Onditas perceptualmente diseñadas para el reconocimiento automático del habla

Alejandro J. Dabin[1], Diego H. Milone[1]y Hugo L. Rufiner[2]

[1] Grupo de Investigación en Señales e Inteligencia Computacional
FICH, Universidad Nacional del Litoral
aledabin@argentina.com, d.milone@ieee.org
[2] Laboratorio de Cibernética
Facultad de Ingeniería, Universidad Nacional de Entre Ríos
lrufiner@bioingenieria.edu.ar

Resumen Se propone un sistema de reconocimiento automático del habla en el que se representa la señal de voz mediante paquetes de onditas. El diseño del árbol de filtros implicado en la transformación se ha orientado en términos de criterios perceptuales. Para ello se ha tenido en cuenta la forma en que el oído humano procesa las señales acústicas. Se han considerado varias alternativas para tratar los coeficientes generados por esta transformación. Para modelar los datos y efectuar el reconocimiento se emplean modelos ocultos de Markov. Los resultados de reconocimiento incluyen señales de habla limpia y ruidosa, para el idioma español y con varios hablantes.

1. Introducción

Actualmente, los sistemas de *reconocimiento automático del habla* (RAH) tienen altas tasas de reconocimiento para conjuntos del orden de 20000 palabras y varios hablantes. Estos resultados corresponden a casos donde la voz es grabada sin ruido ambiental y con sistemas de adquisición de audio de buena calidad [1]. A éstas se las considera como "condiciones de laboratorio" porque a partir de ellas se evita cualquier tipo de distorsión de la señal voz.

En la realidad la situación suele resultar muy distinta. Casi siempre existen interferencias de ruido ambiente producido por distintas fuentes: personas, máquinas, electrodomésticos, etc. En las situaciones más simples se puede suponer un esquema de ruido aditivo. Es posible considerar que la señal de audio $x[n]$, siendo n la variable independiente de tiempo discreto, está formada por la suma de la señal de voz $s[n]$ y el ruido $r[n]$. Sin embargo, en general esta relación no resulta lineal ni conocida. Ésto puede deberse, por ejemplo, al eco que se produce en una habitación y a otros fenómenos como el conocido efecto Lombard [2]. Estos factores afectan considerablemente el rendimiento de los sistemas de RAH, provocando que la tasa de reconocimiento caiga abruptamente a medida que aumenta el ruido.

Se dice que un sistema de RAH es *robusto* cuando la presencia de ruido no afecta significativamente su rendimiento. Los sistemas sensoriales que intervienen

sinc(*i*) Laboratory for Signals and Computational Intelligence (http://fich.unl.edu.ar/sinc)
A. Dabin, D. H. Milone & H. L. Rufiner; "Onditas perceptualmente diseñadas para el reconocimiento automático del habla"
Proc. 7th Argentine Symposium on Artificial Intelligence. pp. 249-260, 2005.

en el proceso de comunicación humana resultan mucho más robustos que sus contrapartes artificiales [1]. Un sistema que emule algunas de estas capacidades tiene muchas más aplicaciones prácticas porque existen más situaciones reales en las que puede funcionar. En este artículo, se presenta un sistema de RAH que utiliza paquetes de onditas y algunos criterios perceptuales sencillos como una alternativa para incrementar la robustez del sistema.

El artículo se organiza de la siguiente manera. En la sección 2 se hace una introducción a la problemática de los sistemas de RAH. Luego, en la sección 3 se comentan las características principales de la transformada paquetes de onditas. El modelado de la voz mediante *modelos ocultos de Markov* (MOM) se describe en la sección 4. Los materiales utilizados se detallan en la sección 5. En la sección 6 se exponen los resultados y discusión de los experimentos. Finalmente, en la sección 7 se presentan las conclusiones de este trabajo.

2. Análisis de la voz y reconocimiento del habla

El reconocimiento del habla es una tarea compleja por los siguientes motivos:

- La pronunciación de un fonema puede tener distintas duraciones, aún para un mismo locutor y dentro de una misma palabra.
- Los locutores tienen diferencias fisiológicas y socio-lingüísticas que afectan sus pronunciaciones.
- Cuando escuchamos creemos percibir claras separaciones entre las palabras y otras unidades del habla, pero en general no existen tales separaciones en la señal real.
- Se trabaja con una señal muy redundante: una señal de voz digitalizada asigna al menos 8000 datos por segundo, pero la cantidad de información relevante que transporta es mucho menor.
- Hay grupos de fonemas que tienen características muy similares, dificultando su clasificación.

Se conoce mucho sobre el funcionamiento interno del oído humano [3]. El mismo realiza una compresión en la escala de las frecuencias altas de las señales de audio, siguiendo la relación conocida como escala de mel para las frecuencias percibidas:

$$F_{mel}(f_{Hz}) = 2595 \log_{10}\left(1 + \frac{f_{Hz}}{700}\right) \quad (1)$$

Varias técnicas de representación de la voz utilizan esta propiedad. La más popular es la de los *coeficientes cepstrales en escala de mel* (CCEM) [4]. Los sistemas de RAH basados en CCEM logran muy buenas tasas de reconocimiento con habla limpia, pero éstas decaen significativamente con niveles crecientes de ruido en la señal.

En este trabajo, la mencionada escala perceptual es considerada para el diseño de la base de paquetes de onditas [5,6], junto con la incorporación de algunos mecanismos para la limpieza del ruido presente en los coeficientes.

La parametrización de la señal de voz se realiza por tramos, utilizando una ventana cuadrada y con solapamientos que permiten incrementar la resolución temporal que "ve" el MOM. A nivel estructural y dentro de los objetivos de este trabajo, es posible descomponer el mensaje hablado en fonemas, palabras y frases. Luego, los fonemas constituyen la unidad mínima a reconocer. La forma en que estos elementos se modelan se detalla en la sección 4.

3. Transformada ondita y paquete de onditas

La *transformada ondita* (TO) (en inglés *wavelets*) se encuentra dentro de los métodos de análisis de tiempo-frecuencia. Sea $\psi \in L^2(\mathbb{R})$ una función definida en cercanías del origen, con media cero y norma unitaria. Por medio de escalamientos y traslaciones se genera el átomo tiempo-frecuencia [7]:

$$\psi_{p,q}(t) = \frac{1}{\sqrt{|p|}}\psi\left(\frac{t-q}{p}\right) \tag{2}$$

donde p es el parámetro de escalamiento y q es el de traslación. La función ψ recibe el nombre de función madre, o simplemente ondita, porque a partir de ella se obtiene la base para la transformación.

Las onditas utilizadas poseen en general soporte compacto, es decir que toman valores no nulos en un rango finito. Esta propiedad es importante para que estén bien localizadas en el dominio temporal. Suele requerirse también una adecuada localización en el dominio frecuencial.

3.1. Transformada ondita discreta

Es posible pensar en la operación realizada por la TO a partir de una división de la señal $A_j(t)$ en dos: una de aproximación (baja frecuencia) $A_{j+1}(t)$, y otra de detalles (alta frecuencia) $D_{j+1}(t)$. La *transformada ondita discreta* (TOD) descompone recursivamente la señal aproximada hasta el nivel deseado. Para formalizar esta descomposición Mallat introdujo el concepto de *análisis multiresolución* (AMR) uniendo los conocimientos de distintas áreas en una teoría compacta [7]. Se define al AMR a partir de una secuencia de subespacios de aproximación $\{V_j\}_{j\in\mathbb{Z}}$ en $L^2(\mathbb{R})$ tal que se satisfacen los siguientes requisitos:

1. $V_j \subset V_{j+1}$
2. $x(t) \in V_j \Leftrightarrow x(2t) \in V_{j+1}$
3. $x(t) \in V_0 \Leftrightarrow x(t+l)_{l\in\mathbb{Z}} \in V_0$
4. $\cup_{j=-\infty}^{j=\infty} V_j$ es densa en $L^2(\mathbb{R})$
5. $\cap_{j=-\infty}^{j=\infty} V_j = \{0\}$
6. Existe una función $\phi \in V_0$, llamada función de escala, tal que la familia $\phi(t-l)_{l\in\mathbb{Z}}$ es una base de Riesz para V_0.

Sea W_j un espacio complementario a V_j en V_{j+1}. Luego, el espacio W_j contiene la información de detalle requerida para ir desde una aproximación en la resolución j a una aproximación en la resolución $j+1$.

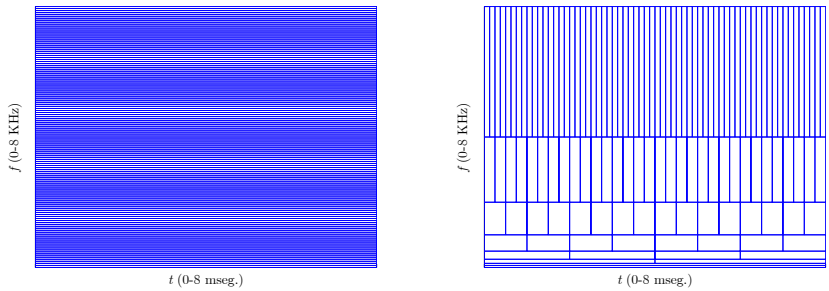

Figura 1. Comparación entre la resolución tiempo-frecuencia para un tramo de la transformada discreta de Fourier (izquierda) y de la TOD (derecha), calculada para tramos de 128 muestras y una frecuencia de muestreo de 16 kHz. Obsérvese la escasa resolución frecuencial relativa de la TOD en la zona entre los 500 y 3000 Hz, que constituye un rango de frecuencias muy importante para la discriminación de las vocales.

A pesar de sus atractivas características para el procesamiento de señales con componentes transitorias, la TOD posee importantes limitaciones para el análisis de la señal de voz [8]. La señal de voz posee una serie de comportamientos complejos que alternan entre trozos cuasi-estacionarios y otros fuertemente transitorios. Se requiere una transformación que tome en cuenta un adecuado compromiso entre ambos aspectos para rescatar las características significativas de la señal.

En la Fig. 1 se muestra una comparación entre la partición tiempo-frecuencia obtenida mediante la *transformada discreta de Fourier* (TDF) y la TOD para un tramo de una señal muestreada a 16 kHz. La TDF no posee ningún tipo de localización temporal (dentro del tramo considerado), pero posee muy buena resolución frecuencial. Por el otro lado la TOD resuelve adecuadamente eventos temporales a altas frecuencias, pero la resolución en frecuencia para el rango bajo y medio resulta insuficiente para discriminar diferencias entre importantes clases fonéticas. Por ello resulta importante plantear algunas alternativas que solucionen estos problemas y, a la vez, contemplen los aspectos positivos del enfoque basado en onditas.

3.2. Transformada paquete de onditas

La *transformada paquete de onditas* (TPO, en inglés *wavelet packet*) generaliza el AMR posibilitando diversas alternativas para el análisis tiempo-frecuencia realizado por la TOD [9]. La TPO no se limita a descomponer sólo la señal aproximada del nivel anterior sino que también puede descomponer la señal de detalle o alta frecuencia. Luego, la TOD constituye en realidad un subconjunto de la TPO.

Un árbol de descomposición específico tiene que ver con la forma en que se procesa la señal. Cada nodo del árbol tiene asignado una profundidad j y una

posición p dentro de ese nivel. Las dos bases ortogonales de TPO para el nodo (j, p) están definidas por [7]:

$$\psi_{j+1}^{2p}(k) = \sum_{n-\infty}^{\infty} h[n]\psi_j^p(k - 2^j n) \tag{3}$$

$$\psi_{j+1}^{2p+1}(k) = \sum_{n-\infty}^{\infty} g[n]\psi_j^p(k - 2^j n) \tag{4}$$

donde $h[n]$ es un filtro pasa-bajos y $g[n]$ es un filtro pasa-altos. Ambos forman un par de filtros espejo conjugados.

En este trabajo se utiliza el árbol de filtros diseñado para tener una resolución en frecuencia similar a la escala de mel. De esta manera, se obtiene una división frecuencial de la señal similar a la que realiza el oído. En contraste con el análisis tiempo frecuencia realizado por la TOD, con este árbol se busca contar con buena resolución para la TPO en el rango 500 Hz a 3000 Hz. Ésto se debe a que en este rango es donde mejor se aprecian las características de los fonemas sonoros.

3.3. Limpieza de ruido mediante onditas

Es posible utilizar las características particulares de los coeficientes derivados del análisis basado en onditas para implementar mecanismos de limpieza de ruido. Uno de los mecanismos más difundidos es la limpieza de ruido por umbralamiento de los coeficientes de la TOD [7].

Se puede decir en forma simplificada que este método consiste en: aplicar la TOD (o la TPO) al m-ésimo tramo de la señal original $x[m, n]$ y luego umbralar los coeficientes obtenidos $a[m, k]$ mediante una función adecuada $a_\theta[m, k] = f(a[m, k], \theta)$. En caso de que se quiera recuperar la señal limpiada en el tiempo, se aplica la transformada inversa a los coeficientes $a_\theta[m, k]$.

Existen varias maneras para calcular y aplicar los umbrales. La función de umbralamiento duro $f_H(a[m, k], \theta)$ involucra igualar a cero a todos los coeficientes cuyos valores absolutos están debajo de un umbral positivo, mientras que el resto permanecen inalterados. La función de umbralamiento suave $f_S(a[m, k], \theta)$ es similar, sólo que los coeficientes por debajo del umbral son "encogidos" hacia cero, mientras que el resto puede también modificarse mediante una función adecuada.

En este trabajo se utiliza la función de umbralamiento suave definida por:

$$f_S(a[m, k], \hat{\theta}) = \begin{cases} 0 & \text{si } |a[m, k]| \leq \hat{\theta}; \\ signo(a[m, k])(|a[m, k]| - \hat{\theta}) & \text{si } |a[m, k]| > \hat{\theta}. \end{cases} \tag{5}$$

El umbral $\hat{\theta}$ se calcula con el método de estimación imparcial del riesgo de Stein [10]. Dados varios valores de umbral posibles, se estima un riesgo asociado a cada uno. La minimización del riesgo en θ arroja el valor óptimo $\hat{\theta}$ a utilizar.

4. Modelos ocultos de Markov

Los MOM son una herramienta estadística que ha mostrado ser de mucha utilidad para el modelado del habla [4].

Un MOM está formado por un conjunto de estados. Entre dos estados i y j está definida una probabilidad de transición que permite pasar de un estado a otro. Existen dos tipos de estados: emisores y no emisores. En los primeros, cada estado m tiene asociado una distribución de probabilidad de observación $b_m(o_t)$, es decir, la probabilidad de generar la observación o_t en el instante t. Como su nombre lo indica, estos estados darán la salida del modelo ante una entrada determinada. El segundo grupo suele utilizarse simplemente para indicar el inicio y fin de cada modelo.

La distribución de probabilidades de observación utilizada es una mezcla de gaussianas, de allí es que se los denomina MOM continuos. Para comenzar se emplea el método de inicialización plana (*flat-start*). Durante el entrenamiento, los parámetros de cada modelo son reestimados con el algoritmo de Baum-Welch [4].

Cada fonema tiene asociado un MOM con 5 estados, de los cuales 3 son emisores. Mediante un diccionario fonético se definen todas las palabras válidas y los fonemas que las componen. Además, se calculan las probabilidades de paso de una palabra a otra, de acuerdo a las ocurrencias en la base de datos, a través de un modelo de bigramática estimado por *backing-off* [11].

Una vez parametrizadas todas las frases, se comienza el entrenamiento de los MOM. En todas las frases se incluyen dos símbolos de silencio, uno al principio y otro al final. De esta manera se modelan las posibles pausas en los extremos de cada grabación. Inicialmente, las probabilidades de cada MOM se reestiman tres veces. Se agrega una pausa corta entre las palabras, que sirve para modelar las separaciones entre estas, y se realizan dos reestimaciones más. Luego, se crean conjuntos de parámetros que comparten sus valores dentro de cada MOM. Con esto se obtiene una mejor estimación de los parámetros y se reduce la cantidad de cálculos. Finalmente, las probabilidades se reestiman 8 veces más.

La cantidad de reestimaciones de cada etapa se determina empíricamente hasta que la probabilidad de las observaciones dado el modelo no varíe significativamente. Una mayor cantidad de reestimaciones produce pocas mejoras en los resultados, aumenta el tiempo del cálculo y puede disminuir la capacidad de generalización del modelo.

Durante el reconocimiento se determina qué MOM más probablemente puede dar como salida la emisión desconocida. Para encontrar la secuencia más probable se emplea el algoritmo de Viterbi.

5. Materiales

La base de datos utilizada es parte del corpus de habla Albayzin [12]. Se tomó un subconjunto que consta de 600 elocuciones en español pronunciadas por 6 hablantes femeninos y 6 masculinos. Este subconjunto posee un total de

200 frases diferentes, con 202 palabras distintas, lo que resulta adecuado como punto de partida para obtener un reconocedor robusto ante varios usuarios y diversos ruidos.

A estos datos, originalmente grabados en condiciones de laboratorio, se le agregaron dos tipos de ruido a partir de la base NOISEX [13]: blanco y "murmullo". Este último está compuesto por la conversación de 100 personas en un bar, siendo las voces individuales apenas audibles. Ambos fueron aplicados en diferentes niveles de relación señal-ruido (SNR), con los valores que se detallan en la sección de resultados.

Las utilidades para trabajar con MOM pertenecen a las herramientas HTK[3]. Las rutinas para el cálculo de TPO son del paquete UviWave[4].

6. Resultados y discusión

Los resultados que se reportan consisten en las tasas de reconocimiento de palabras (RP), promediados sobre 10 particiones (validación cruzada por el método *leave-k-out* [14]). Cada partición se divide en un conjunto para entrenamiento (un 80 % del total de frases disponibles), y otro para prueba (el 20 % restante). Las frases de cada partición son elegidas al azar. En todos los experimentos presentados la ventana de análisis tiene 16 ms (128 muestras) con un solapamiento de 12 ms. Se realizaron pruebas con ventanas de 32 ms pero no resultaron en un mejor desempeño.

Con el fin de explorar las ventajas de cada alternativa y su impacto en el desempeño del sistema, se realizaron experimentos variando el tipo de ondita, el árbol de filtros, la integración de los coeficientes y el umbralamiento para la limpieza. Para poder hacer comparaciones en cuanto a los beneficios de cada representación de la señal, es necesario contar con estructuras comunes en el sistema de aprendizaje maquinal. Es por esto que los experimentos están separados en dos grandes grupos según la cantidad de dimensiones finales o coeficientes de los patrones que se modelan mediante los MOM. En el primer grupo se utilizan patrones de 23 dimensiones y en el segundo de 46. Las características y resultados obtenidos para cada uno se muestran en las Tablas 1 y 2, respectivamente.

En ambos casos se han incluido como referencia los mejores resultados logrados con la TOD. Para obtener 46 coeficientes a partir de las 128 muestras de señal temporal, se integraron las señales en cada escala según el siguiente esquema de reducción de coeficientes ($a[m,k] = \log_{10} \sum_{i \in \Gamma(j,k)} a^2[m,i]$, desde la escala j 1 a la 6): 64 → 16, 32 → 16, 16 → 4, 8 → 4, 4 → 4, 2 → 2. Para la referencia de 23 coeficientes se integró por grupos el doble coeficientes en cada escala.

La ondita utilizada en los experimentos 1, 2, 3, 5 y 6 fue la ondita spline biortogonal de orden 8. En el experimento 4 se empleó la ondita Daubechies de orden 16, manteniéndose el resto de los parámetros iguales a los del experimento

[3] `http://htk.eng.cam.ac.uk`

[4] `ftp://ftp.tsc.uvigo.es/pub/Uvi_Wave/`

Tabla 1. Características y resultados de los experimentos con patrones en $\mathbb{R}^{23}$.

Características		**Ref.**	**Exp.1**	**Exp.2**
Ondita		Spline 8	Spline 8	Spline 8
Árbol de filtros		TOD	Fig. 2	Fig. 3
Integración		directa	directa	TC
Umbralamiento		no	no	no

Resultados	SNR[dB]	RP[%]	RP[%]	RP[%]
Habla limpia	∞	71,1	**85,2**	83,7
Ruido blanco	50	65,2	**85,1**	83,9
	25	29,6	**66,5**	65,7
	15	0,0	43,1	**43,5**
	10	0,0	31,3	31,3
	5	0,0	0,0	0,0
Ruido murmullo	50	64,5	**85,2**	83,6
	25	62,2	**83,1**	81,9
	15	30,1	**65,2**	64,6
	10	0,0	**43,1**	42,4
	5	0,0	**28,7**	26,8

Tabla 2. Características y resultados de los experimentos con patrones en $\mathbb{R}^{46}$.

Características		**Ref.**	**Exp.3**	**Exp.4**	**Exp.5**	**Exp.6**
Ondita		Spline 8	Spline 8	Daub.16	Spline 8	Spline 8
Árbol de filtros		TOD	Fig. 2	Fig. 2	Fig. 2	Fig. 2
Integración		directa	TC	TC	AC	TC
Umbralamiento		no	no	no	no	suave

Resultados	SNR[dB]	RP[%]	RP[%]	RP[%]	RP[%]	RP[%]
Habla limpia	∞	57,0	**84,1**	83,5	59,4	72,8
Ruido blanco	50	56,3	**84,0**	83,4	59,2	72,8
	25	13,0	62,5	61,6	58,3	**63,3**
	15	0,0	32,2	31,4	**58,8**	34,3
	10	0,0	26,0	25,8	**51,4**	23,6
	5	0,0	0,0	0,0	**26,7**	22,3
Ruido murmullo	50	36,7	**84,1**	83,7	59,0	72,7
	25	22,2	80,0	**80,4**	59,0	68,9
	15	15,0	57,3	**57,8**	56,9	38,1
	10	0,0	40,0	40,2	**43,2**	22,7
	5	0,0	0,0	0,0	**28,0**	21,3

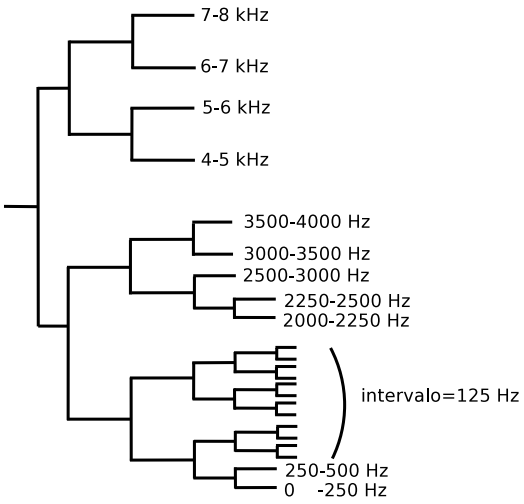

Figura 2. Árbol de filtros para los experimentos 1, 3, 4, 5 y 6. Si bien la idea es obtener resoluciones frecuenciales similares a las de la escala de mel, se tuvieron en cuenta consideraciones prácticas adicionales. Las características distintivas de muchos fonemas se aprecian en las frecuencias medias y bajas. Por lo tanto, el rango de frecuencias de 500 a 2000 Hz fue dividido en bandas pequeñas de 125 Hz. El rango de frecuencias entre 0 y 500 Hz no contiene tanta información importante para la discriminación fonética por lo cual las bandas son más amplias.

3. Entre todas las familias de onditas probadas estas dos brindaron los mejores resultados. El árbol de filtros utilizado en los experimentos 1, 3, 4, 5 y 6 se muestra en la Fig. 2.

Un problema que surge al utilizar directamente los coeficientes de la TOD (o TPO) es que la representación lograda no resulta invariante a la traslación. La ausencia de esta característica puede resultar un aspecto importante en el RAH u otras aplicaciones de aprendizaje maquinal. Esto se debe a que en los coeficientes se codifica la información relacionada con la fase relativa entre la posición del tramo de análisis considerado y las ondas características de la señal de voz. Esta información no tiene relación con la identidad o clase a la que pertenece de dicho tramo y por lo tanto puede perturbar la discriminación fonética.

Para evitar este efecto, en el experimento 1 los coeficientes transformados se integraron en cada escala j de forma de obtener un solo coeficiente por cada una de las 23 bandas frecuenciales ($a[m,j] = \log_{10} \sum_{i \in j} a^2[m,i]$). Este método ha sido utilizado en trabajos anteriores [5] y posee la ventaja adicional de reducir la cantidad de coeficientes a modelar mediante los MOM.

En el experimento 2, con un objetivo similar al anterior, se propone aplicar la *transformada coseno* (TC) a cada banda. Sin embargo, dado que no tendría sentido integrar nuevamente todos los coeficientes de la TC, se utilizó un árbol idéntico al anterior pero realizando una descomposición de una profundidad

sinc(*i*) Laboratory for Signals and Computational Intelligence (http://fich.unl.edu.ar/sinc)
A. Dabin, D. H. Milone & H. L. Rufiner; "Onditas perceptualmente diseñadas para el reconocimiento automático del habla"
Proc. 7th Argentine Symposium on Artificial Intelligence. pp. 249-260, 2005.

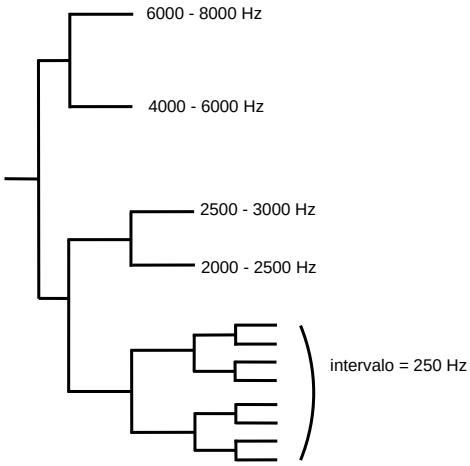

Figura 3. Árbol de filtros para el experimento 2. El esquema de distribución de las bandas de frecuencia sigue consideraciones similares que el árbol de filtros anterior. Sin embargo los filtros tienen un ancho de banda mayor en este caso.

menor (ver Fig. 3). Luego, se integraron las TC de estas bandas por mitades[5] y de esta forma se logró mantener una dimensión de entrada y resolución frecuencial equivalente a la del experimento 1.

Para los experimentos 3, 4 y 6 nuevamente se aplicó la TC a los coeficientes de cada banda, se los dividió en dos partes y se les calculó el logaritmo de su energía, siguiendo el mismo procedimiento antes descripto. Entonces, se obtuvieron dos coeficientes por cada banda, totalizando 46 por tramo de señal. De esta manera, y al igual que en los casos previos, se eliminó la información de la fase relativa del tramo analizado.

En otros trabajos se ha reportado el uso de la TC por sus capacidades de decorrelación bajo supuestos ideales, y por analogía con la forma en la que se obtienen los CCEM [6]. Esta propiedad resulta interesante porque los MOM podrían realizar una mejor estimación de las características de los fonemas. Sin embargo en los trabajos mencionados la TC se calcula sobre todos los coeficientes integrados y no sobre cada banda, lo que no permite aprovechar la mayoría de las propiedades de la TO.

En el experimento 5 se sustituyó la TC por la autocorrelación (AC) de las señales en cada banda. Dado el tramo de señal $x[m, n]$ con N muestras su secuencia de autocorrelación tiene $2N - 1$ coeficientes. Teniendo en cuenta que el resultado es simétrico, sólo se consideraron los primeros N elementos. Al igual que en los experimentos anteriores se dividió en mitades las nuevas secuencias y se tomó el logaritmo de su energía. Para la señal limpia y para SNRs altas los resultados son inferiores al resto. Sin embargo son equiparables con ruido

[5] Excepto la primer banda (0-250 Hz) que se integró entera.

murmullo cuando la SNR cae y significativamente mejores para SNR de 5 dB y para ruido blanco con SNR de 15, 10 y 5 dB.

Entre las pruebas que se realizaron aplicando la técnica de limpieza de ruido descripta anteriormente, se muestra el mejor resultado obtenido en el experimento 6, que provee alguna mejora sólo con cantidades moderadas de ruido.

7. Conclusión y trabajos futuros

Las tasas de reconocimiento disminuyen a medida que aumenta el nivel de ruido, lo cual es un comportamiento normal en todos los sistemas de RAH. Se puede observar que, en términos generales, el ruido tipo murmullo degrada mucho menos el desempeño del sistema que el ruido blanco. Resulta claro que las alternativas propuestas superan ampliamente el desempeño de la TOD.

En el experimento 1 se obtuvieron las mejores tasas de reconocimiento, excepto para los casos con bajas SNR de ruido blanco. Para habla limpia y señales con ruido de baja intensidad, el sistema reconoce en torno al 85 % de las palabras.

Con las familias de onditas Daubechies y spline biortogonal se obtuvieron resultados similares, aunque levemente mejores en el segundo caso.

Las representaciones obtenidas con cualquier TO en un análisis por tramos no son invariantes a la traslación (propiedad básica de toda TO). Por lo tanto, si se las utiliza directamente para alimentar un sistema de aprendizaje maquinal, se está incluyendo también información de la fase relativa entre paso del análisis por tramos y las ondas características de la señal. Para eliminar esta información que no tiene relación con la identidad del fonema que se quiere clasificar, se aplicaron –en cada escala separadamente– distintas transformaciones que no conservan la localización temporal. De esta forma se logró eliminar el efecto introducido por la fase relativa del análisis por tramos, sin perder las propiedades más importantes de la TO.

Si bien los resultados del experimento 6, donde se utiliza limpieza de ruido, no resultan mejores cuando la relación señal-ruido es alta, se aprecia un buen desempeño para la señal sucia a 5 dB. El método de eliminación de ruido utilizado también distorsiona las características de ciertos fonemas, lo que reduce las tasas de reconocimiento cuando el ruido es leve. En trabajos futuros se continuará la búsqueda de criterios y funciones de umbralamiento que eliminen el ruido a la vez que preserven las pistas acústicas significativas de la señal de voz, en combinación con las trasformaciones por escala para eliminar la información de fase relativa introducida mediante el análisis por tramos.

Referencias

1. Lippmann, R.P.: Speech recognition by machines and humans. Speech Communication **22** (1997) 1–15
2. Lombard, E.: Le signe de l'elevation de la voix. Annales Maladies Oreilles, Larynx, Nez, Pharynx **37** (1911) 101–119

3. Deller, J.R., Proakis, J.G., Hansen, J.H.: Discrete-Time Processing of Speech Signals. Macmillan Publishing, NewYork (1993)
4. Rabiner, L.R., Juang, B.H.: Fundamentals of Speech Recognition. Prentice-Hall (1993)
5. Torres, H., Rufiner, H.L.: Clasificación de fonemas mediante paquetes de onditas orientadas perceptualmente. In: Anales del Ier Congreso Latinoamericano de Ingeniería Biomédica, Mazatlán 98. Volume 1., México (1998) 163–166
6. Farooq, O., Datta, S.: Mel filter-like admissible wavelet packet structure for speech recognition. IEEE Signal Processing Letters **8** (2001) 196–198
7. Mallat, S.: A Wavelet Tour of Signal Processing. Second edn. Academic Press (1999)
8. Rufiner, H.: Comparación entre análisis onditas y Fourier aplicados al reconocimiento automático del habla. Master's thesis, Universidad Autónoma Metropolitana (1996)
9. Daubechies, I.: Ten lectures on wavelets. Philadelphia: SIAM (1992)
10. Donoho, D., Johnstone, I.: Adapting to unknown smoothness by wavelet shrinkage. J. American Statist. Assoc. **90** (1995) 1200–1224
11. Jelinek, F.: Statistical Methods for Speech Recognition. MIT Press, Cambrige, Masachussets (1999)
12. Moreno, A., Poch, D., Bonafonte, A., E.Lleida, J.Llisterri, J.B.Marino, Nadeu, C.: Albayzin speech data base: design of the phonetic corpus. In: Proceedings of the 2th European Conference of Speech Communication and Technology, Berlin (1993) 175–178
13. Varga, A., Steeneken, H.: Assessment for automatic speech recognition II NOISEX-92: A database and experiment to study the effect of additive noise on speech recognition systems. Speech Communication **12** (1993) 247–251
14. Guerin-Dugue, A., et al.: Elena II: Enhanced learning for evolutive neural architecture. Technical Report Number 6891, Research Program Report (1995)