3. Adaptación y Normalización

En este capítulo se presentan algunas técnicas que ayudan a mejorar el nivel de reconocimiento o a evitar la degradación en el reconocimiento de voz. Particularmente presentaremos una técnica de normalización de la longitud del tracto vocal (VTLN) y dos técnicas de adaptación que son regresiones lineales de probabilidad máxima *a posteriori* (MAPLR), y regresiones lineales de probabilidad máxima (MLLR).

Uno de los campos de aplicación de los sistemas de reconocimiento de voz es la educación [12,13,14]. Normalmente los usuarios de dichos sistemas son niños. Los reconocedores de voz de propósito general muestran una clara degradación cuando son usados con niños, esto es debido a las diferencias fonéticas y fonológicas de los locutores. Otra razón por la que un reconocedor puede bajar su desempeño es por las diferencias que hay entre los datos de entrenamiento y los datos de evaluación, ya sea por falta de datos para entrenar o la alta variabilidad de los locutores [15,16].

3.1 Reconocimiento Robusto de Voz

Existen diferentes obstáculos para lograr que un reconocedor sea robusto [16,18,19,20,21,22,23,24]. Algunos de estos obstáculos no dependen solamente del sistema sino de la detección de los problemas que degradan el desempeño y la aplicación correcta de las soluciones para nulificar o minimizar dicha degradación en un reconocedor de voz.

Hay métodos que se enfocan a reducir el rango de error de palabras (Word Error Rate), otros mejoran la precisión (accuracy) del reconocimiento modificando elementos en cada uno de los componentes del reconocedor de voz o mejorando la calidad de los datos de entrenamiento.

Los diferentes grupos de investigación han producido tecnología, que, con algunas restricciones, pueden reconocer con precisión el habla. Hoy en día, los sistemas de reconocimiento de voz todavía no pueden igualar el desempeño humano. Aunque se puedan construir reconocedores muy precisos para un locutor en particular, en un idioma y estilo de habla en particular, en un ambiente específico y limitando una tarea en particular, representa un reto construir un reconocedor que pueda esencialmente entender la voz de cualquiera, en cualquier lenguaje, tema, estilo de hablar, y casi en cualquier ambiente [3].

Un sistema de reconocimiento de voz entrenado en el laboratorio con datos (voces) limpios puede degradarse significativamente en el mundo real si los datos limpios usados en el entrenamiento no concuerdan con los datos del mundo real [5].

Robustez en reconocimiento de voz se refiere a la necesidad de mantener una buena precisión en el reconocimiento aún cuando la calidad de la voz de entrada esté degradada, o cuando las características acústicas, articulatorias o fonéticas de la voz en los ambientes de entrenamiento y evaluación difieran [25,26]. Para ser más precisos, los obstáculos que no permiten que un sistema sea robusto incluyen degradaciones acústicas producidas por ruido aditivo, efectos de filtración lineal, fenómenos en transducción o transmisión, así como fuentes impulsivas de interferencias, y precisión disminuida causada por cambios en la articulación producida por la presencia de fuentes de ruido de alta intensidad [27]. Las

diferencias de locutor a locutor imponen un distinto tipo de variabilidad, produciendo variaciones en el rango de voz, coarticulación, contexto y dialecto [28].

3.1.1 Variabilidad del Locutor

Cada locutor es diferente, la voz que produce refleja el tamaño físico del tracto vocal, longitud y anchura del cuello, un rango de características físicas, edad, sexo, dialecto, salud, educación, y estilo personal [29]. Aún si excluimos estas diferencias entre locutores, el mismo locutor es a menudo incapaz de producir precisamente la misma pronunciación. En la figura 3.1 podemos ver 4 diferentes espectrogramas de la vocal /**a**/ de niños de 8 y 9 años de edad. Se puede ver en la figura que los dos primeros pertenecen a hombres con distinta edad. Los siguientes dos pertenecen a mujeres de la misma edad. La pronunciación de la misma vocal es diferente en cada uno de los casos.

3.1.2 Variabilidad del entorno

La mayoría de los lugares están llenos de sonidos a distinto volumen con diferentes fuentes. Cuando interactuamos con computadoras, podemos tener a personas hablando en el fondo. Alguien puede azotar la puerta, o el aire acondicionado puede estar puesto sin notarlo. Si el reconocimiento de voz es implantado en dispositivos móviles, como PDAs (asistentes personales digitales) o teléfonos celulares, el espectro de ruidos varía significativamente porque el dueño se mueve alrededor [30]. Estos parámetros externos, como las características del ruido ambiental y el tipo y colocación del micrófono, pueden realmente afectar el desempeño del sistema de reconocimiento de voz [31]. Además de los

ruidos de fondo, tenemos que tratar con el ruido hecho por los locutores, como ruido con los labios y palabras que no significan nada.

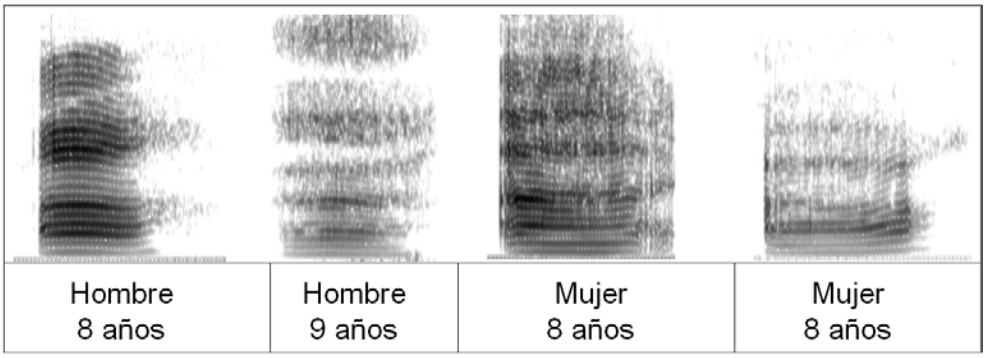

Figura 3.1 Espectrogramas para la vocal /a/ de 4 diferentes niños

3.1.3 Variabilidad del contexto

La interacción del lenguaje hablado entre personas requiere conocimiento del significado de las palabras, contexto de la comunicación, y sentido común. Las palabras con diferentes significados y patrones de uso pueden tener la misma realización fonética. Además de la variabilidad a nivel palabra y oración, podemos encontrar variabilidad a nivel fonético.

3.1.4 Adaptación Dinámica de Parámetros

La adaptación dinámica ya sea de las características que son la entrada en el sistema de reconocimiento, o de las representaciones del sistema almacenadas internamente de las posibles pronunciaciones, es el enfoque más directo hacia la adaptación del ambiente y del

locutor [32]. Existen tres diferentes enfoques de adaptación del locutor y del ambiente: (1) el uso de los procedimientos óptimos de estimación para obtener nuevos valores de los parámetros en las condiciones de evaluación; (2) el desarrollo de procedimientos de compensación basados en comparaciones empíricas de voz en ambientes de entrenamiento y evaluación; y (3) el uso de filtros de los valores de los parámetros para mejorar la robustez [3].

La solución de los problemas de estimación frecuentemente requiere aproximaciones analíticas o numéricas o el uso de técnicas iterativas de estimación, como el algoritmo EM[1] (estimar-maximizar). Estos enfoques han sido exitosos en aplicaciones donde las suposiciones de los modelos son razonablemente válidas, pero están limitados en algunos casos por la complejidad computacional.

Se pueden usar las comparaciones empíricas de características derivadas de voz de alta calidad con características de voz que están simultáneamente grabadas bajo condiciones degradadas para compensar las desigualdades entre condiciones de entrenamiento y evaluación.

La mayor técnica de adaptación es el filtrado de paso-alto cepstral, el cual provee una remarcable cantidad de robustez con cero costo computacional [33].

Las técnicas de normalización aplicadas en el entrenamiento de reconocedores automáticos de voz apuntan a separar fonéticamente variaciones relevantes de variaciones irrelevantes causadas por particularidades del locutor o del entorno acústico de los datos de entrenamiento [34,35].

[1] [1, pp. 147-163]

Las técnicas de adaptación pueden ser usadas modificando los parámetros del sistema para mejorar la correspondencia de variaciones en contextos de micrófonos, canal de transmisión, ruido ambiental, locutor, estilo, y aplicación [1].

3.2 Normalización de la Longitud del Tracto Vocal (VTLN)

Las variaciones en el locutor, y en el entorno son los mayores retos para los sistemas de reconocimiento de voz. El desempeño de un sistema podría cambiar en gran manera debido a estas variaciones en un uso práctico. Además, cómo hacer posible que los sistemas de reconocimiento de voz sean los más precisos y robustos, este es uno de los mayores problemas en reconocimiento de voz. Desde un punto de vista de las técnicas más actuales para reconocimiento de voz, el principal origen de dependencia del locutor está dado desde la señal de voz. La razón de que la señal de voz es dependiente del locutor es muy compleja. No está solamente relacionado con las diferencias fisiológicas de los locutores, como la forma y longitud del tracto vocal, sino también relacionado con las diferencias lingüísticas, como el acento, dialecto y estrés, entre otros, o hasta las condiciones físicas y mentales del locutor. Pero se está generalmente de acuerdo en que uno de los orígenes de variación entre locutores es la longitud del tracto vocal (VTL[2]). Además, la técnica de normalización de la longitud del tracto vocal (VTLN[3]) ha sido muy investigada para eliminar la variación VTL [34, 36, 37, 38, 39].

En general dos puntos están envueltos en VTLN: (1) Dados los datos de voz de un locutor, cómo obtener el factor de normalización; (2) Dado el factor de normalización, cómo hacer

[2] Vocal Tract Length
[3] Vocal Tract Length Normalization

la normalización. Estrictamente hablando, el paso (1) debería ser cómo obtener la longitud del tracto vocal desde un punto de vista exacto de VTLN. Ya que, con el presente método, uno generalmente calcula un factor que refleja las diferencias en el tracto vocal entre diferentes locutores, en lugar de medir directamente la longitud del tracto vocal de cada locutor [39]. Nos referiremos al factor de normalización como factor de deformación (*warping factor*), ya que el factor es usado para deformar el espectro en el eje de la frecuencia con el propósito de normalizarlo. Para obtener el factor, existen básicamente dos métodos, esto es, obteniendo el factor calculando VTL o con una búsqueda de línea.

Se ha mostrado que VTL tiene relación con las posiciones formantes y por lo tanto podría ser calculado con la frecuencia de la formante basado en el modelo lineal predictivo. Las desventajas de este método son:

(a) La frecuencia de la formante y su relación con VTL son altamente dependientes del contexto, y podría variar con un contexto diferente para el mismo locutor.

(b) Es necesario separar y seleccionar los datos que contengan voz, ya que no tiene sentido calcular la frecuencia de la formante basado en datos que no tienen voz (como consonantes y ruidos).

(c) El criterio para calcular el factor no es consistente con el criterio de estimación de los otros parámetros de los modelos acústicos los cuales son de hecho calculados bajo el criterio de entrenamiento de probabilidad máxima (ML), y por lo tanto no garantiza que la normalización con ese factor pueda incrementar la puntuación de correspondencia de ML.

Debido a la falta de una técnica para calcular con alta precisión, independiente del contexto, y robusta, el método para obtener el factor de deformación basado en el cálculo de la longitud del tracto vocal parece difícil para su uso práctico.

Dado lo anterior, el método de búsqueda de línea fue propuesto e investigado con anterioridad en otros estudios. La ventaja de este método es que no necesita considerar la relación entre VTL y la frecuencia formante o la separación de datos con voz o sin voz, y además es consistente con el criterio de entrenamiento del modelo acústico.

Para normalizar la característica de la voz dado un factor de deformación, se han propuesto dos métodos, estos son, la deformación de la frecuencia (*frequency warping*) y deformación de la escala *Bark/Mel*[4].

3.2.1 Preprocesamiento

Se asume que la señal grabada es transmitida vía algún tipo de canal y se recibe en un dispositivo. En el proceso de transmisión y recepción, la señal de voz limpia es alterada por distorsiones en el canal y con algunos ruidos aditivos. Generalmente, la distorsión se multiplica en el dominio de la frecuencia, así que la señal recibida puede ser expresada como:

$$X(\mathrm{w}) = H(\mathrm{w})S(\mathrm{w}) + N(\mathrm{w}) \tag{3.1}$$

Donde $X(\mathrm{w})$, $S(\mathrm{w})$, $H(\mathrm{w})$, y $N(\mathrm{w})$ son el espectro recibido de la señal, la señal de voz limpia, la respuesta del canal, y la señal con ruido aditivo respectivamente. En el proceso del banco de filtro Bark/Mel, $X(\mathrm{w})$ es integrado con un banco de filtros[5]de paso de banda,

[4] Este tema forma parte del análisis de frecuencias el cual no se verá en este estudio, para mayor información se puede consultar [5] pp. 32,34.

[5] Un banco de filtros es una colección de filtros que atraviesan todo el espectro de la frecuencia [5].

espaciado de acuerdo a la escala Bark/Mel, y normalmente tiene una forma triangular o trapezoidal. La integración con el banco de filtros puede ser formulada como sigue:

$$O(n) = \sum_{w=l_n}^{w=h_n} T_n(w)X(w) \qquad 0 \le n \le N-1 \tag{3.2}$$

Donde O(n) es la n-ésima salida del banco de filtros, N es el número de filtros, l_n y h_n son los límites inferior y superior del n-ésimo filtro $T_n(\omega)$. El ancho de banda de cada $T_n(\omega)$, i.e., $h_n - l_n$, dependen de la escala Bark/Mel.

3.2.2 VTLN Basada en la Deformación de la Frecuencia

En la Figura 3.1 podemos ver el diagrama de la VTLN basado en el método de la deformación de la frecuencia [39].

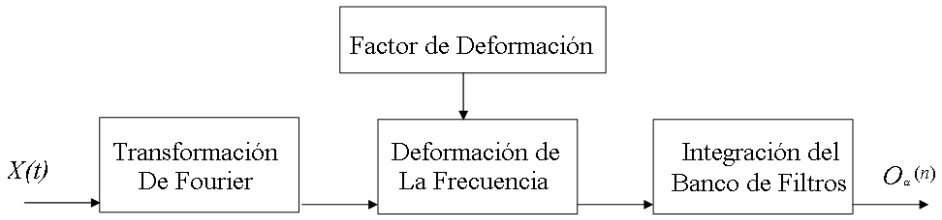

Figura 3.2 Deformación de la Frecuencia VTLN

Donde x(t) es la entrada de la señal, $O_\alpha(n)$ es la n-ésima salida del banco de filtros, y α es el factor de deformación. Se debe notar que α es dependiente del locutor. La figura 3.1 puede ser expresada como:

$$O_\alpha(n) = \sum_{\omega=l_n}^{\omega=h_n} T_n(\omega)X(\varphi_\alpha(\omega)) \qquad 0 \le n \le N-1 \tag{3.3}$$

Donde $\varphi_\alpha(\omega)$ es la función de deformación. Si $\varphi_\alpha(\omega) = \omega$, entonces la ecuación (3.3) es igual a la ecuación (3.2) lo cual significa que no hay deformación. Comparado a *O*(n) en la ecuación (3.2), $O_\alpha(n)$ depende del factor de deformación en un locutor específico y la regla de deformación. Por lo general se usan dos reglas en este método:

(1) regla de intervalos discretos (*piecewise*)*:*

$$\varphi_\alpha(\omega) = \begin{cases} \alpha^{-1}\omega & si \quad \omega < \omega_0 \\ b\omega + c & si \quad \omega \geq \omega_0 \end{cases} \tag{3.4}$$

(2) regla bilineal:

$$\varphi_\alpha(\omega) = \omega + 2\tan^{-1}\left(\frac{(1-\alpha)\sin(\omega)}{1-(1-\alpha)\cos(\omega)}\right) \tag{3.5}$$

Donde α es el factor de deformación de un locutor específico, ω_0 en la ecuación (3.4) es una frecuencia fija la cual se pone para manejar el problema de desigualdad de ancho de banda, y *b*, *c* pueden ser calculados con W_0 conocido. Se debe notar que de acuerdo a la ecuación (3.4) y (3.5), α >1.0 corresponde a estirar o ensanchar el espectro, y α<1.0 corresponde a comprimir el espectro, y α=1.0 corresponde al caso donde no hay deformación.

En las figuras 3.2 y 3.3[6], son presentadas 3 curvas en cada figura, las cuales reflejan el rango de los factores de deformación en el proceso de entrenamiento. Las curvas inferior y superior corresponden a los factores mínimos y máximos, y la de en medio corresponde a la unidad del factor de deformación (sin deformación). En los experimentos que se han conducido en otros trabajos [37,38,39], los locutores femeninos son dominantes en el área entre la curva inferior y la curva media, lo cual corresponde a comprimir el espectro, y los factores de deformación para los locutores masculinos son dominantes en el área entre la curva media y la superior, lo cual corresponde a ensanchar el espectro en el eje de la frecuencia. Esto es consistente con el hecho de que el VTL de las mujeres es generalmente más corto que el de los hombres y las posiciones de las formantes son más altas que el de los hombres.

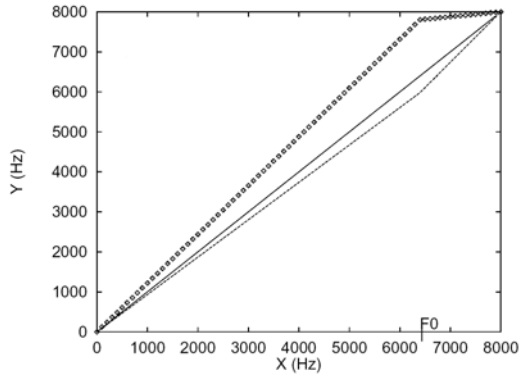

Figura 3.3 Curvas de deformación de intervalos discretos

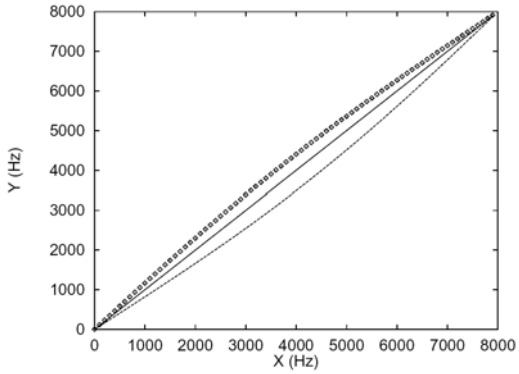

Figura 3.4 Curvas de deformación bilineales

[6] Las gráficas mostradas en estas figuras corresponden a [12].

3.2.3 Procedimiento de Entrenamiento

Suponemos que $O_\alpha(t)$ es la secuencia de vectores de características de la voz de la pronunciación de entrada con un factor α. Sea W la transcripción de la pronunciación de entrada, Λ la densidad de la probabilidad de la característica de voz. Usemos el HMM como modelo acústico, y la densidad de mezclas-gausianas (mixture-gaussian) como la densidad de la probabilidad de salida para cada estado de un HMM. El siguiente procedimiento es usado para entrenar un sistema VTLN:

1. Colocar el factor de deformación inicial α=1.0 para todos los locutores.
2. Hacer un alineamiento de Viterbi con la transcripción W para obtener el mejor segmento de estado S_t^*:

$$S_t^* = \arg\max_{s_t} P(O_\alpha(t), s_t \mid \Lambda, W) \tag{3.6}$$

3. Buscar el mejor factor de deformación en un mapa basado en S_t^*:

$$\alpha^* = \arg\max_{l \le \alpha \le h} P(O_\alpha(t) \mid s_t^*, \Lambda, W) \tag{3.7}$$

4. Hacer una alineación de Viterbi basado en W y el mejor factor de deformación α_t^* para colectar las suficientes estadísticas y actualizar los parámetros del modelo.

$$\Lambda^* = \arg\max_{\Lambda} P(O_{\alpha^*}(t) \mid \Lambda, W) \tag{3.8}$$

5. Igualar $\alpha = \alpha^*$ y $\Lambda = \Lambda^*$, e ir al paso 2.

El procedimiento anterior se detiene si no hay una diferencia significante en los factores de deformación entre dos iteraciones consecutivas de entrenamiento.

3.2.4 Procedimiento de Descifrado (evaluación)

En el procedimiento de descifrado para VTLN, primero la pronunciación de entrada se descifra con el factor de deformación α=1.0 y la oración de salida (hipótesis) es usada para alinear la señal de voz con la alineación de Viterbi para obtener el segmento de estado [39]. Basados en el segmento de estado, se calcula la puntuación de correspondencia con todos los posibles factores, y el que tenga la mejor puntuación es seleccionado como el factor de deformación para esa pronunciación de entrada. Con el mejor factor de deformación, la pronunciación de entrada es descifrada otra vez para obtener la hipótesis final. Este es el procedimiento de descifrado:

1. Colocar el factor inicial de deformación α=1.0
2. Descifrar la pronunciación de entrada $O_\alpha(t)$:

$$\hat{W} = \arg\max_{w} P(W \mid O_\alpha(t), \Lambda) \tag{3.9}$$

3. Hacer una alineación de Viterbi con la hipótesis $\hat{W}$ para obtener el mejor segmento de estado.

$$S_t^* = \arg\max_{S_t} P(O_\alpha(t), s_t \mid \Lambda, \hat{W}) \tag{3.10}$$

4. Encontrar el mejor factor de deformación basado en el segmento S_t^*.

$$\alpha^* = \arg\max_{\alpha} P(O_\alpha(t) \mid s_t^*, \Lambda) \tag{3.11}$$

5. Descifrar otra vez basados en el mejor factor de deformación α^*.

$$\widetilde{W} = \arg\max_{W} P(W \mid O_{\alpha^*}(t), \Lambda) \tag{3.12}$$

Donde $O_\alpha(t)$, s_t, y Λ tienen el mismo significado que en el procedimiento de entrenamiento. Ŵ es la hipótesis de la primera iteración del descifrado sin VTLN. $\widetilde{W}$ es la hipótesis con VTLN. Como en el procedimiento de entrenamiento, buscamos el mejor factor de deformación basados en el segmento fijo de estado S_t^* con el cual aumentamos dramáticamente la velocidad del descifrado sin pérdida significante de precisión en el reconocimiento.

3.3 Regresión Lineal de Probabilidad Máxima (MLLR)

MLLR fue originalmente desarrollada para adaptación [18,40,41,42,43,44] pero puede ser también aplicada a situaciones donde haya diferencias en el entorno o ambiente. Un conjunto de matrices de transformaciones es estimado para los parámetros de los HMM Gausianos lo cual maximiza la probabilidad de los datos de adaptación. El conjunto de transformaciones es relativamente pequeño comparado con el número total de Gausianos en el sistema y con esto un número de Gausianos comparte las mismas matrices de transformación. Esto significa que los parámetros de transición pueden ser estimados de manera robusta con solamente una cantidad limitada de datos, lo cual permite actualizar a los Gausianos en el conjunto del HMM. Para una cantidad pequeña de datos (o estimación de la transformación muy robusta) sólo se usa una simple transformación. Entre más datos

estén disponibles se pueden estimar más transformaciones específicas. Originalmente las transformaciones fueron estimadas sólo para los parámetros de la media pero recientemente el enfoque ha sido extendido para que las varianzas Gausianas puedan ser actualizadas también. En la siguiente sección se dará un breve resumen de la teoría básica de MLLR para los parámetros de la media y las varianzas.

Las medias y las varianzas son adaptadas en dos etapas separadas. Inicialmente se encuentran nuevas medias. Entonces, dadas estas nuevas medias, actualizamos las varianzas. Por lo tanto, se modifican los HMMs tal que

$$L(O_T \mid \tilde{M}) \geq L(O_T \mid \hat{M}) \geq L(O_T \mid M)$$

donde M es el conjunto de modelos originales, el conjunto de modelos $\hat{M}$ tiene sólo actualizado los parámetros de la media (a $\hat{\mu}_1,...,\hat{\mu}_M$) y el modelo colocado $\tilde{M}$ tiene ambos las medias y las varianzas $\hat{\Sigma}_1,...,\hat{\Sigma}_M$ actualizadas, y O_T son los datos adaptados,

$$O_T = \{o(1), o(2),..., o(T)\}$$

3.3.1 Adaptación MLLR de las Medias

El objetivo de MLLR es obtener un conjunto de matrices de transformación que maximice la probabilidad de los datos de adaptación [18]. Se usa una matriz de transformación para dar un nuevo estimado de la media, donde

$$\hat{\mu}_m = \hat{W}_m \xi_m \tag{3.13}$$

y $\hat{W}_m$ es la matriz de transformación n x (n+1) (para n datos dimensionales) y ξ_m es el vector extendido de medias

$$\xi_m = [1 \ \ \mu_1, ..., \mu_n]^T \tag{3.14}$$

Para asegurar una estimación de los parámetros de transformación, las matrices de transformación se conectan a través de un número de Gausianos, de acuerdo al árbol de clases de regresión [45,46]. Este árbol contiene a todos los Gausianos en el sistema, con las estadísticas reunidas en las hojas (la cual puede contener un número de Gausianos). La transformación más específica que puede ser estimada robustamente usando la adaptación se genera para todos los Gausianos en el sistema.

Una transformación particular $\hat{W}_m$ se relaciona a través de R Gausianos $\{m_1, ..., m_R\}$. Para la salida Gausiana de la función de probabilidad de densidad considerada, $\hat{W}_m$ puede ser encontrada resolviendo

$$\sum_{T=1}^{T} \sum_{r=1}^{R} L_r(\tau) \Sigma_r^{-1} o(\tau) \xi_r^T = \sum_{T=1}^{T} \sum_{r=1}^{R} L_r(\tau) \Sigma_r^{-1} \hat{W}_m \xi_r \xi_\tau^T \tag{3.15}$$

donde

$$L_r(\tau) = p(q_r(\tau) \mid M, O_T)$$

y $q_r(\tau)$ indica el Gausiano m_r en el tiempo τ. Para el caso de la matriz completa de covarianzas es computacionalmente tratable y descrito en [40]. Cada transformación puede ser una matriz completa o restringida que está en bloque diagonal.

3.3.2 Adaptación MLLR de las Varianzas

Los vectores de varianzas de los Gausianos, o en general las matrices de covarianza, son actualizadas usando la siguiente transformación

$$\hat{\Sigma}_m = B_m^T \hat{H}_m B_m \tag{3.16}$$

donde $\hat{H}_m$ es la transformación lineal que es estimada y B_m es el inverso del factor de Choleski de Σ_m^{-1}, así que

$$\Sigma_m^{-1} = C_m C_m^T \tag{3.17}$$

y $B_m = C_m^{-1}$

En una forma similar para las medias, una transformación de varianzas es repartida sobre un número de Gausianos, $\{m_1,...,m_R\}$. Podemos mostrar que el estimado de la probabilidad máxima está dada por

$$\hat{H}_m = \frac{\sum_{r=1}^{R} C_r^T [\sum_{r=1}^{T} L_r(\tau)(o(\tau) - \hat{\mu}_r)(o(\tau) - \hat{\mu}_r)] C_r}{\sum_{r=1}^{R} \sum_{r=1}^{T} L_r(\tau)} \tag{3.18}$$

donde $\hat{\mu}_r$ es la media previamente calculada. Se puede ver que la matriz de transformación de varianzas estará completa, produciendo matrices completas de covarianzas para cada Gausiano. Se puede obtener una transformación diagonal para las varianzas simplemente haciendo cero los términos de la diagonal cancelada. Esto todavía garantiza el incremento de la probabilidad [18].

3.3.3 MLLR Basado en Lattices

Un problema con el resultado de la confidencia basado en MLLR es que una cantidad de datos de adaptación razonable puede necesitar ser descartada lo cual limita la precisión del estimado de las matrices de transformación. Como una alternativa, se ha desarrollado un método para usar directamente una representación de un *lattice* de cada pronunciación el cual es recorrido para proveer las estadísticas necesarias para la adaptación MLLR. Esto, en principio, significa que ningún dato necesita ser descartado sino que es incluido para que cada una de las estructuras de una contribución de peso a las estadísticas reunidas para muchos estados de un HMM.

El método MLLR estándar usa un paso forward-backward a través de una secuencia de un modelo HMM, cuando se calcula la probabilidad posterior de cada Gausiano en cada estructura y acumulando las estadísticas necesarias para MLLR. La idea detrás de la adaptación MLLR basada en lattices es que el paso forward-backward es hecho a través del lattice de reconocimiento de las rutas alternativas. Además la probabilidad posterior de un

estado en un tiempo en particular incluirá contribuciones de peso de todas las instancias de palabras relevantes que estuvieron en el lattice en ese momento [47].

3.4 Regresión Lineal de Probabilidad Máxima a Posteriori (MAPLR)

En algunas situaciones donde faltan datos de adaptación pueden conducirnos a un desempeño pobre para el algoritmo de adaptación MLLR. Para grandes cantidades de datos de adaptación, el uso de clases de regresión fijas pueden no proveer un desempeño óptimo [16].

El escenario que presenta la adaptación MLLR es bien conocida por sus pobres propiedades asintóticas y el aumento del desempeño se satura rápidamente cuando la cantidad de datos de adaptación se incrementa [16]. Esto pasa porque una transformación afina sin restricciones no es propiamente estructurada y nos puede llevar a soluciones inaceptables afectando la estructura fundamental del espacio acústico. Una posible solución a este problema es introducir algunas restricciones directamente sobre los parámetros de transformación, bajo la forma $g(n)=0$, pero este enfoque depende mucho sobre la elección ad-hoc de la función $g(.)$ y además debería ser evitada. Pero existe una solución alternativa que consiste en introducir conocimiento adicional sobre los valores que pueden tomar los parámetros de la transformación [32,48,49].

Esta restricción puede ser formalmente insertada en el proceso de estimación usando un criterio de probabilidad máxima *a posteriori* (MAP) para derivar n:

$$\hat{n} = \arg\max_{n} p(n \mid Y, \Lambda) \tag{19}$$

$$= \arg\max_{n} p(Y \mid n, \Lambda)\, p(n),$$

donde p(n) es la distribución *a priori* de los parámetros n.

A continuación se da una breve explicación de la transformación del modelo.

Sea $\Lambda = \{w_{n,m}, \mu_{n,m}, R_{n,m}\}$ un conjunto de mezclas de Gausianos de HMM donde $w_{n,m}, \mu_{n,m}$ y $R_{n,m}$ son el peso de la mezcla, el vector de medias y la matriz de precisión [7]del m-ésimo componente de la mezcla en el estado n, con $\mu_{n,m} \in \Re^{p}$ y $R_{n,m} \in \Re^{pxp}$.

El vector de medias de una distribución Gausiana es adaptado a una transformación afina, definida por el conjunto de parámetros n = {A,b} donde A es una matriz de transformación *pxp* y b es un vector de translación. Se asume que diferentes vectores de medias pueden compartir la misma transformación, y denotar n_c la transformación asociada a un grupo c de parámetros de un HMM. Dado algunos datos de adaptación, $Y = \{y_t\}, y_t \in \Re^{p}$, el objetivo es derivar los parámetros de transformación n_c usando el criterio de estimación de probabilidad máxima *a posteriori* descrito en [1]. La derivación de los parámetros de transformación esta dado en [32].

La mayoría de las técnicas desarrolladas para el reconocimiento robusto de la voz han sido aplicadas en adultos. Se ha demostrado que los cambios en las características espectral y temporal de la señal de la voz a través de varios grupos con diferentes edades afectan el desempeño de un reconocedor. Los sistemas de reconocimiento de voz para adultos han tenido una clara degradación en su desempeño cuando son utilizados con niños

[7] La matriz de precisión está definida como la matriz inversa de la covarianza.

[12,41,50]. Esto se atribuye principalmente a las diferencias anatómicas y morfológicas en la geometría del tracto vocal, al menos control de los articuladores y a las habilidades menos refinadas para controlar aspectos como la prosodia [51]. Existen diferencias importantes en las características espectrales de las voces de los niños comparadas con la de los adultos.

Dados los diferentes algoritmos para mejorar el desempeño de un reconocedor, se supone que técnicas como VTLN, MLLR y MAPLR mejorarán el reconocimiento de voz de niños. Con los resultados obtenidos se podrá entender mejor qué tipo problemas se presentan en cada una de las etapas del reconocimiento de voz así como nuevas soluciones posibles a estos problemas.

En los siguientes capítulos se presentará la implementación de los experimentos y los resultados obtenidos en cada uno de ellos.