación se muestran los resultados para los tres criterios subjetivos siguiendo la estructura presentada en la sección 4.1.3, se presenta un gráfico por cada bloque para ver el desempeño en cada etapa.

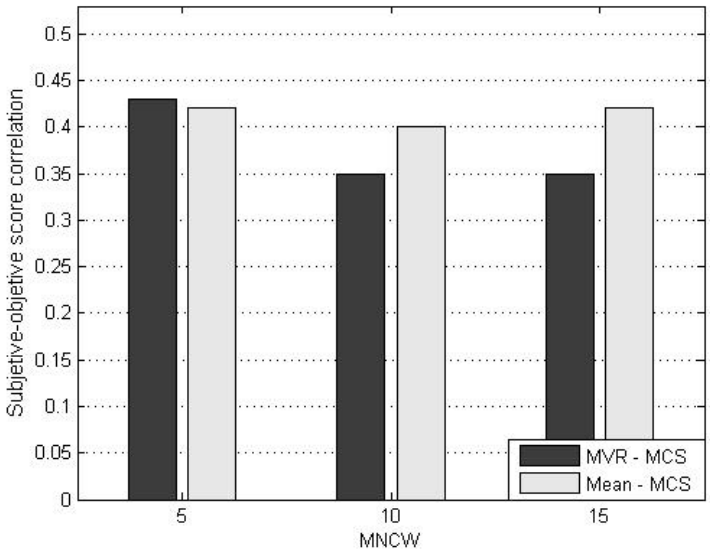

Figura 15: Correlación entre la calificación objetiva y subjetiva v/s MNCW.

La Figura 15 muestra la correlación entre la calificación objetiva y subjetiva donde la clase competidora $Class_{m,l}$ presentada en la sección 3.5 es compuesta por la pronunciación de la palabra *target* $W_{m,l}$ y el vocabulario competitivo $CL_{m,l}$. MNCW definido en la sección 3.6, es igual a 5, 10 y 15 palabras. $SubSentenceScore_{S_m}$ y $ObjSentenceScore_{S_m}$ fueron estimados de acuerdo a SubCrit2 de la sección 3.1 y ObjMetrComb2 de la sección 3.2

respectivamente. Los umbrales $D_{\min}^{CL}$ y $D_{\max}^{CL}$ fueron 8 y 25 respectivamente, tomados de [Molina *et al*,2008]. Como se puede ver en la Figura 15 variar el número máximo de palabras competidoras no muestra diferencias muy significativas en la correlación entre la calificación objetiva y subjetiva cuando la calificación objetiva de las palabras son estimadas combinando las características del ASR con MCS promedio. Por otro lado hay una diferencia mayor en la correlación entre la calificación objetiva y subjetiva con MNCW=5 cuando la calificación objetiva de las palabras es calculada con MCS MVR (*Majority vote rule*). Este resultado sugiere que MCS promedio puede ser más robusto en cuanto a la precisión del ASR que MVR.

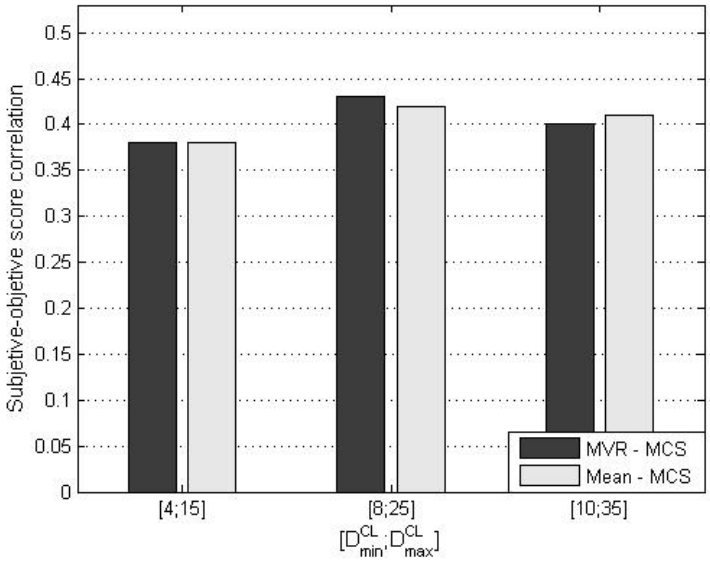

Figura 16: Correlación entre la calificación objetiva y subjetiva v/s $[D_{\min}^{CL}, D_{\max}^{CL}]$

En la Figura 16 tres pares $[D_{\min}^{CL}, D_{\max}^{CL}]$ evaluados cuando la clase competidora $Class_{m,l}$ presentada la sección 3.5 es compuesta por la pronunciación de la palabra *target* $W_{m,l}$ y el vocabulario competitivo $CL_{m,l}$: [4;15]; [8;25]; y [10;35]. MNCW es igual a cinco palabras competidoras, mejor resultado obtenido en la Figura 15. $SubSentenceScore_{S_m}$ y $ObjSentenceScore_{S_m}$ fueron estimados de acuerdo a SubCrit2 y ObjMetrComb2 respectivamente. La mayor correlación entre la calificación objetiva y subjetiva tiene lugar cuando $D_{\min}^{CL} = 8$ y $D_{\max}^{CL} = 25$ con ambos MCS. Observemos que MNCW, $D_{\min}^{CL}$ y $D_{\max}^{CL}$ definen un *trade-off* entre la precisión de la tecnología ASR y la capacidad de discriminar entre una pronunciación correcta o incorrecta.

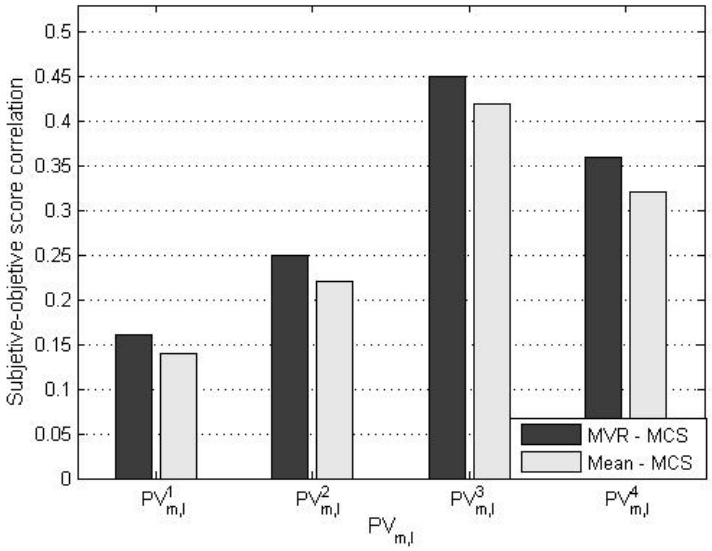

Figura 17: Correlación entre la calificación objetiva y subjetiva v/s $PV_{m,l}$

La Figura 17 muestra la correlación entre la calificación objetiva y subjetiva cuando la clase competidora $Class_{m,l}$ es compuesta por la pronunciación de la palabra *target* $W_{m,l}$ y solo la variante fonética $PV_{m,l}$. $SubSentenceScore_{S_m}$ y $ObjSentenceScore_{S_m}$ fueron estimados de acuerdo a SubCrit2 y ObjMetrComb2 respectivamente. Cuatro combinaciones de $PV_{m,l}$ fueron evaluadas: $PV_{m,l} = \left\{ PV_{m,l}^{1} \right\}$; $PV_{m,l} = \left\{ PV_{m,l}^{2} \right\}$; $PV_{m,l} = \left\{ PV_{m,l}^{3} \right\}$; y $PV_{m,l} = \left\{ PV_{m,l}^{4} \right\}$. Como se puede ver en la Figura 17, la mayor correlación entre la calificación objetiva y subjetiva con ambos MCS se obtiene con $PV_{m,l} = \left\{ PV_{m,l}^{3} \right\}$ Estos resultados sugieren que $PV_{m,l}^{3}$ puede modelar mejor los errores de pronunciación en 2LL (*Second Language Learning* - aprendizaje de segundo idioma) que las otras variantes fonéticas. Como se explico antes, $PV_{m,l}^{3}$ corresponde a la descomposición de la palabra *target* $W_{m,l}$ de acuerdo a las reglas fonéticas del lenguaje materno del estudiante reemplazando los fonemas del lenguaje materno del estudiante con los fonemas del Inglés más similares. La pregunta de por qué este tipo de errores puede ser más frecuente en 2LL que otros esta fuera del alcance de este trabajo.

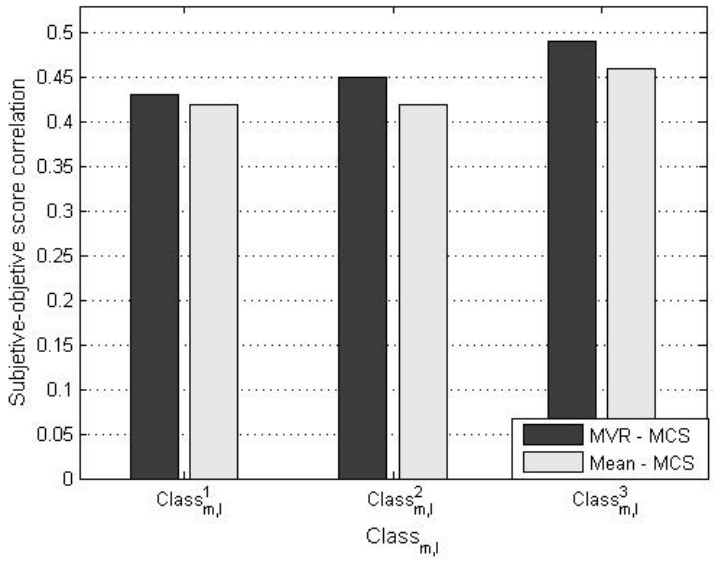

Figura 18: Correlación entre la calificación objetiva y subjetiva v/s $Class_{m,l}$

Los experimentos en la Figura 18 corresponden a las clases competidora $Class_{m,l}$ compuesta por la pronunciación de la palabra *target* $W_{m,l}$, vocabulario competitivo $CL_{m,l}$ y la variante fonética $PV_{m,l}$. $CL_{m,l}$ fue generado con MNCW=5 y $[D^{CL}_{\min}=8, D^{CL}_{\max}=25]$ y $PV_{m,l}=\left\{PV^3_{m,l}\right\}$; $SubSentenceScore_{S_m}$ y $ObjSentenceScore_{S_m}$ fueron estimados de acuerdo a SubCrit2 y ObjMetrComb2 respectivamente. En la Figura 18 se observa $Class^1_{m,l}=\left\{W_{m,l},CL_{m,l}\right\}$, $Class^2_{m,l}=\left\{W_{m,l},PV^3_{m,l}\right\}$ y $Class^3_{m,l}=\left\{W_{m,l},CL_{m,l},PV^3_{m,l}\right\}$. $Class^3_{m,l}$ muestra la mayor correlación entre la calificación objetiva y subjetiva con ambos MCS.

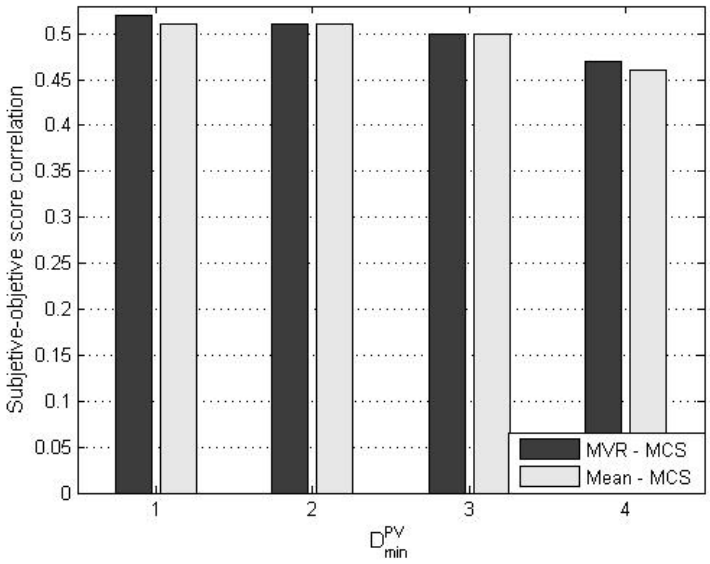

Figura 19: Correlación entre la calificación objetiva y subjetiva v/s $D^{PV}_{\min}$

La Figura 19 muestra la correlación entre la calificación objetiva y subjetiva v/s el umbral $D^{PV}_{\min}$ definido en la sección 3.7. Como fue explicado $D^{PV}_{\min}$ define el umbral mínimo de semejanza entre la pronunciación correcta y la variante fonética debido al hecho de que la precisión de la tecnología ASR impone una habilidad máxima de discriminación. $CL_{m,l}$ fue generado con MNCW=5 y $[D^{CL}_{\min}=8, D^{CL}_{\max}=25]$ y $PV_{m,l}=\left\{PV^{3}_{m,l}\right\}$; $SubSentenceScore_{S_m}$ y $ObjSentenceScore_{S_m}$ fueron estimados de acuerdo a SubCrit2 y ObjMetrComb2 respectivamente. De acuerdo a la Figura 19 la máxima correlación entre la calificación objetiva y subjetiva es alcanzada con $D^{PV}_{\min}=1$ con ambos clasificadores. $D^{PV}_{\min}$ introduce un aumento de 6.1% y 10.8% en la correlación entre la calificación objetiva y subjetiva con los clasificadores MCS MVR y

MCS promedio respectivamente cuando comparamos los resultados con la Figura 18 sin $D_{\min}^{PV}$.

Los experimentos mostrados en las Figura 20-Figura 24 son similares a las Figura 15-Figura 19 excepto por el hecho que $SubSentenceScore_{S_m}$ y $ObjSentenceScore_{S_m}$ fueron estimados de acuerdo a SubCrit1 en la sección 3.1 y ObjMetrComb1 en la sección 3.2 respectivamente.

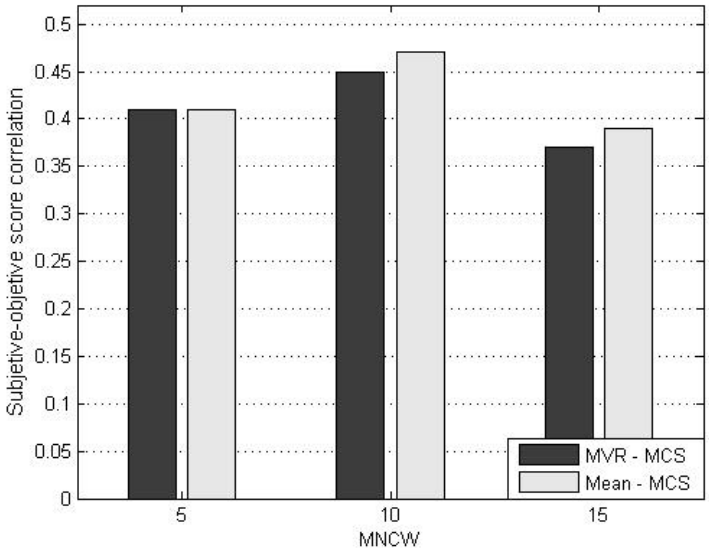

Figura 20: Correlación entre la calificación objetiva y subjetiva v/s MNCW

A diferencia de la Figura 15, la Figura 20 muestra que el optimo de la correlación entre la calificación objetiva y subjetiva se alcanza cuando MNCW es igual a 10 con ambos clasificadores.

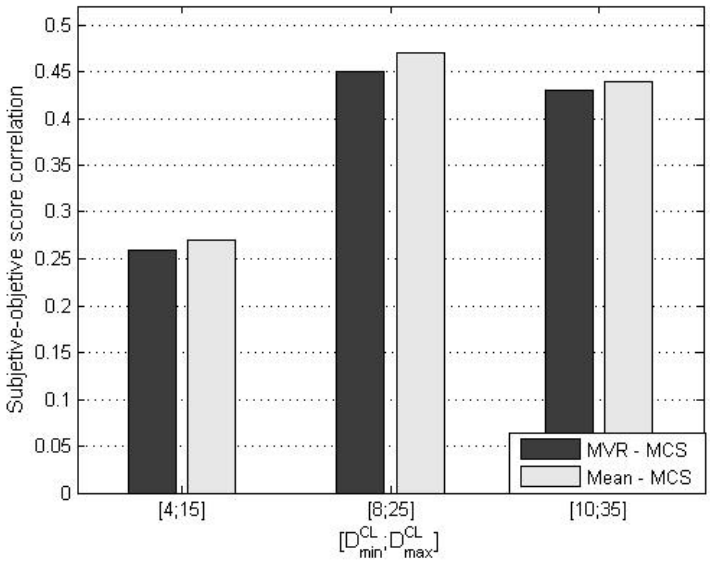

Figura 21: Correlación entre la calificación objetiva y subjetiva v/s $[D_{\min}^{CL}, D_{\max}^{CL}]$

En la Figura 21 la mayor correlación entre la calificación objetiva y subjetiva ocurre en el rango $[D_{\min}^{CL} = 8, D_{\max}^{CL} = 25]$ con ambos clasificadores al estimar $ObjSentenceScore_{S_m}$.

De acuerdo a la Figura 22 la mayor correlación entre la calificación objetiva y subjetiva en ambos clasificadores toma lugar con $PV_{m,l} = \left\{ PV_{m,l}^2 \right\}$. Este resultado no es consistente con el alcanzado en la Figura 17 y podría deberse al hecho que $PV_{m,l}^2$ modela mejor los errores de pronunciación más extremos que $PV_{m,l}^3$.

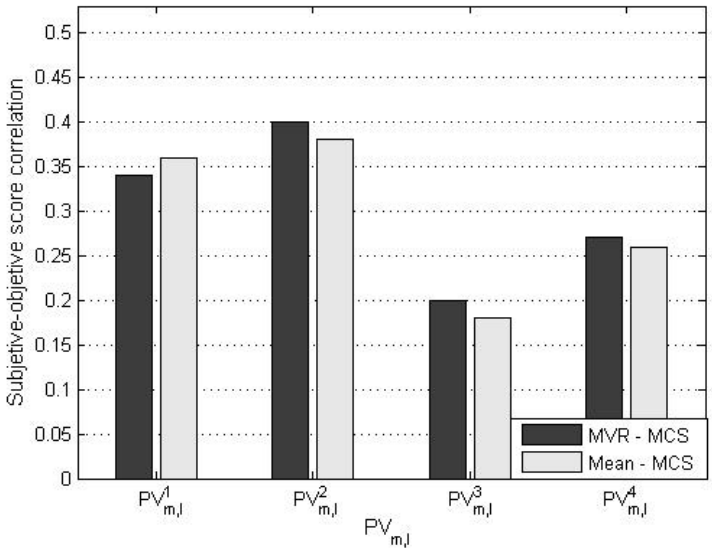

Figura 22: Correlación entre la calificación objetiva y subjetiva v/s $PV_{m,l}$

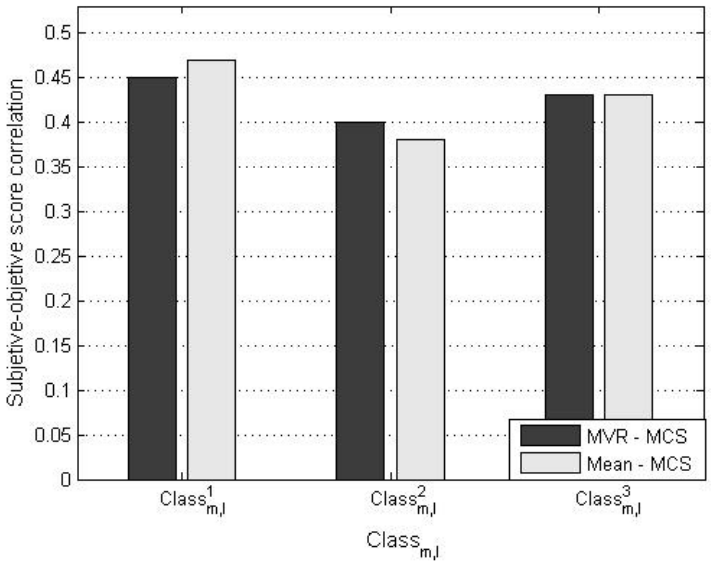

Figura 23: Correlación entre la calificación objetiva y subjetiva v/s $Class_{m,l}$

Como se puede ver en la Figura 23 $Class^{1}_{m,l} = \{W_{m,l}, CL_{m,l}\}$ alcanza una mayor correlación entre la calificación objetiva y subjetiva que $Class^{2}_{m,l} = \{W_{m,l}, PV^{2}_{m,l}\}$ y $Class^{3}_{m,l} = \{W_{m,l}, CL_{m,l}, PV^{2}_{m,l}\}$ con ambos clasificadores. En otras palabras, la introducción de la variante fonética en $Class_{m,l} = \{W_{m,l}, CL_{m,l}\}$ no incorpora una mejora con SubCrit1/ObjMetrComb1.

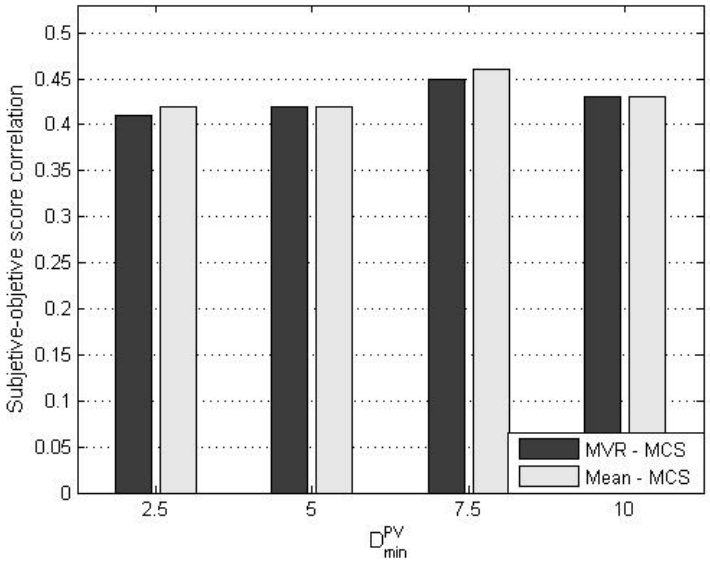

Figura 24: Correlación entre la calificación objetiva y subjetiva v/s $D^{PV}_{\min}$.

A pesar que la incorporación de la alternativa de pronunciación produce perjuicio en cuanto a la correlación, se continua con $PV_{m,l} = \{PV^{2}_{m,l}\}$ para completar el ciclo de experimentos presentado en la sección 4.1.3. En la Figura 24 el umbral mínimo $D^{PV}_{\min}$ definido en la sección 3.7 es evaluada con

$Class_{m,l} = \{W_{m,l}, CL_{m,l}, PV^2_{m,l}\}$. Cuando $D^{PV}_{\min}$ aumenta $Class_{m,l} = \{W_{m,l}, CL_{m,l}, PV^2_{m,l}\}$ tiende a $Class_{m,l} = \{W_{m,l}, CL_{m,l}\}$.

Continuando con la estructura descrita en la sección 4.1.3, los experimentos mostrados en las Figura 25-Figura 29 son similares a las Figura 15-Figura 19 y Figura 20-Figura 24 excepto por el hecho que $SubSentenceScore_{S_m}$ y $ObjSentenceScore_{S_m}$ fueron estimados de acuerdo a SubCrit3 en la sección 3.1 y ObjMetrComb3 en la sección 3.2 respectivamente.

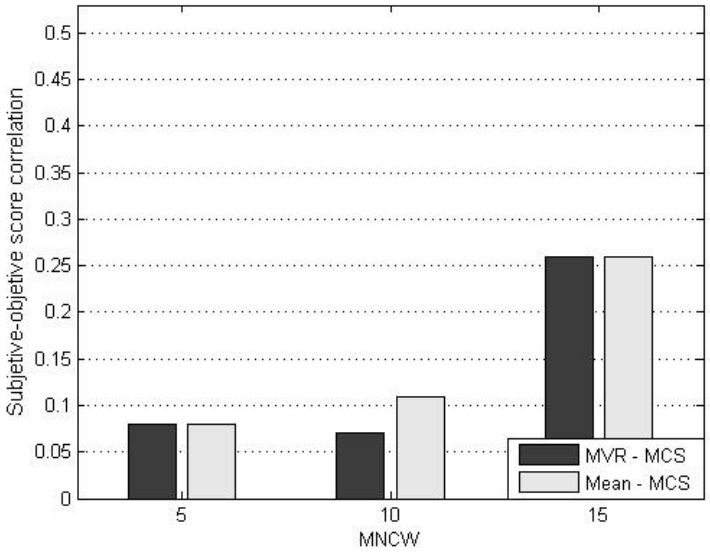

Figura 25: Correlación entre la calificación objetiva y subjetiva v/s MNCW

En la Figura 25 se observa que la correlación entre la calificación objetiva y subjetiva depende de manera más fuerte de MNCW con SubCrit3/ObjMetrComb3 que en los casos anteriores. El máximo se obtiene

cuando MNCW es igual a 15 con ambos MCS. Cabe mencionar que la máxima correlación entre la calificación objetiva y subjetiva en la Figura 25 es mucho menor que para los otros dos criterios.

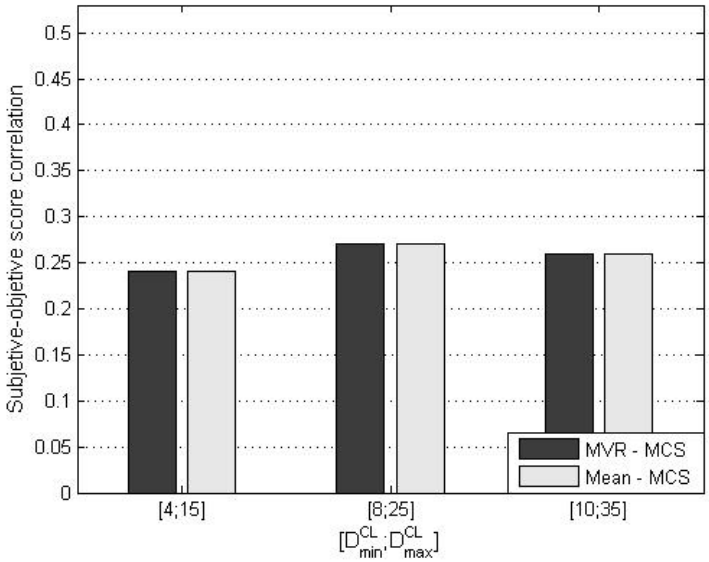

Figura 26: Correlación entre la calificación objetiva y subjetiva v/s $[D_{\min}^{CL}, D_{\max}^{CL}]$

En la Figura 26 la mayor correlación entre la calificación objetiva y subjetiva se obtiene cuando $[D_{\min}^{CL} = 8, D_{\max}^{CL} = 25]$ con ambos clasificadores.

En la Figura 27 se observa que la mayor correlación entre la calificación objetiva y subjetiva es observada con $PV_{m,l} = \left\{ PV_{m,l}^{3} \right\}$ y $PV_{m,l} = \left\{ PV_{m,l}^{4} \right\}$ con MCS MVR y MCS promedio respectivamente.

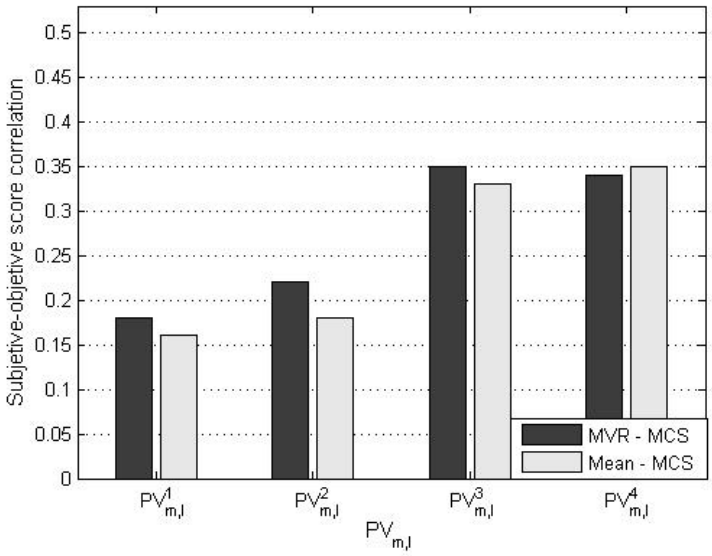

Figura 27: Correlación entre la calificación objetiva y subjetiva v/s $PV_{m,l}$

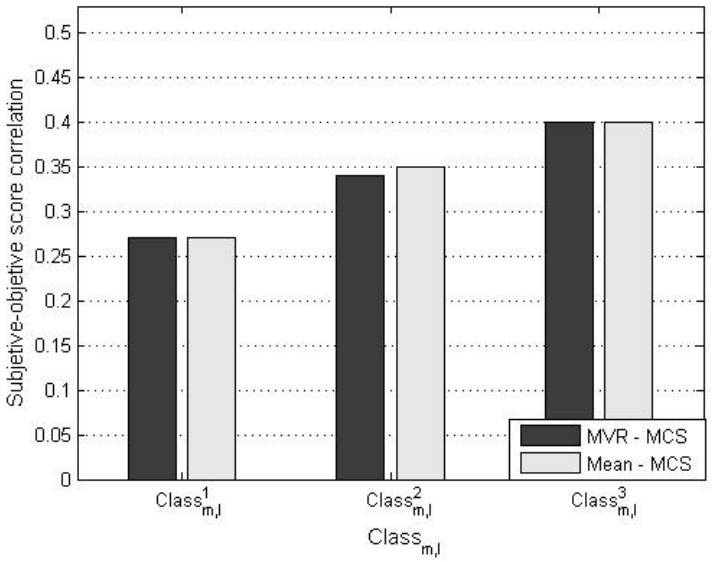

Figura 28: Correlación entre la calificación objetiva y subjetiva v/s $Class_{m,l}$

En la Figura 28, se observa $Class^{1}_{m,l} = \{W_{m,l}, CL_{m,l}\}$, $CL_{m,l}$ fue generado con MNCW=15 y $[D^{CL}_{\min} = 8, D^{CL}_{\max} = 25]$; $Class^{2}_{m,l} = \{W_{m,l}, PV^{4}_{m,l}\}$ y $Class^{3}_{m,l} = \{W_{m,l}, CL_{m,l}, PV^{4}_{m,l}\}$. $Class^{3}_{m,l}$ presenta la mayor correlación entre la calificación objetiva y subjetiva con ambos MCS.

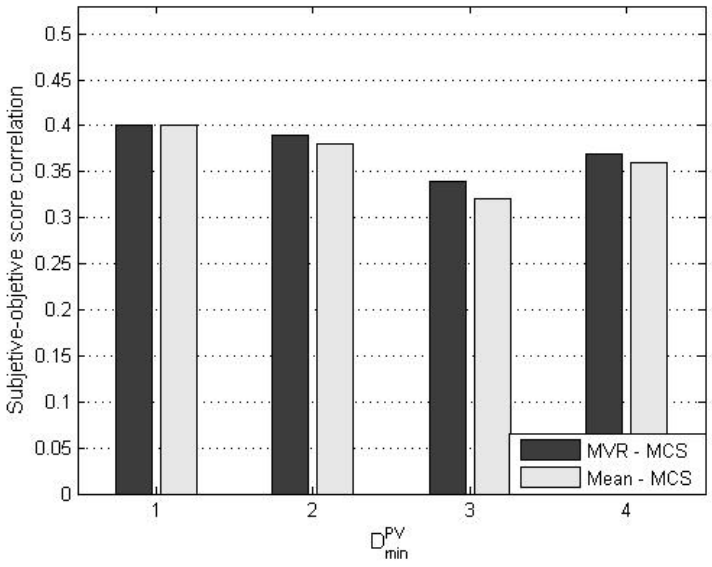

Figura 29: Correlación entre la calificación objetiva y subjetiva v/s $D^{PV}_{\min}$

En la Figura 29 el umbral $D^{PV}_{\min}$ no introduce una mejora en la correlación entre la calificación objetiva y subjetiva comparada con $Class^{3}_{m,l} = \{W_{m,l}, CL_{m,l}, PV^{4}_{m,l}\}$ en la Figura 28.

Los resultados óptimos obtenidos para la correlación entre la calificación objetiva y subjetiva con SubCrit3/ObjMetrComb3 mostradas en las

Figura 25-Figura 29 son bastante menores a las obtenidas con SubCrit1/ObjMetrComb1 y con SubCrit2/ObjMetrComb2. Este resultado puede deberse al hecho que el SubCrit3, que corresponde a la primera impresión, es más difícil de modelar.

A continuación se presentan gráficos matriciales calculados con las mejores configuraciones logradas mostradas en las figuras anteriores, las configuraciones son: $Class_{m,l} = \{W_{m,l}, CL_{m,l}, PV_{m,l}\}$, $CL_{m,l}$ fue generado con MNCW=5, $[D_{\min}^{CL} = 8, D_{\max}^{CL} = 25]$ y $PV_{m,l} = \{PV_{m,l}^{3}\}$ con $D_{\min}^{PV} = 1$ para el promedio; $Class_{m,l} = \{W_{m,l}, CL_{m,l}\}$, $CL_{m,l}$ fue generado con MNCW=10 y $[D_{\min}^{CL} = 8, D_{\max}^{CL} = 25]$ sin alternativa de pronunciación para el mínimo; y $Class_{m,l} = \{W_{m,l}, CL_{m,l}, PV_{m,l}\}$, $CL_{m,l}$ fue generado con MNCW=15, $[D_{\min}^{CL} = 8, D_{\max}^{CL} = 25]$ y $PV_{m,l} = \{PV_{m,l}^{4}\}$ con $D_{\min}^{PV} = 1$ para la moda.

En las Figura 30 y Figura 31 cada criterio subjetivo es modelado con ObjMetrComb1, ObjMetrComb2 y ObjMetrComb3 usando MCS promedio y MVR, respectivamente.

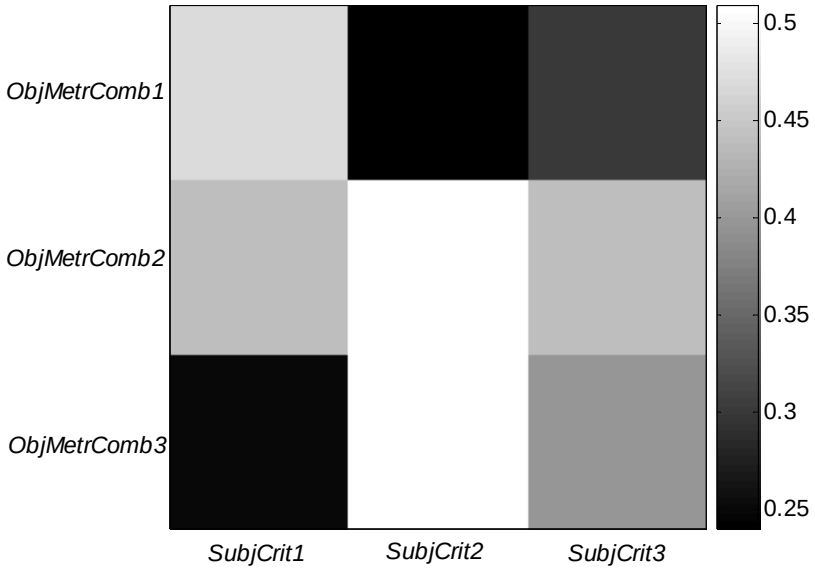

Figura 30: Correlación subjetiva-objetiva con MCS promedio. Cada criterio subjetiva es modelado con ObjMetrComb1, ObjMetrComb2, y ObjMetrComb3 ocupando la mejor configuración obtenida para cada criterio subjetivo.

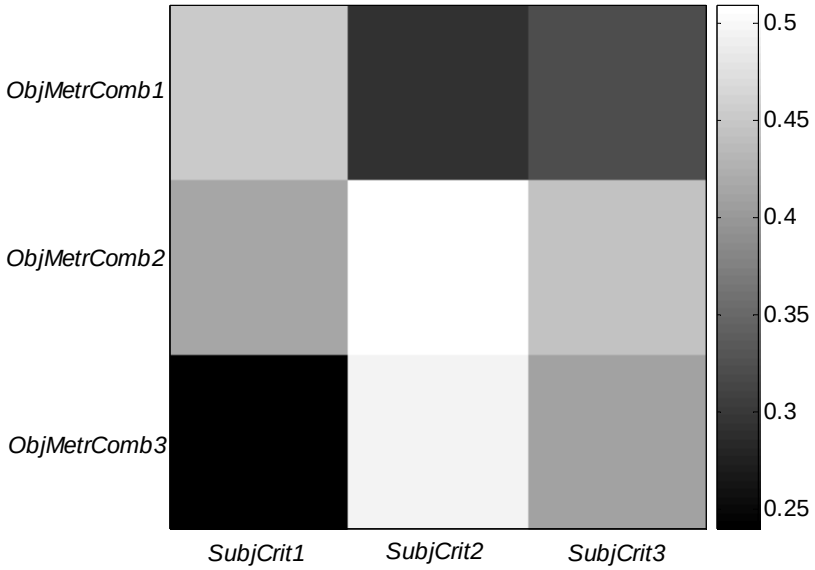

Figura 31: Correlación subjetiva-objetiva con MCS MVR. Cada criterio subjetiva es modelado con ObjMetrComb1, ObjMetrComb2, y ObjMetrComb3 ocupando la mejor configuración obtenida para cada criterio subjetivo

Como se pude ver en las figuras 30 y 31 ObjMetrComb1 entrega una mayor correlación con SubCrit1 que ObjMetrComb2 y ObjMetrComb3. En un análisis similar SubCrit2 es mas preciso con ObjMetrComb2 y ObjMetrComb3. Este resultado sugiere que ObjMetrComb2(promedio) y ObjMetrComb3(moda) entregan una calificación objetiva similar para la frase, esto se debe al hecho que las calificaciones subjetivas de las palabras en las frases muestran una baja desviación estándar y por lo tanto las calificaciones objetivas de las palabras en la frase también presentan esta baja desviación estándar. Similarmente ObjMetrComb2 y ObjMetrComb3 muestran una mayor correlación con SubCrit3 que ObjMetrComb1. Se puede ver que ObjMetrComb3 ocupando MCS MVR o MCS promedio no logra obtener la mayor correlación, se puede desprender de esto que ObjMetrComb3 no modela de buena forma el SubCrit3.

CAPITULO V

5.Conclusiones

5.1.Conclusiones y análisis finales.

En esta tesis es abordado el problema de estimación de calidad de pronunciación para frases en 2LL sin la necesidad de estudios previos de errores comunes. El objetivo principal es proponer y evaluar una nueva estrategia para la evaluación de pronunciación de frases, la cual presenta novedosamente un estudio de diferentes criterios subjetivos obtenidos por expertos calificados en 2LL con el fin de obtener varios niveles de evaluación para ser aplicados en los sistemas CAPT enfocados a diferentes grados de dificultad. Para lograr esto la evaluación de pronunciación de la frase se obtiene evaluando las unidades que la componen, las palabras.

Siguiendo los pasos de [Molina et al.,2009] fue implementado un método no supervisado para generar las frases competidoras en base a la distancia acústica entre palabras. Esto permite que el sistema sea más escalable y no requiere de un detallado estudio para ser implementado, adicionalmente disminuye los costos.

En la línea de CAPT, el concepto de utilizar varios criterios subjetivos para la evaluación de pronunciación presenta posibilidades como una opción a

las alternativas ya existentes. En este trabajo los criterios subjetivos utilizados fueron: primera impresión, mínimo y promedio, los criterios subjetivos fueron modelados por moda, mínimo y el promedio respectivamente. Las correlaciones máximas obtenidas fueron: primera impresión – moda 0.4; mínimo - mínimo 0.47; y promedio - promedio 0.52 con cinco niveles de evaluación.

Los resultados alcanzados ubican esta nueva alternativa para la evaluación de pronunciación de frases dentro de los rangos de correlación encontrados en la literatura y la hacen una alternativa viable para seguir explorando posibilidades y aumentar así la correlación entre las calificaciones subjetivas y objetivas.

Las correlaciones obtenidas para frases no son muy altas, lo que significa que las evaluaciones obtenidas por el reconocedor no son tan similares a las entregadas por un calificador experto, esto ratifica que el sistema sirve como una herramienta complementaria a las clases tradicionales siendo una ayuda para el profesor y nunca un reemplazo para este.

También se observa de los resultados obtenidos que la moda no modela el criterio subjetivo de primera impresión tan bien como se esperaba. En efecto, la correlación fue bastante menor que los logrados con los otros dos criterios. En el otro extremo el promedio alcanza la correlación más alta lograda con las diferentes configuraciones. Por el momento la forma de modelar los criterios subjetivos del mínimo y primera impresión no son lo suficientemente precisos y requieren más trabajo a futuro.

Contrario a lo que se esperaba la incorporación de variantes fonéticas no contribuyo al aumento de la correlación en todos los casos por lo cual no podrá ser considerada como una regla general. En particular la incorporación de las variantes fonéticas de pronunciación incrementó la correlación solo en el caso del promedio, 6.1% y 10,8% con MCS promedio y MCS MVR respectivamente al compararla con la ausencia de variantes fonéticas, en la moda se mantuvo estable la correlación y en el mínimo se obtuvieron resultados adversos al incorporarla.

5.2. Trabajo propuestos a futuro.

Dada las posibilidades de evaluación de pronunciación para CAPT, la utilización de distintos criterios para evaluar distintos niveles de pronunciación se ve como una buena opción. Existen varias líneas donde enfocar nuevos trabajos, desde probar nuevas características extraídas del reconocedor hasta modelar algunos criterios subjetivos de forma distinta para obtener mejores correlaciones. En particular, la moda no modelo el criterio de primera impresión como se esperaba y este es un punto a mejorar.

Al realizar el trabajo se noto que se podría tener en cuenta como una opción considerar la nota de la palabra anterior para obtener la nota de la palabra actual.

6.Glosario

2LL: *Second Language Learning.*

Alineamiento: Proceso para asociar a cada vector de parámetros acústicos un estado de los modelos que describen el ASR (HMMs).

ASR: *Automatic Speech Recognition* – Reconocedor Automático de Voz.

BBCM: *Bayes Based Confidence Measure*.

Calificación objetiva: Nota que entrega el ASR con el fin de imitar la evaluación realizada por un experto.

CALL: *Computed Aided Language Learning*.

CAPT: *Computed Aided Pronunciation Training*.

CM: *Confidence Measure*.

Coeficientes Cepstrales: Parámetros acústicos que caracterizan una señal de voz. Se basan en un análisis en frecuencia de la señal.

Conjunto de Entrenamiento: Señales acústicas que se utilizan para determinar los parámetros de los modelos que describen el ASR.

Conjunto de Test: Señales acústicas que evalúan el reconocedor y que no fueron utilizadas para el entrenamiento de los modelos que describen el ASR.

Criterio Subjetivo: Juicio otorgado por un evaluador experto.

DCT: *Discrete Cosine Transform.*

DFT: *Discrete Fourier Transform*.

Estado: Etapa de un HMM que presenta un período estacionario de una señal acústica. Su valor es escalar.

Feature: También llamada característica en este trabajo, sirve para distinguir una palabra de sus semejantes.

Fonema: Cada una de las unidades fonológicas mínimas que en el sistema de una lengua pueden oponerse a otras en contraste significativo.

***Frame*:** Ventana o segmentación de la señal acústica, unidad mínima de análisis.

HMM: *Hidden Markov Models* – Modelos ocultos de Markov.

LDC: *Linguistic Data Consortium.*

Lenguaje Materno: Lengua Nativa o primera lengua, es el primer idioma o lengua que una persona aprende.

Lenguaje Natural: Situación que se da cuando una aplicación de dialogo conversacional permite que el usuario exprese una solicitud usando mas palabras que las que requiere la aplicación.

LogWD: *Logarithmic word density confidence measure.*

LPTV: Laboratorio de Procesamiento de la Voz.

Matlab: Software matemático que ofrece un entorno de desarrollo integrado (IDE) con un lenguaje de programación propio (lenguaje M).

MFCC: *Mel Frecuency Cepstral.*

MNCW: *Maximun Number of Competitive Words.*

MVR: *Majority Vote Rule.*

Nativo: En el contexto de este trabajo se refiere al idioma propio, innato del locutor, también se suele llamar lengua Materna.

Palabra *target:* Corresponde a la palabra que se quiere reconocer y evaluar.

POS: *N-best position* – Posición en la lista N-best.

REC: *Recognition flag* – Flag de Reconocimiento.

***Score* Objetivo:** Nota asociada a una uteración obtenida a partir de la evaluación realizada por el CAPT.

***Score* Subjetivo:** Nota asociada a una uteración obtenida a partir de la evaluación realizada por un evaluador experto del idioma Inglés.

***Trade-off*:** Cambiar una cosa por otra.

Variante fonética: Es una alternativa de pronunciación, corresponde a una pronunciación errada ya sea por su descomposición en fonemas o por los fonemas utilizados.

WDCM: *Word Density Confidence Measure*.

WSJ: *Wall Street Journal*, Diario de publicación internacional de finanzas y negocios.

7.Referencias

- ***ANGUITA J., HERNANDO J., PEILLON S., BRAMOULLÉ A., 2005.*** *Deteccion of confusable Words in Automatic Speech Recognition, IEEE Signal processing Letters, vol 12 Nº 8.*

- ***BECCHETTI C. AND PRINA L. 1999.*** *Speech Recognition, Theory and C++ Implementation. Wiley E. London, UK.*

- ***BONAVENTURA P., HOWARTH P., MENZEL W., - ,*** *Phonetic annotation of a non-native speech corpus.*

- ***CINCAREK T., GRUHN R., HACKER C., NÖTH E., NAKAMURA S., 2008.*** *Automatic pronunciation scoring of words and sentences independent from the non-native's first language.*

- ***CUCCHIARINI C., STRIK H. AND BOVES L., 1997,*** *Automatic Evaluation of Dutch Pronunciation by using Speech Recognition Technology, IEEE workshop ASRU, pp. 622-629.*

- ***DESHMUKH O., JOSHI S., VERMA A., 2008,*** *Automatic Pronunciation Evaluation and Classification, IBM India Research Lab, New Delhi, India.*

- ***DUDA R. O. AND HART P. E., 1973,*** *Pattern Classification and Scene Analysis. John Wiley & Sons Press, New Cork.*

- ***FORSYTH M. 1995.*** *Discriminating observation probability (dop) hmm for speaker verification. Speech Comm., vol. 17, pp. 117-129.*

- ***FRANCO H., NEUMEYER L., DIGALAKIS V., & RONEN O. 1998****. Combination of Machine Score for Automatic Grading of Pronunciation Quality. Speech Technology and Research Laboratory SRI International.*

- ***FRANCO, H., NEUMEYER, L., KIM, Y., & RONEN, O. 1997****. Automatic Pronunciation Scoring for Language Instruction. In: Ieee international conference on acoustics speech and signal processing. Institute of Electrical Engineers Inc (IEE).*

- ***FUMERA G. AND ROLI F., 2005,*** *A theoretical and experimental analysis of linear combiners for multiple classifier systems, IEEE Trans. Pattern Analysis and Machine Intelligence. 27(6). pp. 942-956.*

- ***GAROFALO J., GRAFF D., PAUL D., AND PALLETT D., 1993.*** *Continous Speech Recognition (CSR-I) Wall Street Journal (WSJ0) news, complete, Linguistic Data Consortium, Philadelphia.*

- ***HAMID S. AND RASHWAN M., 2004*** *Automatic Generation of Hypotheses for Automatic Diagnosis of Pronunciation Errors, in Proceedings of NEMLAR Conference on Arabic Language Resources and Tools, Cairo, Egypt.*

- ***HSU, BO-JUNE, 2007,*** *Generalized linear interpolation of language models, Workshop on Automatic Speech Recognition & Understanding, 2007. ASRU. IEEE, pp. 136 - 140*

- ***JELINEK, F. 1997****. Statistical Methods for Speech Recognition. Massachusetts Institute of Technology (MIT) Press.*

- ***KIM Y., FRANCO H., NEUMEYER L., - ,*** *Automatic Pronunciation scoring of specific phone segments for language instruction. Speech Technology and Research Laboratory SRI International.*

- ***KITTLER J. AND ALKOOT F.M., 2003,*** *Sum versus vote fusion in multiple classifier systems, IEEE Trans. Pattern Analysis and Machine Intelligence. 25, pp. 110-115.*

- ***KITTLER, J., HATEF, M., DUIN, R.P.W., MATAS, J., 1998,*** *On combining classifiers, IEEE Trans. on Pattern Analysis and Machine Intelligence. 20, pp. 226-239.*

- ***KUNCHEVA, L. I., BEZDECK, J. C. AND DUIN, R. P. W. 2001.*** *Decision templates for multiple classifier fusion: an experimental comparison, Pattern Recognition, vol. 34(2), pp. 299-314.*

- ***KUNCHEVA, L.I., 2002,*** *A theoretical study on six classifier fusion strategies, IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 24, pp. 281-286.*

- ***KWAN K. Y., LEE T. AND YANG C., 2002*** *Unsupervised N-best based model adaptation using model-level confidence measures, in Proc. of ICSLP, 2002, pp. 69-72.*

- ***LAMEL, L., RABINER, L., ROSENBERG, A., & WILPON, J. 1981.*** *An improved endpoint detector for isolated word recognition. Acoustics, speech, and signal processing [see also ieee transactions on signal processing], ieee transactions on, 29(4), 777–785.*

- ***LAURILA K., VASILACHE M. AND VIIKKI O. 1998.*** *A Combination of Discriminative and Maximum Likelihood Techniques for Noise Robust Speech Recognition. IEEE Conference on Acoustics, Speech and Signal Processing. 12-15 May, 1998. Seattle, Washington.*

- ***LDC. 1995.*** *Latino database provided by Linguistic Data Consortium. Univ. of Pennsylvania: http://www.ldc.upenn.edu/Catalog/LDC95S28.html.*

- ***MAK M., CHEUNG M., AND KUNG S.**, Robust speaker verification from GSM-transcoded speech based on decision fusion and feature transformation, International Conference on Acoustics, Speech, and Signal Processing (ICASSP). pp. 745 -748, 2003a.*

- ***MOLINA C., 2005,** compensación dependiente de parámetros en reconocimiento de voz con señales distorsionadas por codificación.*

- ***MOLINA C., YOMA N.B., WUTH J., & VIVANCO H. 2008**. ASR based pronunciation evaluation with automatically generated competing vocabulary and classifier fusion. Speech communication.*

- ***MOUSTROUFAS N., DIGALAKIS V., 2006**, Automatic pronunciation evaluation of foreign speakers using unknown text.*

- ***NEUMEYER L., FRANCO H., DIGALAKIS V., WEINTRAUB M., 1999**, Automatic scoring of pronunciation quality Speech Communication.*

- ***PICONE, JW, INC, T.I., & DALLAS, TX. 1993**. Signal modeling techniques in speech recognition. Proceedings of the ieee, 81(9), 1215–1247.*

- ***RABINER, L.R. 1989.** A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the ieee, 77(2), 257–286.*

- ***RANAWANA R., AND PALADE V.**, Multi-Classifier Systems: Review and a roadmap for developers, Int. J. Hybrid Intell. Syst., pp. 35-61., 2006.*

- ***RAVEST P., ABRIL 2009,** Aplicación de tecnologías de robustez en reconocimiento de voz a la enseñanza de segundo idioma.*

- ***ROBLES I.,ABRIL 2009**, Estimación de la curva de entonación para aprendizaje de segundo idioma.*

- ***SCHWARTZ, R., CHOW, Y., KIMBALL, O., ROUCOS, S., KRASNER, M., MAKHOUL, J., BERANEK, B., NEWMAN, I., & CAMBRIDGE, MA. 1985**. Context-dependent modeling for acoustic-phonetic recognition of continuous speech. In: Acoustics, speech, and signal processing, ieee international conference on icassp'85., vol. 10.*

- ***SENEFF, S., HURLEY, E., LAU, R., PAO, C., SCHMID, P., & ZUE, V. 1998.** GALAXY-II: A Reference Architecture for Conversational System Development. In: Fifth international conference on spoken language processing. ISCA.*

- ***SHUANG XU, DENGFENG KE, JIE JIANG, XI YANG, HONGYAN LI, BO XU., 2008**, Automatic Pronunciation Evaluation Based on Feature Extraction and Combination IEEE.*

- ***SI WEI, YI-QIAN PAN, GUO-PING HU, YU HU AND REN-HUA WANG, 2008**, Pronunciation space models for pronunciation evaluation. IEEE.*

- **SOOFUL J., AND BOTHA E., 2002**. *"Comparison of acoustic distance measures for automatic cross-language phoneme mapping," in Proc. of ICSLP, pp. 521-524. Denver, USA.*

- ***WARD W., ISSAR S., 1996**, A Class Based Language Model for Speech Recognition.*

- ***WARSCHAUER M. 1998.** Computer Assisted Language Learning: an Introduction.*

- ***WITT S.M., YOUNG S. J., 2000**. Phone-level pronunciation scoring and assessment for interactive language learning. Speech Communication 30, 95-108.*

- ***WUTH J., 2007**, Desarrollo de aplicación de reconocimiento de voz para telefonía celular.*

- ***XU, L., KRZYZAK, A., SUEN, C.Y., 1992,** Methods of combining multiple classifiers and- their applications to handwriting recognition, IEEE Trans. Syst.. Man, Cybern. SMC-22(3), pp. 418-435.*

❖ ***ZHANG J., WARD W., PELLOM B., YU X., HACIOGLU K., 2001,*** *Improvements in Audio Processing and Language Modeling in the CU Communicator. Center for Spoken Language Research University of Colorado, USA.*

8. Anexo, "On modeling criteria for ASR based pronunciation quality evaluation of sentences"

On modeling criteria for ASR based pronunciation quality evaluation of sentences

Néstor Becerra Yoma[1], Leopoldo Benavides[1], Hiram Vivanco[2]

(1) Speech Processing and Transmission Laboratory

Department of Electrical Engineering, Universidad de Chile, Santiago, Chile

(2) Department of Linguistics, Universidad de Chile, Santiago, Chile

nbecerra@ing.uchile.cl

Telephone: +56-2-978 4205 Fax: +56-2-695 3881

Abstract

In this paper the problem of assessing pronunciation quality of sentences in 2LL without the need of a priori studies of common mistakes is tackled. The proposed method combines objective word scores to approximate subjective evaluations of sentences. The subjective evaluation of sentences is defined with three criteria: the minimum subjective

word score; the mean subjective word score; and, first impression. Then, three combinations of objective word scores are proposed to emulate the subjective criteria: min; mean; and, mode. Class based language models are used to incorporate students`s mother and *target* language phonetic rules to model wrong and correct pronunciations. The results presented here show that the highest subjective-objective score correlation for sentences is achieved when the subjective evaluation criterion based on the mean word-based subjective score is emulated with the mean word-based objective score. In this case the subjective-objective score correlation for sentences is equal to 0.5 with five levels of pronunciation quality. The proposed method does not require any analysis of common mispronunciations. Consequently, the integration of new *target* sentences with the ASR based pronunciation quality evaluation technology is more efficient and requires less human assistance than those approaches based on previous study of pronunciation mistakes.

I. Introduction

The problem of pronunciation quality evaluation in CAPT for 2LL has been addressed by several authors in the last years. Pronunciation quality scores have usually been based on duration, syllabic-timing, and hidden Markov model (HMM) log-likelihoods (Neumeyer et al., 1996; Franco et al., 1997). Initially, those features or confidence metrics attempted to compare the observed signal with native and non-native models by making

use of the forced Viterbi algorithm (Franco et al., 1997). Also in (Franco et al., 1997) phoneme log-posterior score, based on Bayes classification rule for a single feature, leads to a higher correlation between subjective and objective evaluations than the ordinary features or confidence metrics by themselves. In addition, the use of non-native *a priori* acoustic information can be considered as an important source of information to improve the evaluation accuracy by increasing the discriminability between correct and wrong pronunciations (Cucchiarini et al., 1997; Neri et al., 2003).

Combining several features or confidence metrics to evaluate pronunciation quality has also been used in recent papers. In (Nakagawa & Ohta, 2007), a statistical combination of some measures is implemented to make more robust automatic score pronunciation. It is based on the linear combination of scores and the weighting parameters are estimated by using the minimum square error method. Some of the confidence measures that are usually employed in CAPT are: log-likelihood ratio between native English and non-native English HMMs given a spoken sentence; log-likelihood ratio between native English and non-native English HMMs at phoneme level; phoneme recognition rate; and, word recognition rate. In (Tepperman et al., 2007), a Bayes network structure with four metrics was proposed to take into consideration the possible pronunciation errors to jointly evaluate pronunciation quality and reading skills. In (Molina el al., 2009) ASR technology was employed to model non-native speaker pronunciation mistakes as phonetic variants in

automatically generated competitive lexicons without the use of a detailed study of common pronunciation mistakes in isolated-word pronunciation assessment.

Surprisingly, despite the fact that most of the papers on pronunciation quality evaluation are in the framework of isolated words, the problem of assessing the pronunciation quality in sentences has not been addressed as a different task. In (Deshmukh et al., 2008; Cincarek et al., 2008; Xu S et al.,2008; Moustroufas et al., 2006; Neumeyer et al., 2000; Franco et al., 1998,1997; Cuchiarini et al., 1997), several features such as average phone confidence, time normalized phone confidence, fricative confidence, etc, are combined to evaluate the pronunciation quality of sentences. Also, the pronunciation quality of sentences has not been modeled explicitly as the combination of the pronunciation quality of words. For instance, in (Cincarek et al., 2008; Wei el al., 2008) sentences are considered as scoring units. However, it is very noticeable that a 2LL teacher can score an utterance differently depending on the adopted criterion based on the pronunciation of each word. For instance, the whole sentence subjective evaluation may be determined by the worst pronounced word.

The contribution of this paper concerns: a) an analysis of the subjective evaluation of sentences based on the combination of word based evaluations; b) the generation of objective sentence pronunciation scores as the combination of word based scores; c) a method to assess the pronunciation quality of sentences in 2LL without the need of a

priori studies of common mistakes; and, d) the use of class-based language model to incorporate students´s mother and *target* language phonetic rules to represent wrong and correct pronunciations. The results presented here suggest that subjective-objective score correlation for sentences as high as 0.5 with five levels of pronunciation quality can be achieved. As mentioned above, the proposed method does not require any analysis of common mistakes. Also, this correlation is very comparable with those published elsewhere with methods that require a priori studies of pronunciation errors, which in turn also requires a complete definition of the *target* utterances. Consequently, the integration of new *target* sentences with the ASR based pronunciation quality evaluation technology is more efficient and requires less human assistance in the scheme presented here.

II. Automatic pronunciation assessment of sentences with ASR

As discussed below, assessing the pronunciation of sentences in 2LL corresponds to a much more complex problem than the pronunciation evaluation of single words. When listening to a whole sentence pronounced by a student, the human expert on the *target* language can apply one out of several criteria to assess pronunciation quality. For instance, the subjective score associated to a single word, $SubjWordScore_W$, could be defined based on the quality acoustic production of phonemes (Molina el al., 2009). In contrast, the reference subjective score associated to a whole sentence, $SubjSentenceScore_S$, depends on

at least one of the three following criteria that can be employed by the teacher: the minimum $SubjWordScore_W$ within the sentence; the perceived average of $SubjWordScore_W$ within the sentence; and, first impression. On the other way round, three possible combinations of objective word score, $ObjWordScore_W$, in the sentence can be considered to estimate the objective sentence score, $ObjSentenceScore_S$: the minimum $ObjWordScore_W$; the averaged $ObjWordScore_W$; and, the mode of $ObjWordScore_W$.

2.1. Subjective score criteria in sentences

Consider a *target* sentence $S_m = \{W_{m,1}, W_{m,2}, W_{m,3}, ..., W_{m,l}, ..., W_{m,L_m}\}$ composed of L_m words, where $w_{m,l}$ denotes the l^{th} word. As mentioned above, $SubjSentenceScore_S$ could be the result of one of the following criteria applied by the *target* language human expert:

Subjective criterion 1 (SubjCrit1)- The subjective pronunciation score of *target* sentence S_m correspond to the lowest subjective score associated to one of the words $W_{m,l}$ where $1 \le l \le L_m$:

$$SubjSentenceScore_{S_m} = \min_{1 \le l \le L_m} \left\{ SubjWordScore_{W_{m,l}} \right\} \quad (1)$$

Subjective criterion 2 (SubjCrit2) - The subjective pronunciation score of *target* sentence S_m correspond to the perceived average of the subjective scores associated to words $w_{m,l}$ where $1 \le l \le L_m$:

$$SubjSentenceScore_{S_m} = \underset{1 \le l \le L_m}{\text{perceived average}} \left\{ SubjWordScore_{W_{m,l}} \right\} \quad (2)$$

By "perceived average" one denotes the averaged subjective assessment of all the pronounced words that compose the *target* sentence. In this paper the perceived average of $SubjWordScore_{w_{m,l}}$ is modeled as:

$$\underset{1 \le l \le L_m}{\text{perceived average}} \left\{ SubjWordScore_{W_{m,l}} \right\} = \frac{1}{L_m} \cdot \sum_{l=1}^{L_m} SubjWordScore_{W_{m,l}} \quad (3)$$

Subjective criterion 3 (SubjCrit3) - The subjective pronunciation score of *target* sentence S_m is determined by the first impression without an explicit analysis on each pronounced word. Basically, in this case the subjective evaluation is given after having heard a recorded utterance only once.

2.2. Objective sentence scores as a combination of word based objective scores

If each word $W_{m,l}$ in sentence S_m, where $0 \le l \le L_m$, is associated to an objective score $ObjWordScore_{W_{m,l}}$, then objective sentence score $ObjSentenceScore_{S_m}$ can be estimated by employing one of the following metric combination:

Objective metric combination 1 (ObjMetrComb1)- The objective pronunciation score of *target* sentence S_m correspond to the lowest objective score associated to one of the words $W_{m,l}$ where $1 \le l \le L_m$:

$$ObjSentenceScore_{S_m} = \min_{1 \le l \le L_m} \left\{ ObjWordScore_{W_{m,l}} \right\} \quad (4)$$

Objective metric combination 2 (ObjMetrComb2)- The objective pronunciation score of *target* sentence S_m correspond to the average word-based objective score:

$$ObjSentenceScore_{S_m} = \frac{1}{L_m} \cdot \sum_{l=1}^{L_m} ObjWordScore_{W_{m,l}} \quad (5)$$

Objective metric combination 3 (ObjMetrComb3)- The objective pronunciation score of *target* sentence S_m is equal to the statistic mode or most frequent word within S_m:

$$ObjSentenceScore_{S_m} = \max_{1 \le l \le L_m} \left\{ Frequency\left[ObjWordScore_{W_{m,l}} \right] \right\} \quad (6)$$

where $Frequency\left[ObjWordScore_{W_{m,l}} \right]$ indicates how many times word score $ObjWordScore_{W_{m,l}}$ appears in S_m.

2.3. Subjective and objective score correlation in sentences

At this point the correspondence between subjective score criteria in subsection 2.1 and objective score combinations in subsection 2.2 is straightforward in the following cases: SubjCrit1 and ObjMetrComb1; and, SubjCrit2 and ObjMetrComb2. However, SubjCrit3, which is defined as first impression, is much more difficult to define. A possibility is to model SubjCrit3 with the mode of the objective metrics within the *target* utterance (i.e.

ObjMetrComb3). In this paper the accuracy of objective metric combinations is estimated by means of the correlation between the subjective scores provided by human experts and the ASR based objective metrics.

III. ASR based objective metric with competitive vocabulary and class based language model

In this paper the objective quality pronunciation score associated to a sentence S_m, $ObjSentenceScore_{S_m}$, is estimated as a combination of word based objective scores, $ObjWordScore_{W_{m,l}}$, as proposed in subsection 2.3. Scores $ObjWordScore_{W_{m,l}}$ are estimated by a continuous speech recognition system with class based language model (Ward and Issar, 1996; Zhang et al., 2001). Given a sentence $S_m = \{W_{m,1}, W_{m,2}, ..., W_{m,l}, ..., W_{m,L_m}\}$, competitive class $Class_m = \{Class_{m,1}, Class_{m,2}, ..., Class_{m,l}, ..., Class_{m,L_m}\}$ where each class $Class_{m,l}$ can be composed of: *target* word $W_{m,l}$; a competitive lexicon; and, phonetic variants of $W_{m,l}$. Both the competitive lexicon and phonetic variants require no a priori analysis of common mistakes and are automatically generated. The continuous speech recognition system is employed to make compete the *target* pronunciation of sentence S_m with the pronunciation of sentences composed of similar words and phonetic variants of *target* words. To do so, a class-based trigram language model is generated for each

sentence S_m by estimating $\Pr(Class_{m,l} \mid Class_{m,l-1}, Class_{m,l-2})$. Figure 1 shows the block diagram of the proposed scheme to assess pronunciation quality of sentences based on continuous speech recognition. Per each *target* word N-best list analysis resulting from Viterbi decoding delivers a set of word features, $\overline{WF}_{W_{m,l}} = \left[WF^1_{W_{m,l}}, WF^2_{W_{m,l}}, ..., WF^j_{W_{m,l}}, ..., WF^J_{W_{m,l}} \right]$, where J is the number of features. As explained later, examples of word features are the position in the N-best list where a *target* word $W_{m,l}$ is contained and the word density confidence measure (Kwan el al., 2002). Each word feature $WF^j_{W_{m,l}}$ delivered by the Viterbi decoding is mapped to an objective $ObjWordScore^j_{W_{m,l}}$ by making use of Bayes decision rule. Then the objective pronunciation metric associated to word $W_{m,l}$, $ObjWordScore_{W_{m,l}}$, is obtained by combining $ObjWordScore^j_{W_{m,l}}$, where $1 \le j \le J$, by employing multi-classifier fusion techniques. Finally, the objective pronunciation score that corresponds to sentence S_m is obtained by combining $ObjWordScore_{W_{m,l}}$, where $1 \le l \le L_m$, according to subsection 2.2.

3.1. Automatic generation of competitive class

Each word within the *target* sentence generates a class composed of: a) *target* word $W_{m,l}$ with the correct pronunciation; b) a competitive lexicon similar to the *target* word with the correct pronunciation; and, c) phonetic variants of the *target* word according to the *target* or student´s mother language. As a result, class $Class_{m,l}$ can be represented as:

$$Class_{m,l} = \left\{ W_{m,l}, CL_{m,l}, PV_{m,l} \right\} \quad (7)$$

where $CL_{m,l} = \left\{ CL^1_{m,l}, CL^2_{m,l}, CL^3_{m,l}, ..., CL^k_{m,l}, ..., CL^{K_{m,l}}_{m,l} \right\}$ denotes the competitive lexicon composed of words $CL^k_{m,l}$, where $1 \le k \le K_{m,l}$ and $K_{m,l}$ is the number of words in $CL_{m,l}$; and, $PV_{m,l}$ denotes the phonetic variants of the *target* word according to the *target* and student´s mother language within $Class_{m,l}$. No previous analysis based on errors made by students is required in order to achieve an efficient integration of didactic material to ASR technology without human assistance. Observe that the definition and generation of $Class_{m,l}$ attempt to find a trade-off between the accuracy of the pronunciation assessment and the limitation of the ASR technology: the higher the number of competing words and phonetic variants, the more difficult the recognition task itself. The automatic generation of $CL_{m,l}$ and $PV_{m,l}$ is described as follows.

3.1.1 Automatic generation of competitive lexicon

Competitive vocabulary $CL_{m,l}$ helps to force the simultaneous competition of the correct and wrong pronunciation and is crucial to make ASR technology successful in CAPT. This paper employs the same approach proposed in (Molina el al., 2009). First of all, the K-L (Kullback-Leibler) distance defined in (Sooful and Botha, 2002) between *target* word $W_{m,l}$ and words from a lexicon are estimated. The lexicon should be complete

enough and representative of the *target* language in order to include a significant range of word distances. Second, the lexicon whose distance to the *target* word is within an interval defined by a minimum, D_{min}^{CL}, and a maximum, D_{max}^{CL}, thresholds is stored. Then, the lexicon within the interval $\left[D_{min}^{CL}; D_{max}^{CL} \right]$ is sorted with respect to the distance to the *target* word and uniformly sampled to reduce the number of selected words to *MNCW* (Maximum Number of Competitive Words). Parameters $\left[D_{min}^{CL}; D_{max}^{CL} \right]$ define a trade-off between the discrimination ability resulting from the distance between the competitive lexicon and the *target* word, and the accuracy of the speech recognition technology.

3.1.2. Automatic generation of Spanish phonetic variant of target words

To improve the accuracy of the pronunciation quality evaluation, variants of the phonetic realization of target word $W_{m,l}$ according to the target and student's mother language (i.e. Spanish in this case), $PV_{m,l} \subset \left\{ PV_{m,l}^{1}, PV_{m,l}^{2}, PV_{m,l}^{3}, PV_{m,l}^{4} \right\}$, are included in competitive class $Class_{m,l}$. This strategy attempts to incorporate information on user's mother language without implementing a detailed study of pronunciation mistakes made by students. The phonetic variants are generated as follows. According to Fig. 2, target $W_{m,l}$ can be decomposed according to English or student´s mother language (i.e. Spanish) phonetic rules. In the case of English phonetic decomposition, there are two possibilities: using English phonemes; and, replacing English phonemes with the most similar phoneme

in student´s mother language according to Table 1. In the case of decomposition according to student´s mother language phonetic rules, there also two possibilities: employing Spanish phonemes; and, replacing Spanish phonemes with the most similar phonetic units in English according to Table 2. It is worth highlighting that Table 1 and 2 were generated by an expert in English language and phonetics. As a result, the phonetic variant component $PV_{m,l}$ in $Class_{m,l}$ according to (7) is defined as follows:

$PV^1_{m,l}$ - Decomposition of target word $W_{m,l}$ according to English language phonetic rules by replacing English phonemes with the most similar student´s language ones according to Table 1.

$PV^2_{m,l}$ - Decomposition of target word $W_{m,l}$ according to student´s language phonetic rules and phonemes.

$PV^3_{m,l}$ - Decomposition of target word $W_{m,l}$ according to student´s language phonetic rules by replacing phonemes with the most similar English one according to Table 2.

$PV^4_{m,l}$ - It includes $PV^1_{m,l}$, $PV^2_{m,l}$ and $PV^3_{m,l}$ simultaneously in $Class_{m,l}$.

Then, $PV_{m,l}^{i}$, where $1 \le i \le 4$, is included in $PV_{m,l}$ if the K-L distance between $PV_{m,l}^{i}$ and target word $W_{m,l}$ is greater or equal than D_{min}^{PV}. Threshold D_{min}^{PV} defines a trade-off between the discrimination ability resulting from the distance between the phonetic variants and the target word, and the accuracy of the speech recognition technology.

3.2. Class based language model in ASR

Continuous speech recognition is run by using class based language model (Zhang et al., 2001;Ward and Issar, 1996). As mentioned above, competitive class $Class_m$ is generated from sentence S_m. Then, trigrams $Pr\left(Class_{m,p} \mid Class_{m,q}, Class_{m,\gamma}\right)$ are defined as follows:

$$Pr\left(Class_{m,p} \mid Class_{m,q}, Class_{m,\gamma}\right) = \begin{cases} 1 & \text{if } q = p-1 \text{ and } r = q-1 \text{ and } p \ge 3 \\ 0 & \text{otherwise} \end{cases} \quad (8)$$

When $p = 2$, $Pr\left(Class_{m,p} \mid Class_{m,q}, Class_{m,\gamma}\right)$ is replaced with bigram $Pr\left(Class_{m,p} \mid Class_{m,q}\right)$:

$$Pr\left(Class_{m,p} \mid Class_{m,q}\right) = \begin{cases} 1 & \text{if } q = p-1 \text{ and } p = 2 \\ 0 & \text{otherwise} \end{cases} \quad (9)$$

The use of class based language model attempts to identify one or more mispronounced words within the target sentence.

3.3. Word based N-best list feature extraction

Given a target utterance, ASR with class based language model makes possible the efficient extraction of several features per word. In this paper four confidence measures delivered by the ASR procedure are employed: N-best position; recognition flag; word density confidence measure and logarithmic word density confidence measure.

3.3.1. Position in the N-best list

Position in the N-best list of word $W_{m,l}$ in target sentence S_m, $POS_{m,l}$, corresponds to the index of the most likely hypothesis where $W_{m,l}$ is recognized:

$$POS_{m,l} = argmax_r \left\{ \left[Q(h_r) \right] \middle| r \in E(W_{m,l}, H) \right\} \tag{10}$$

where $Q(h_r) = P(h_r)^{\gamma} \cdot P(O/h_r)$; h_r is the r^{th} hypothesis in the N-Best Viterbi list; $Q(h_r)$ is the likelihood score given by the Viterbi search; $P(h_r)$ is the language model probability of h_r; $P(O/h_r)$ is the observation probability of h_r; γ is the acoustic model scaling factor; $E(W_{m,l}, H)$ corresponds to the indices of the hypotheses where word $W_{m,l}$ is contained; and finally, H denotes all the N-best alignments or hypotheses obtained from Viterbi decoding.

3.3.2. Recognition Flag

This binary confidence measure associated to word $W_{m,l}$ in target sentence S_m, denoted by $REC_{m,l}$, is defined as:

$$REC_{m,l} = \begin{cases} 1 & if\ W_{m,l} \subset h_1 \\ 0 & if\ W_{m,l} \not\subset h_1 \end{cases} \quad (11)$$

where h_1 is the first hypothesis in the N-Best Viterbi list.

3.3.3. Word density confidence measure

Word density confidence measure of word $W_{m,l}$ in target sentence S_m, $WDCM_{m,l}$, is defined as:

$$WDCM_{m,l} = \frac{\sum_{r \in E(W_{m,l},H)} Q(h_r)}{\sum_{l=1}^{N} Q(h_l)} \quad (12)$$

3.3.4. Logarithmic Word density confidence measure in the logarithmic domain

Logarithmic Word density confidence measure of target word $W_{m,l}$, $LogWDCM_{m,l}$, is defined as:

$$LogWDCM_{m,l} = \frac{\sum_{r \in E(W_{m,l},H)} \log(Q(h_r))}{\sum_{l=1}^{N} \log(Q(h_l))} \quad (13)$$

3.4. Word based objective pronunciation score estimation

As in (Molina el al., 2009), word based objective pronunciation score, $ObjWordScore_{W_{m,l}}$, is estimated by employing MCS (multi-classifier system) techniques as shown in Fig. 1. As described above, four word features or confidence metrics are evaluated per each word in target sentences: $POS_{m,l}$, $REC_{m,l}$, $WDCM_{m,l}$ and $LogWDCM_{m,l}$. The problem of word based pronunciation quality evaluation is modeled as a mapping between confidence metrics and score $ObjWordScore_{W_{m,l}}$ that emulate the opinion given by a human instructor, $SubjWordScore_{W_{m,l}}$. Suppose that subjective score $SubjWordScore_{W_{m,l}}$ is quantized in M levels (in this paper $M = 5$). Consequently, every confidence metrics could be assumed as a score delivered by a given classifier and every subjective score level would be a class. Consider that O is the sequence of observation vectors corresponding to target sentence S_m uttered by a student. By using the Bayes rule, $ObjWordScore_{W_{m,l}}$ can be estimated as:

$$\begin{aligned} ObjWordScore_{W_{m,l}}(O) &= \underset{C_m}{argmax}\, P\left[C_m / \overrightarrow{WF}_{W_{m,l}}(O)\right] \\ &= \underset{C_m}{argmax}\left\{\frac{P(\overrightarrow{WF}_{W_{m,l}}(O)/C_m)P(C_m)}{P(\overrightarrow{WF}_{W_{m,l}}(O))}\right\} \end{aligned} \tag{14}$$

where $ObjWordScore_{W_{m,l}}(O)$ is the final decision for $W_{m,l}$ that corresponds to signal O. Theoretically, the classification error is optimally minimized by (14). $P(C_m)$ is assumed

uniformly distributed and equal to $\frac{1}{M}$. However, the *a priori* multivariable p.d.f.'s $P(\overrightarrow{WF_{W_{t,n}}}(O)/C_m)$ and $P(\overrightarrow{WF}_{W_{m,l}}(O))$ may require an unmanageable amount of training data to be estimated reliably (Kittler et al., 1998). As a consequence, the problem is substantially simplified if maximization in (14) could be expressed in terms of computations performed by individual classifiers. The classical techniques to simplify the Bayesian Fusion (Kittler et al., 1998; Kuncheva et al., 2001, 2002) are: Product Rule; Max Rule; Min rule; Mean Rule; and, Majority Vote Rule. Among the several MCS combinations rules in the literature, Mean Rule and Majority Vote Rule are the most frequently employed approximations to simplify the Bayesian Fusion (Kittler & Alkoot, 2003; Fumera & Roli, 2005). Product Rule corresponds to the optimal Bayesian fusion if the classifiers are statistically independent. Vote Rule allows combining local decision of individual classifiers. The mean rule is defined as:

$$\begin{aligned} OPrS_{W_{t,n}}(O) &= argmax_{C_m}\left\{\frac{1}{J}\sum_{j=1}^{J}P(C_m/WF^{j}_{W_{t,n}}(O))\right\} \\ &= argmax_{C_m}\left\{\frac{1}{J}\sum_{j=1}^{J}\frac{P(WF^{j}_{W_{t,n}}(O)/C_m)P(C_m)}{P(WF^{j}_{W_{t,n}}(O))}\right\} \end{aligned} \tag{15}$$

As mentioned above, $1 \le m \le M$ and M is the total number of possible levels of pronunciation quality and J is the total number of word features or confidence metrics.

Majority Vote Rule (MVR) is a straightforward scheme to combine the output of individual classifiers (Kittler & Alkoot, 2003). Given a set of individual classifier decisions $OPrS_{W_{t,n}}(O) = [OPrSWF^1_{W_{t,n}}(O), OPrSWF^2_{W_{t,n}}(O), \ldots, OPrSWF^j_{W_{t,n}}(O)]$, where $OPrSWF^j_{W_{t,n}}(O)$ is the decision provided by the classifier corresponding to word feature *j*, then the final decision, $OPrS_{W_{t,n}}(O)$, will be the class which receives the largest number of votes as the consensus (majority).

IV. Experiments

The native American English acoustic models were trained with CSR-I WSJ0[1] corpus (Garafalo et al., 1993). In CSR-I WSJ0 speech data was recorded with a high-quality microphone and the sample rate was equal to 16 kHz. All the training signals (20055 utterances) were used to train the CDHMMs. Also, LATINO 40 (LDC, 1995) was employed to train the Spanish phonetic units used to generate the phonetic variants according to subsection 3.1.2. This database is composed of continuous speech from 40 Latin American native speakers, with each speaker reading 125 sentences from newspapers in Spanish. The training utterances were 4500 uncoded sentences provided by 36 speakers and context-dependent phoneme HMMs were employed. The vocabulary is composed of almost 6000 words. Thirty-three MFCC parameters per frame were

[1] CSR-I (WSJ0) Sennheiser, Publisher by LDC, ISBN: 1-58563-006-3

computed: the frame energy plus ten static coefficients and their first and second time derivatives. Cepstral Mean Normalization (CMN) was also employed. Each monophoneme and triphoneme was modeled with a three-state left-to-right topology without skip-state transition, with eight multivariate Gaussian densities per state with diagonal covariance matrices. Language model is estimated according to subsection 3.2. As explained above, a competitive class, as defined in (7), is generated for each target word within a given target sentence. The competitive lexicon for each target word was chosen from the vocabulary that composes the CSR-I WSJ0 corpus as in (Molina el al., 2009). The phonetic variants of each target word are generated according to subsection 3.1.2 and Fig. 2. In this paper four confidence measures delivered by the ASR procedure are employed: $POS_{m,l}$; $REC_{m,l}$; $WDCM_{m,l}$; and, $LogWDCM_{m,l}$.

The data base is composed of the following sentences: “It was nice to see my relatives”; “I was so happy to see my parents”; “I missed my kiwi family”; “Lots of people play soccer”; “People live a healthy life here”; “New Zealand is a country in Oceania”; “She gently shows me my seat”; “I have never had this experience”; “To meet different people”; and, “The airport is so big”. These sentences were extracted from a lesson designed for a web based 2LL system designed at LPTV (see Fig. 3) and were revised by an expert in English language and phonetics in order to achieve a phonetically balanced evaluation data set. Each sentence was pronounced by 43 speakers, which showed different levels of English proficiency, and was recorded with inexpensive desktop

microphones. Seven utterances were discarded due to the fact that they showed a high percentage of clipped samples. Finally, the database, which is composed of 423 utterances, was divided in training (212 utterances) and testing (211 utterances) data. The training data was employed to estimate the *a priori* p.d.f.´s in (15).

The subjective scores were determined with seven experts in English language. The pronunciation quality of each word within target utterance l, $SubjWordScore_{W_{m,l}}$, was evaluated by two experts in English language as in (Molina el al., 2009). If both evaluations diverged, the opinion of a third expert was taken into consideration. Five different categories of pronunciation errors were defined per each target word from 1 to 5: score 5 corresponds to the correct pronunciation of the target word; and, score 1 denotes the worst possible pronunciation, i.e. the result of the application of Spanish pronunciation rules. Then, subjective scores SubjCrit1 and SubjCrit2 in each sentence were estimated as explained in subsection 2.1 with $SubjWordScore_{W_{m,l}}$. In contrast, subjective score SubjCrit3, which corresponds to the first impression, was determined by asking the experts in English language to give their opinion after listening each sentence only once.

V. Discussion

Figure 4 shows the subjective-objective score correlation when competitive class $Class_{m,l}$ in (7) is composed of the target pronunciation of $W_{m,l}$ and competitive lexicon

$CL_{m,l}$. *MNCW*, defined in subsection 3.1.1, is made equal to 5, 10 and 15 words. $SubjSentenceScore_{S_m}$ and $ObjSentenceScore_{S_m}$ were estimated according to SubjCrit2 in (3) and ObjMetrComb2 in (5), respectively. Thresholds D_{min}^{CL} and D_{max}^{CL} were made equal to 8 and 25, respectively. As can be seen in Fig. 4, varying the maximum number of competitive words does not lead to significant differences in the subjective-objective score correlation when the objective word scores are estimated by combining ASR features with the MCS mean rule (Fig. 1). In contrast, a more distinguishable optimum subjective-objective score correlation is obtained at *MNCW*=5 when the objective word scores are computed with the MCS majority vote rule (MVR). This result suggests that the MCS mean rule could be more robust to the ASR precision than MVR.

In Fig. 5 three pairs $\left[D_{min}^{CL};\ D_{max}^{CL}\right]$ were evaluated when competitive class $Class_{m,l}$ in (7) is composed of the target pronunciation of $W_{m,l}$ and competitive lexicon $CL_{m,l}$: $[4;\ 15]$; $[8;\ 25]$; and $[10;\ 35]$. *MNCW* was made equal to five competitive words. $SubjSentenceScore_{S_m}$ and $ObjSentenceScore_{S_m}$ were estimated according to SubjCrit2 in (3) and ObjMetrComb2 in (5), respectively. The highest subjective-objective score correlation takes place when $D_{min}^{CL} = 8$ and $D_{max}^{CL} = 25$ with both MCS mean and majority vote rule to estimate $ObjSentenceScore_{S_m}$. Observe that *MNCW*, D_{min}^{CL} and D_{max}^{CL} define a trade-off between the ASR technology accuracy and the target capability to discriminate between correct and wrong pronunciations. This trade-off is clearly observed in Fig. 5.

Figure 6 shows the subjective-objective score correlation when competitive class $Class_{m,l}$ in (7) is composed of the target pronunciation of $W_{m,l}$ and phonetic variants $PV_{m,l}$ only. $SubjSentenceScore_{S_m}$ and $ObjSentenceScore_{S_m}$ were estimated according to SubjCrit2 in (3) and ObjMetrComb2 in (5), respectively. Four configurations of $PV_{m,l}$ were evaluated: $PV_{m,l} = \left\{ PV_{m,l}^1 \right\}$; $PV_{m,l} = \left\{ PV_{m,l}^2 \right\}$; $PV_{m,l} = \left\{ PV_{m,l}^3 \right\}$; and, $PV_{m,l} = \left\{ PV_{m,l}^4 \right\}$. As can be seen in Fig. 6, the highest subjective-objective score correlation with both MCS mean and majority vote rule takes place with $PV_{m,l} = \left\{ PV_{m,l}^3 \right\}$. This result suggest that $PV_{m,l}^3$ could better model pronunciation mistakes in 2LL than $PV_{m,l}^1$ and $PV_{m,l}^2$. As explained above, $PV_{m,l}^3$ corresponds to the decomposition of target word $W_{m,l}$ according to student´s mother language phonetic rules where each phoneme is replacing with the most similar English phoneme. The question of why this type of mistake could be more frequent in 2LL than the others is out of the scope of the current paper.

Experiments in Fig. 7 correspond to when the competitive class $Class_{m,l}$ is composed of the target pronunciation of $W_{m,l}$, competitive lexicon $CL_{m,l}$ and phonetic variance $PV_{m,l}$. $CL_{m,l}$ was generated with *MNCW*=5 and $\left[D_{min}^{CL} = 8;\ D_{max}^{CL} = 25 \right]$, and $PV_{m,l} = \left\{ PV_{m,l}^3 \right\}$. $SubjSentenceScore_{S_m}$ and $ObjSentenceScore_{S_m}$ were estimated according to SubjCrit2 in (3) and ObjMetrComb2 in (5), respectively. As can be seen in Fig. 7, $Class_{m,l} = \left\{ W_{m,l}, CL_{m,l}, PV_{m,l} \right\}$ leads to a higher subjective-objective score correlation than

$Class_{m,l} = \{W_{m,l}, CL_{m,l}\}$ and $Class_{m,l} = \{W_{m,l}, PV_{m,l}\}$ with both MCS mean and majority vote rule.

Figure 8 shows the subjective-objective score correlation vs. threshold D_{min}^{PV} defined in subsection 3.1.2. As explained above, D_{min}^{PV} defines a minimum similarity threshold between the target pronunciation and the phonetic variance due to the fact that the ASR technology accuracy imposes a maximum discrimination ability. $CL_{m,l}$ was generated with *MNCW*=5 and $\left[D_{min}^{CL} = 8;\ D_{max}^{CL} = 25\right]$, and $PV_{m,l} = \{PV_{m,l}^{3}\}$. $SubjSentenceScore_{S_m}$ and $ObjSentenceScore_{S_m}$ were estimated according to SubjCrit2 in (3) and ObjMetrComb2 in (5), respectively According to Figure 8, the maximum subjective-objective score correlation is achieved when $D_{min}^{PV} = 1$ with the MCS mean and majority vote rule. D_{min}^{PV} can introduce an increases of 10.8% and 6.1% in the subjective-objective score correlation with MCS mean and majority vote rule, respectively, when results in Fig. 8 are compared with those in Fig. 7.

Experiments in Figs. 9-13 are similar to those in Figs. 4-8 except from the fact that $SubjSentenceScore_{S_m}$ and $ObjSentenceScore_{S_m}$ were estimated according to SubjCrit1 in (2) and ObjMetrComb1 in (4), respectively. In contrast to Fig. 4, Figure 9 shows that the optimum subjective-objective score correlation is achieved when *MNCW* is equal to 10 with both the MCS mean and majority vote (MVR) rules. In Fig. 10 the highest subjective-objective score correlation takes place when $D_{min}^{CL} = 8$ and $D_{max}^{CL} = 25$ with both MCS mean

and majority vote rule to estimate $ObjSentenceScore_{S_m}$. According to Fig. 11, the highest subjective-objective score correlation with both MCS mean and majority vote rule takes place with $PV_{m,l} = \left\{PV_{m,l}^{1}\right\}$. This result is not consistent with that achieved in Fig. 6 and could be due to the fact that $PV_{m,l}^{1}$ models better extreme pronunciation mistakes than $PV_{m,l}^{3}$, which in turn enhances ObjMetrComb1 based on the lowest objective word score. As can be seen in Fig. 12, $Class_{m,l} = \left\{W_{m,l}, CL_{m,l}\right\}$ leads to a higher subjective-objective score correlation than $Class_{m,l} = \left\{W_{m,l}, CL_{m,l}, PV_{m,l}\right\}$ and $Class_{m,l} = \left\{W_{m,l}, PV_{m,l}\right\}$, where $PV_{m,l} = \left\{PV_{m,l}^{1}\right\}$, with both MCS mean and majority vote rule. On other words, the introduction of phonetic variants in $Class_{m,l} = \left\{W_{m,l}, CL_{m,l}\right\}$ does not lead to any improvement with SubjCrit1/ObjMetrComb1. In Fig. 13 minimum bound D_{min}^{PV} as defined in subsection (3.1.2) is evaluated when $Class_{m,l} = \left\{W_{m,l}, CL_{m,l}, PV_{m,l}\right\}$ and $PV_{m,l} = \left\{PV_{m,l}^{1}\right\}$. When D_{min}^{PV} increases $Class_{m,l} = \left\{W_{m,l}, CL_{m,l}, PV_{m,l}\right\}$ tends to $Class_{m,l} = \left\{W_{m,l}, CL_{m,l}\right\}$. As a consequence, Fig. 13 corroborates the result obtained in Fig. 12.

Experiments in Figs. 14-18 are similar to those in Figs. 4-8 and Figs. 9-13 except from the fact that $SubjSentenceScore_{S_m}$ and $ObjSentenceScore_{S_m}$ were estimated according to SubjCrit3 in subsection 2.1 and ObjMetrComb3 in (6), respectively. Figure 14 suggests that the subjective-objective score correlation is highly dependent on *MNCW* with SubjCrit3 - ObjMetrComb3, and with a maximum when *MNCW* is equal to 15 with both

the MCS mean and majority vote rules. However, the maximum subjective-objective score correlation in Fig. 14 is much lower than those in Figs. 4 and 9. In Fig. 15 the highest subjective-objective score correlation takes place when $D_{min}^{CL}=8$ and $D_{max}^{CL}=25$ with both MCS mean and majority vote rule to estimate $ObjSentenceScore_{S_m}$. According to Fig. 16, the highest subjective-objective score correlation is observed with $PV_{m,l}=\left\{PV_{m,l}^{4}\right\}$ and $PV_{m,l}=\left\{PV_{m,l}^{3}\right\}$ with MCS mean and majority vote rules, respectively. As can be seen in Fig. 17, $Class_{m,l}=\left\{W_{m,l},CL_{m,l},PV_{m,l}\right\}$, where $PV_{m,l}=\left\{PV_{m,l}^{4}\right\}$, leads to a higher subjective-objective score correlation than $Class_{m,l}=\left\{W_{m,l},CL_{m,l}\right\}$ and $Class_{m,l}=\left\{W_{m,l},PV_{m,l}\right\}$ with both MCS mean and majority vote rule. As shown in Figure 18, the introduction of lower bound D_{min}^{PV} does not lead to any improvement in subjective-objective score correlation when compared with $Class_{m,l}=\left\{W_{m,l},CL_{m,l},PV_{m,l}\right\}$ in Fig. 17. It is worth highlighting that the optimum subjective-objective scores in Figs 14-18 are lower than those in Figs 4-8 and Figs. 9-13. This result should be a consequence of the fact that SubjCrit3, which corresponds to the first impression, is more difficult to model than SubjCrit1 and SubjCrit2 by combining objective word scores.

In Figs. 19 and 20 each subjective criterion is modeled with ObjMetrComb1, ObjMetrComb2 and ObjMetrComb3 by using MCS mean and majority vote rules, respectively. As can be seen in Figs. 19 and 20, ObjMetrComb1 provides a higher correlation with SubjCrit1 than ObjMetrComb2 and ObjMetrComb3. In a similar analysis,

SubjCrit2 is more precisely approximated with ObjMetrComb2 and ObjMetrComb3. This result suggests that ObjMetrComb2 (mean) and ObjMetrComb3 (mode) gives similar objective sentence scores, which in turn should be due to the fact that the subjective word scores within a sentence shows a low standard deviation. Similarly, ObjMetrComb2 and ObjMetrComb3 present higher correlations with SubjCrit3 than ObjMetrComb1. Also ObjMetrComb2 gives a slightly better approximation of SubjCrit3 than ObjMetrComb3.

VI. Conclusion

The problem of assessing pronunciation quality of sentences in 2LL without the need of a priori studies of common mistakes is addressed in this paper. The proposed technique combines objective word scores to approximate subjective evaluations of sentences. First, the subjective evaluation of sentences is defined with three criteria: the minimum subjective word score; the mean subjective word score; and, first impression. Second, three combinations of objective word scores are proposed to emulate the subjective criteria: min; mean; and, mode. Then, class based language models are used to incorporate student's mother and target language phonetic rules to represent wrong and correct pronunciations.

The highest subjective-objective score correlation for sentences is achieved when the subjective evaluation criterion based on the mean word-based subjective score is emulated with the mean word-based objective score. In this scenario the subjective-objective score correlation for sentences is equal to 0.5 with five levels of pronunciation quality. This

correlation is very comparable with that achieved elsewhere with a priori studies of pronunciation errors, which in turn also requires a complete definition of the target utterances. As a consequence, the integration of new target sentences with the ASR based pronunciation quality evaluation technology is more efficient and requires less human assistance with the approach proposed in this paper. A slightly worse result is achieved when the subjective evaluation criterion based on the minimum word-based subjective score is emulated with the minimum word-based objective score. The most difficult subjective criterion to model is the first impression. In this case, the mode and mean of word-based objective scores provided subjective-objective score correlations for sentences equal to 0.40 and 0.44, respectively. This must be due to the fact that the model represented by the combination of word-based subjective evaluations is less applicable to the first impression criteria. Improving the discrimination ability provided by ASR based technology between correct and wrong pronunciations, exploring the correlation of word-based subjective scores within utterances, proposing more accurate models for the subjective first impression criteria and proposing new subjective criterion (e.g. semantic) are proposed for future research.

VII. References

A., Narayanan, Sh., 2007. A Bayesian network classifier for word-level reading assessment. In: Proc. InterSpeech ICSLP, August, Antwerp, Belgium.

Cincarek T, Gruhn R., Hacker C., Nöth E., Nakamura S. "Automatic pronunciation scoring of words and sentences independent from the non-native's first language" 2008

Cucchiarini C., Strik H., Binnenpoorte D., Boves L. Pronunciation Evaluation in read and spontaneous speech: A comparison between human ratings and automatic scores, 1997.

Cucchiarini C., Strik H., Boves L."Automatic evaluation of dutch pronunciation by using speech recognition technology"IEEE 1997.

Deshmukh Om., Joshi S., Verma A."Automatic Pronunciation Evaluation and Classification", IBM India Research Lab, New Delhi, India, 2008

Franco H., Neumeyer L., Digalakis V., and Ronen O. "Combination of Machine scores for automatic grading of pronunciation quality". SRI International 1998.

Franco H., Neumeyer L., Kim Y., and Ronen O., "Automatic pronunciation scoring for language instruction," in Proc. of ICASSP,1997.

Fumera G. and Roli F., 2005. **"**A theoretical and experimental analysis of linear combiners for multiple classifier systems," IEEE Trans. Pattern Analysis and Machine Intelligence. 27(6). pp. 942-956.

Kittler J. and Alkoot F.M., 2003. "Sum versus vote fusion in multiple classifier systems," IEEE Trans. Pattern Analysis and Machine Intelligence. 25, pp. 110-115.

Kittler, J., Hatef, M., Duin, R.P.W., Matas, J., 1998. "On combining classifiers," IEEE Trans. on Pattern Analysis and Machine Intelligence. 20, pp. 226-239.

Kuncheva, L. I., Bezdeck, J. C. and Duin, R. P. W. 2001. "Decision templates for multiple classifier fusion: an experimental comparison", Pattern Recognition, vol. 34(2), pp. 299-314.

Kuncheva, L.I., 2002. "A theoretical study on six classifier fusion strategies," IEEE Trans. Pattern Analysis and Machine Intelligence,vol. 24, pp. 281-286.

Kwan K.Y., Lee T., Yang C., 2002. "Unsupervised N-best based model adaptation using model-level confidence measures. In: Proc. ICSLP, pp. 69-72.

Molina C., Yoma N.B., Wuth J., & Vivanco H. 2008. ASR based pronunciation evaluation with automatically generated competing vocabulary and classifier fusion. Speech communication.

Moustroufas N., Digalakis V.. "Automatic pronunciation evaluation of foreign speakers using unknown text". 2006

Nakagawa, S., Ohta, K., 2007. A statistical method of evaluating pronunciation proficiency for resentation in English. In: Proc. InterSpeech ICSLP, August, Antwerp, Belgium.

Neri, A., Cucchiarini, C., Strik, W., 2003. Automatic speech recognition for second language learning: how and why it actually works. In: Proc. 15th Internat. Congress of Phonetic Sciences, Barcelona, Spain, pp. 1157–1160.

Neumeyer L., Franco H., Digalakis V., Weintraub M. "Automatic scoring of pronunciation quality" Speech Communication 2000.

Neumeyer, L., Franco, H., Weintraub, M., Price, P., 1996. Automatic text-independent pronunciation scoring of foreign language student speech. In: Proc. ICSLP'96.

Sooful J., and Botha E., 2002. "Comparison of acoustic distance measures for automatic cross-language phoneme mapping," in Proc. of ICSLP, pp. 521-524. Denver, USA.

Tepperman, J., Black, M., Price, P., Lee, S., Kazemzadeh, A., Gerosa, M., Heritage, M., Alwan, Ward W., Issar S., "A Class Based Language Model for Speech Recognition", 1996.

Wei S., Pan Y., Hu G., Hu Y. and Wang R. "Pronunciation space models for pronunciation evaluation".IEEE 2008.

Xu S., Ke D., Jiang J., Yang X., Li H., Xu B.. "Automatic Pronunciation Evaluation Based on Feature Extraction and Combination" IEEE 2008.

Zhang J., Ward W., Pellom B., Yu X., Hacioglu K. "Improvements in Audio Processing and Language Modeling in the CU Communicator", 2001.

Table 1: English phonemes are replaced with the most similar Spanish phonemes to generate phonetic variant $PV^1_{m,l}$.

English	Spanish	English	Spanish
Ah, Ae,	a_s	Ow	o_s u_s
Aa, Ao	o_s	P	p_s
B, V	b_s	R	r_s
Ch, Sh	ch_s	S, Z, Th	s_s
D, Dh	d_s	Zh, Jh, Y	y_s
Eh, Er	e_s	T	t_s
F	f_s	Uh, Uw	u_s
G	g_s	Oy	o_s i_s
Ih, Iy	i_s	Aw	a_s u_s
Hh	j_s	W	g_s u_s
K	k_s	Ng	n_s g_s , n_s
L	l_s	Ay	a_s i_s
M	m_s	Ey	e_s i_s
N	n_s		

Table 2: Spanish phonemes are replaced with the most similar English phonemes to generate phonetic variant $PV^3_{m,l}$.

Spanish	English	Spanish	English
a_s	Ah	m_s	M
b_s	B	n_s	N
ch_s	Ch	o_s	Aa
d_s	D, Dh	p_s	P
e_s	Eh	r_s	R
f_s	F	s_s	S
g_s	G	t_s	T
i_s	Ih	u_s	Uh
j_s	Hh	w_s	W
k_s	K	x_s	KS
l_s	L	y_s	Y

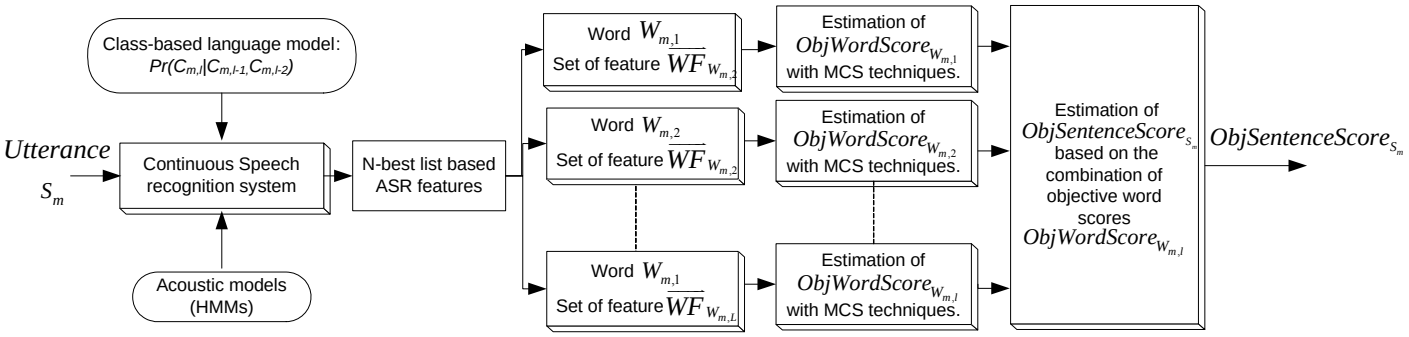

Figure 1: Block diagram of the proposed method for pronunciation evaluation of sentences.

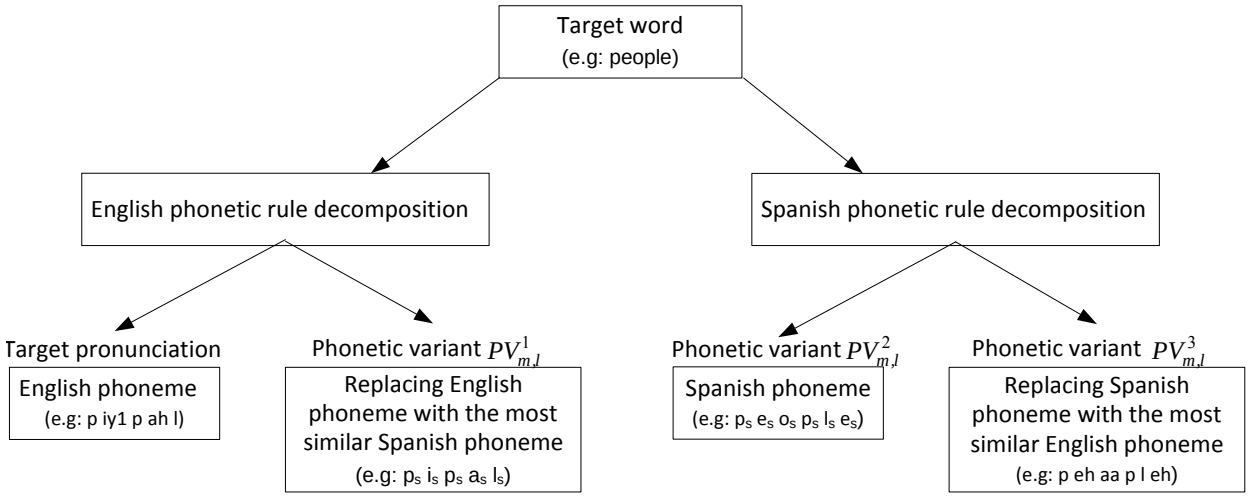

Figure 2: Generation of phonetic variants.

Figure 3: Web based 2LL system employed to record the database.

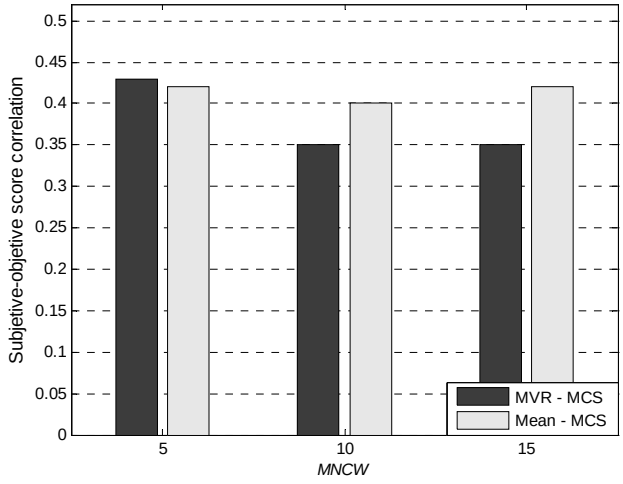

Figure 4: Subjective-objective score correlation when $Class_{m,l} = \{W_{m,l}, CL_{m,l}\}$ and *MNCW* is made equal to 5, 10 and 15 words. $SubjSentenceScore_{S_m}$ and $ObjSentenceScore_{S_m}$ were estimated according to SubjCrit2 and ObjMetrComb2.

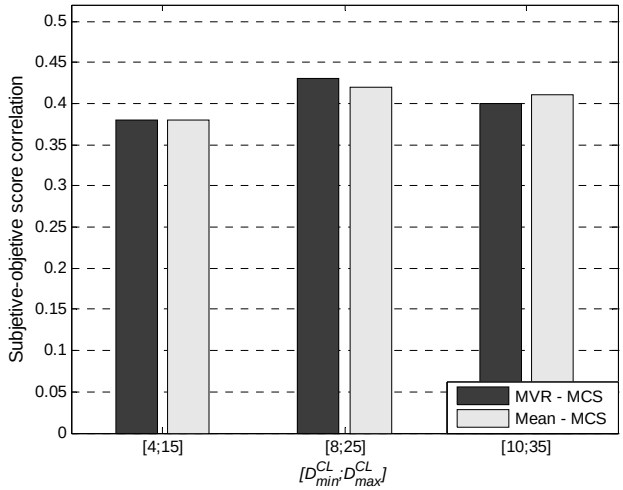

Figure 5: Subjective-objective score correlation when $Class_{m,l} = \{W_{m,l}, CL_{m,l}\}$ and *MNCW* was made equal to 5 competitive words. Three pairs $\left[D_{min}^{CL}; D_{max}^{CL}\right]$ were evaluated: $[4; 15]$; $[8; 25]$; and $[10; 35]$. $SubjSentenceScore_{S_m}$ and $ObjSentenceScore_{S_m}$ were estimated according to SubjCrit2 and ObjMetrComb2.

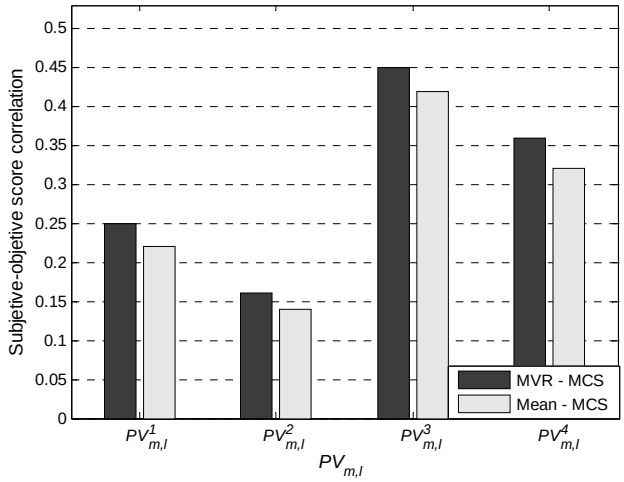

Figure 6: Subjective-objective score correlation when $Class_{m,l} = \{W_{m,l}, PV_{m,l}\}$. Four configurations of $PV_{m,l}$ were evaluated: $PV_{m,l} = \{PV_{m,l}^1\}$; $PV_{m,l} = \{PV_{m,l}^2\}$; $PV_{m,l} = \{PV_{m,l}^3\}$; and, $PV_{m,l} = \{PV_{m,l}^4\}$. $SubjSentenceScore_{S_m}$ and $ObjSentenceScore_{S_m}$ were estimated according to SubjCrit2 and ObjMetrComb2.

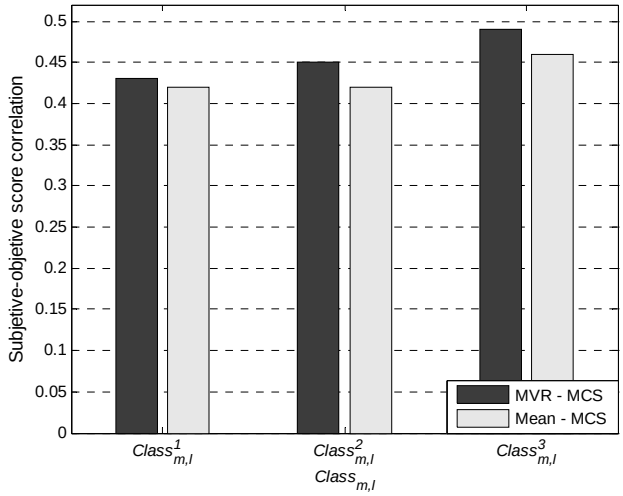

Figure 7: Subjective-objective score correlation when $Class_{m,l}$ is made equal to: $Class^{1}_{m,l} = \{W_{m,l}, CL_{m,l}\}$; $Class^{2}_{m,l} = \{W_{m,l}, PV^{3}_{m,l}\}$; and, $Class^{3}_{m,l} = \{W_{m,l}, CL_{m,l}, PV^{3}_{m,l}\}$. $CL_{m,l}$ was generated with *MNCW*=5 and $\left[D^{CL}_{min} = 8;\ D^{CL}_{max} = 25\right]$. $SubjSentenceScore_{S_m}$ and $ObjSentenceScore_{S_m}$ were estimated according to SubjCrit2 and ObjMetrComb2.

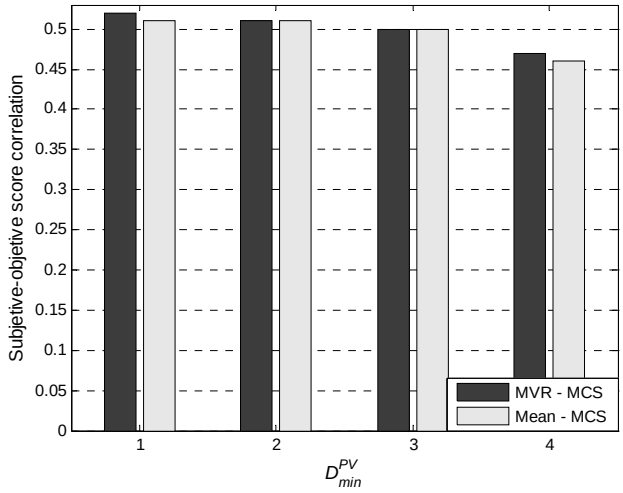

Figure 8: subjective-objective score correlation vs. threshold D^{PV}_{min}. $CL_{m,l}$ was generated with *MNCW*=5 and $\left[D^{CL}_{min} = 8;\ D^{CL}_{max} = 25\right]$. $PV_{m,l} = \{PV^{3}_{m,l}\}$. $SubjSentenceScore_{S_m}$ and $ObjSentenceScore_{S_m}$ were estimated according to SubjCrit2 and ObjMetrComb2.

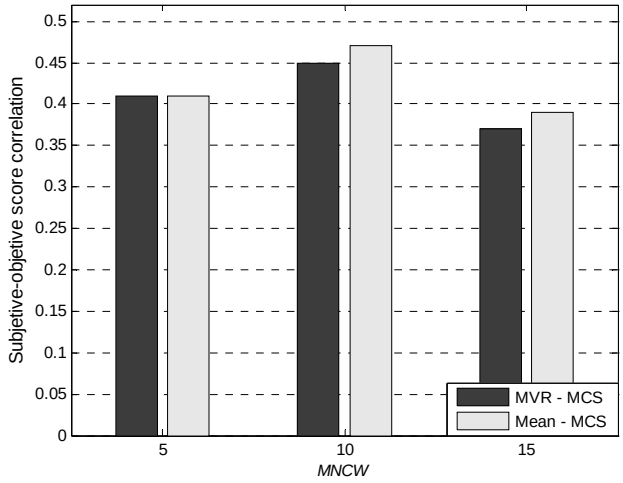

Figure 9: Subjective-objective score correlation when $Class_{m,l} = \{W_{m,l}, CL_{m,l}\}$ and *MNCW* is made equal to 5, 10 and 15 words. $SubjSentenceScore_{S_m}$ and $ObjSentenceScore_{S_m}$ were estimated according to SubjCrit1 and ObjMetrComb1.

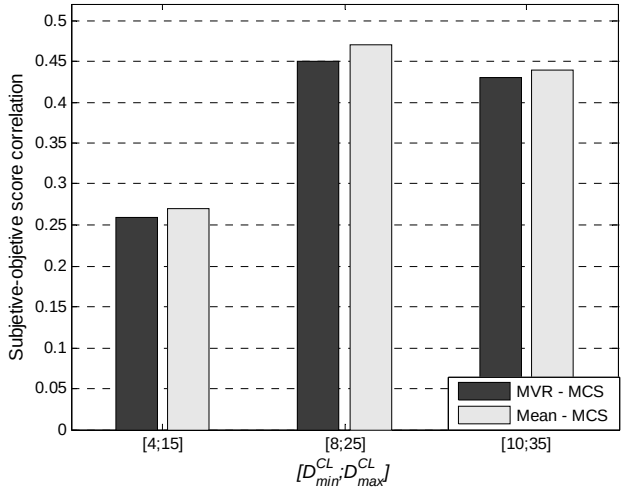

Figure 10: Subjective-objective score correlation when $Class_{m,l} = \{W_{m,l}, CL_{m,l}\}$ and *MNCW* was made equal to 10 competitive words. Three pairs $\left[D_{min}^{CL}; D_{max}^{CL}\right]$ were evaluated: $[4; 15]$; $[8; 25]$; and $[10; 35]$. $SubjSentenceScore_{S_m}$ and $ObjSentenceScore_{S_m}$ were estimated according to SubjCrit1 and ObjMetrComb1.

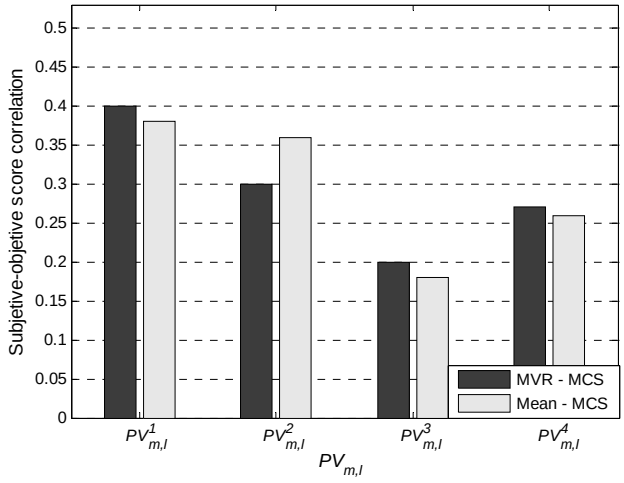

Figure 11: Subjective-objective score correlation when $Class_{m,l} = \{W_{m,l}, PV_{m,l}\}$. Four configurations of $PV_{m,l}$ were evaluated: $PV_{m,l} = \{PV^1_{m,l}\}$; $PV_{m,l} = \{PV^2_{m,l}\}$; $PV_{m,l} = \{PV^3_{m,l}\}$; and, $PV_{m,l} = \{PV^4_{m,l}\}$. $SubjSentenceScore_{S_m}$ and $ObjSentenceScore_{S_m}$ were estimated according to SubjCrit1 and ObjMetrComb1.

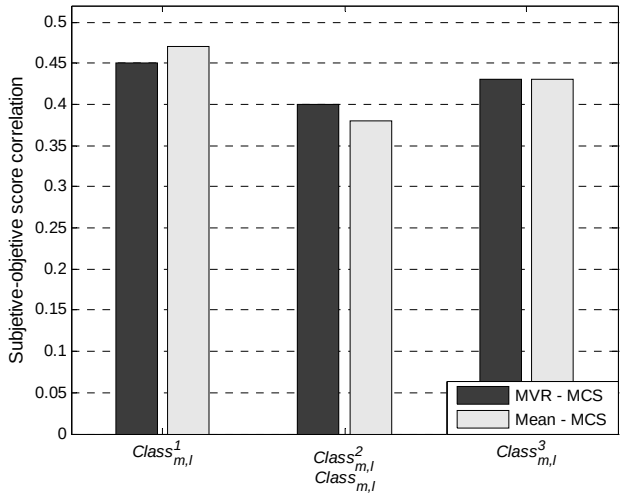

Figure 12: Subjective-objective score correlation when $Class_{m,l}$ is made equal to: $Class^1_{m,l} = \{W_{m,l}, CL_{m,l}\}$; $Class^2_{m,l} = \{W_{m,l}, PV^1_{m,l}\}$; and, $Class^3_{m,l} = \{W_{m,l}, CL_{m,l}, PV^1_{m,l}\}$. $CL_{m,l}$ was generated with $MNCW$=10 and $\left[D^{CL}_{min} = 8;\ D^{CL}_{max} = 25\right]$. $SubjSentenceScore_{S_m}$ and $ObjSentenceScore_{S_m}$ were estimated according to SubjCrit1 and ObjMetrComb1.

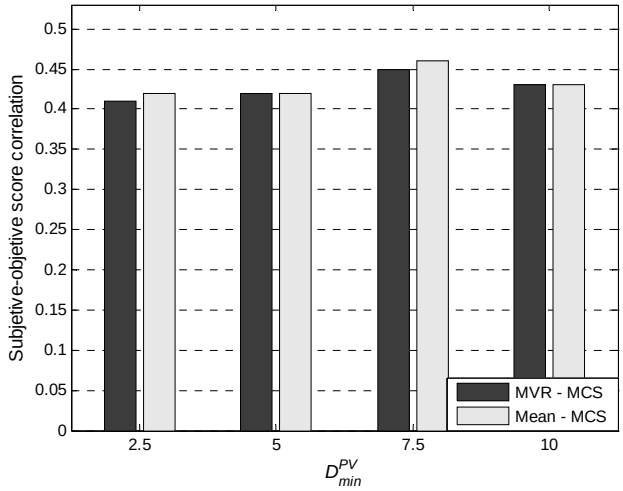

Figure 13: Subjective-objective score correlation vs. threshold D_{min}^{PV}. $CL_{m,l}$ was generated with *MNCW*=10 and $\left[D_{min}^{CL}=8;\ D_{max}^{CL}=25\right]$. $PV_{m,l}=\left\{PV_{m,l}^{1}\right\}$. $SubjSentenceScore_{S_m}$ and $ObjSentenceScore_{S_m}$ were estimated according to SubjCrit1 and ObjMetrComb1.

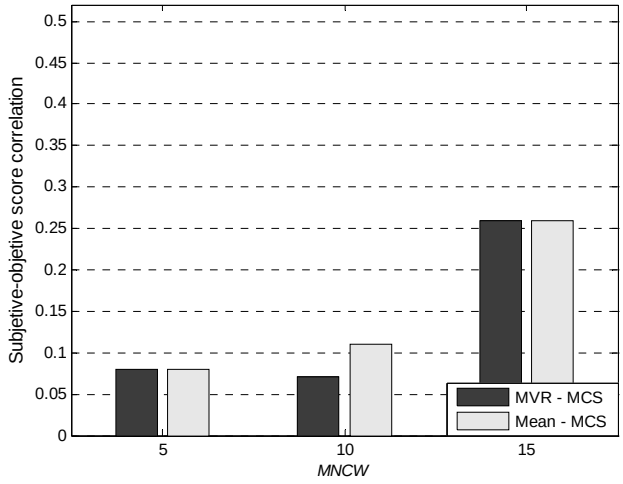

Figure 14: Subjective-objective score correlation when $Class_{m,l}=\left\{W_{m,l},CL_{m,l}\right\}$ and *MNCW* is made equal to 5, 10 and 15 words. $SubjSentenceScore_{S_m}$ and $ObjSentenceScore_{S_m}$ were estimated according to SubjCrit3 and ObjMetrComb3.

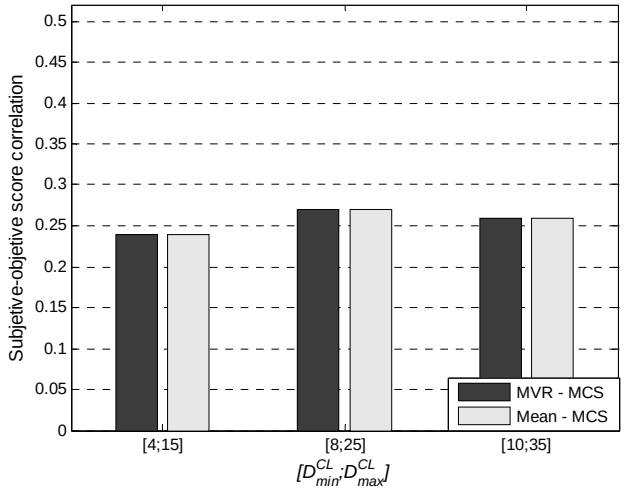

Figure 15: Subjective-objective score correlation when $Class_{m,l} = \{W_{m,l}, CL_{m,l}\}$ and *MNCW* was made equal to 15 competitive words. Three pairs $[D_{min}^{CL}; D_{max}^{CL}]$ were evaluated: $[4; 15]$; $[8; 25]$; and $[10; 35]$. $SubjSentenceScore_{S_m}$ and $ObjSentenceScore_{S_m}$ were estimated according to SubjCrit3 and ObjMetrComb3.

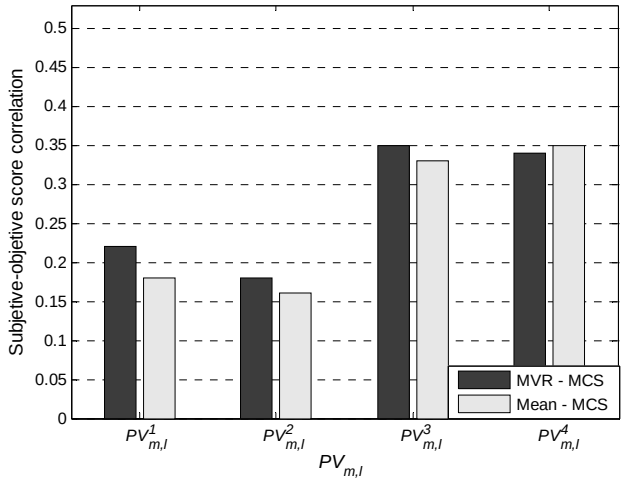

Figure 16: Subjective-objective score correlation when $Class_{m,l} = \{W_{m,l}, PV_{m,l}\}$, Four configurations of $PV_{m,l}$ were evaluated: $PV_{m,l} = \{PV_{m,l}^1\}$; $PV_{m,l} = \{PV_{m,l}^2\}$; $PV_{m,l} = \{PV_{m,l}^3\}$; and, $PV_{m,l} = \{PV_{m,l}^4\}$. $SubjSentenceScore_{S_m}$ and $ObjSentenceScore_{S_m}$ were estimated according to SubjCrit3 and ObjMetrComb3.

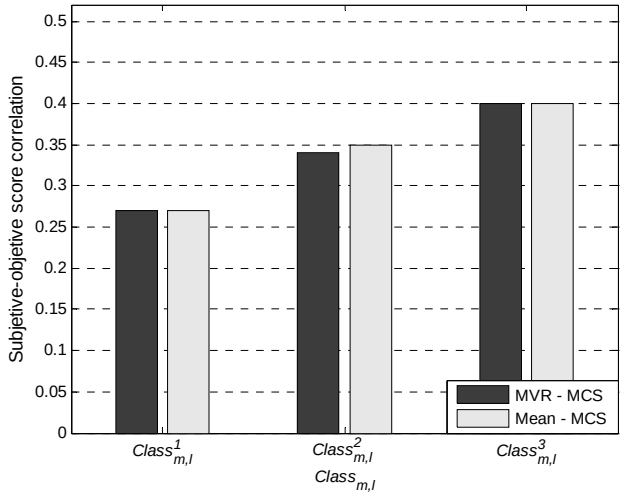

Figure 17: Subjective-objective score correlation when $Class_{m,l}$ is made equal to: $Class^1_{m,l} = \{W_{m,l}, CL_{m,l}\}$; $Class^2_{m,l} = \{W_{m,l}, PV^4_{m,l}\}$; and, $Class^3_{m,l} = \{W_{m,l}, CL_{m,l}, PV^4_{m,l}\}$. $CL_{m,l}$ was generated with *MNCW*=15 and $\left[D^{CL}_{min} = 8;\ D^{CL}_{max} = 25\right]$. $SubjSentenceScore_{S_m}$ and $ObjSentenceScore_{S_m}$ were estimated according to SubjCrit3 and ObjMetrComb3.

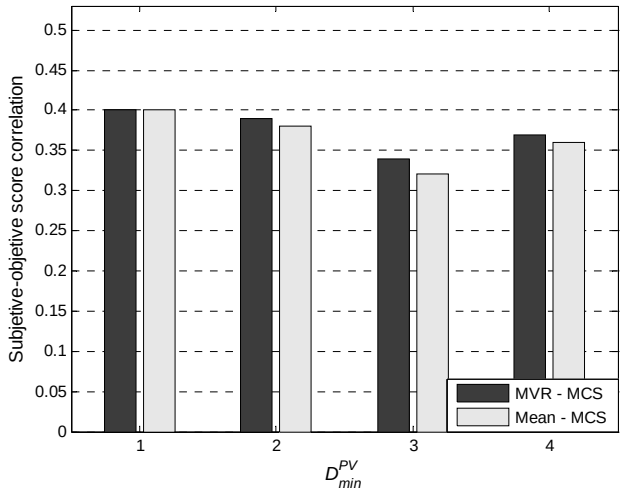

Figure 18: Subjective-objective score correlation vs. threshold D^{PV}_{min}. $CL_{m,l}$ was generated with *MNCW*=15 and $\left[D^{CL}_{min} = 8;\ D^{CL}_{max} = 25\right]$. $PV_{m,l} = \{PV^4_{m,l}\}$. $SubjSentenceScore_{S_m}$ and $ObjSentenceScore_{S_m}$ were estimated according to SubjCrit3 and ObjMetrComb3.

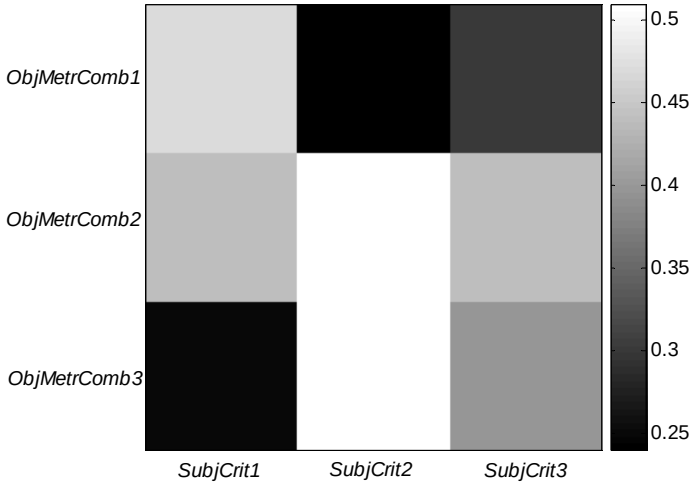

Figure 19: Subjective-objective score correlation with MCS mean rule. Each subjective criterion is modeled with ObjMetrComb1, ObjMetrComb2 and ObjMetrComb3. In each case the best configuration was employed.

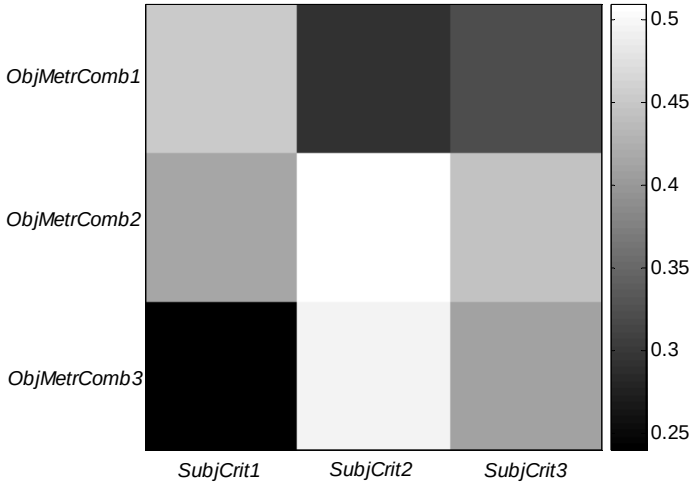

Figure 20: Subjective-objective score correlation with MCS MVR. Each subjective criterion is modeled with ObjMetrComb1, ObjMetrComb2 and ObjMetrComb3. In each case the best configuration was employed.