TECNIACUSTICA
BARCELONA • 1996

Aplicación de las representaciones tiempo-frecuencia de clase Cohen al reconocimiento automático del habla

[2] Juan Carlos Llorente Bujosa y [1,2] Juan Luis Navarro Mesa

[1] Escola Universitària Politècnica de Mataró. Avda. Puig i Cadafalch, 101. 08303 Mataró, Barcelona
[2] Departament de Teoria del Senyal i Comunicacions. Universitat Politècnica de Catalunya (UPC)
C/ Gran Capitán. Campus Nord, UPC. Módulo-D5. 08034 Barcelona. email: navarro@gps.tsc.upc.es.

RESUMEN

La mayoría de métodos de parametrización para Reconocimiento Automático del Habla (RAH) funcionan bajo el presupuesto de que la señal se puede considerar casi estacionaria. Sin embargo, es bien sabido que hay numerosas situaciones en las que este presupuesto falla. También se trabaja con la idea de que el método de parametrización cumplir una serie de propiedades deseables. En los últimos años han aparecido unas representaciones adecuadas para señales no es estacionarias y que en buena medida cumplen ciertas propiedades de interés. En este artículo estudiamos la aplicación al RAH de dichas representaciones donde el espectrograma es la referencia como método de obtención de la información tiempo-frecuencia.

REPRESENTACIONES CUADRÁTICAS APLICADAS A VOZ.

A la hora de analizar una señal de voz normalmente se asume que sus características evolucionan lentamente con el tiempo permitiendo suponer una estadística estacionaria. Así, podemos hacer el análisis tomando la señal en tramas cortas, típicamente entre 10 y 60 ms. Esto nos permite utilizar una serie de técnicas para señales estacionarias, p.e., codificación LPC, análisis cepstral, etc. Sin embargo, la voz es un proceso complejo en el que es incorrecto asumir no estacionariedad dado que su dinámica induce cambios rápidos como transiciones sonoro/sordo, etc. que llevan asociada mucha información. En consecuencia un método basado en el supuesto de no estacionariedad es incapaz de describir correctamente tales situaciones. Un enfoque exclusivo desde el punto de vista estadístico o de evolución temporal da una visión incompleta de la voz misma ya que su contenido espectral es muy rico. Este contenido evoluciona con el tiempo incluso cuando se puede asumir estacionariedad. Por tanto, cualquier técnica de análisis de voz debe estar basada en Representaciones Tiempo-Frecuencia (RTF) específicamente diseñadas para procesos no estacionarios.

Muchas de las técnicas de extracción de información de señales como la voz están basadas en operadores cuadráticos. Esto es formal e intuitivamente razonable si tenemos en cuenta que las distribuciones de energía (o potencia) son magnitudes cudráticas que ofrecen mucha información y permiten hacer una interpretación física de la información. Además, su utilidad ha sido ampliamente probada en multitud de aplicaciones. Tal punto de vista ha dado lugar en los últimos años a una serie de RTF, llamadas de clase Cohen, que por su variedad y las propiedades que permiten conseguir se muestran muy sugerentes de cara al reconocimiento.

REPRESENTACIONES DE CLASE COHEN Y APLICACIÓN A VOZ.

Las Representaciones Tiempo-Frecuencia de clase Cohen [1] $C_x(t,f)$ son de tipo cuadrático. Hay varias formas de expresarlas matemáticamente. A nivel práctico, una de las más interesantes es mediante la autocorrelación instantánea $R_x(t,\tau) = x(t + \tau / 2)x^*(t - \tau / 2)$ porque permite cálculos relativamente rápidos. A partir de $R_x(t,\tau)$ se obtiene la autocorrelación generalizada de una señal compleja x(t) como,

$$R_x(t,\tau) = 1 / 2\pi \int R_x(t,\tau)\phi(t-u,\tau)du \quad (1)$$

donde $\phi(t,\tau)$ es el núcleo de la representación en el dominio tiempo-retardo. Aplicando la TF sobre $R_x(t,\tau)$ la expresión de $C_x(t,f)$ queda,

$$C_x(t,f) = \int R_x(t,\tau)e^{-i2\pi f\tau}d\tau \quad (2)$$

Dada la gran variedad de posibles núcleos $\phi(t,\tau)$ hay una gama amplia de representaciones. Las propiedades de las mismas parten del núcleo. De cara a su aplicación a reconocimiento de voz, las propiedades que podríamos desear en una RTF son: 1) la representación es real; 2) la RTF está limitada al mismo intervalo temporal que una señal limitada en tiempo (soporte temporal) y que esté limitada al mismo intervalo frecuencial de una señal de banda limitada (soporte frecuencial); 3) reduce o elimina los términos cruzados inherentes a toda representación cuadrática; 4) no es negativa, 5) tiene buena resolución temporal y frecuencial, etc., además de otras. En la tabla 1 presentamos las RTF que utilizamos en nuestros experimentos y el grado de cumplimiento de las propiedades mencionadas (todas son reales).

Tabla 1: relación de RTF y el grado de cumplimiento de varias propiedades.

RTF	No Negatividad	Resol. Tem./Frec.	Soporte Tem./Frec.	Términos Cruzados
Espectrograma	Si	BA/BE[1]	BA/BE[1]	Bien
Wigner	No	Bien/Muy Bien	Débil / Débil	Mal
Stankovic[2] [2]	No	Bien/Muy Bien	Débil / Débil	Muy Bien
Choi-Williams	No	Regular	Variable[3]	Variable[3]
Núcleo Cónico	No	Bien/Bien	Débil / Fuerte	Bien

[1] Compromiso fuerte entre espectrograma de banda ancha (BA) y de banda estrecha (BE).

[2] En esta representación [2] hay un parámetro L (=1,2,3) del que dependen sus propiedades. Se puede ver como un Wigner suavizado.

[3] Depende de α [1]. Si es α <1 la supresión de términos cruzados es buena, pero el soporte es malo. α =10 en las pruebas

Ninguna RTF tiene un comportamiento *bonito* para todos los propósitos. Cada una tiene sus propias ventajas y desventajas y es apropiada para ciertos tipos de señales y aplicaciones. Una de las propiedades que merecen mayor atención es la de no negatividad. En tratamiento de señal, una RTF debe ser (semi)definida positiva si queremos que se la pueda interpretar como una distribución de potencia pues los valores negativos no tienen significado físico. Desde este punto de vista el espectrograma es la única RTF de clase Cohen definida positiva y, por tanto, es una estimación plausible de la distribución de potencia de una señal. Esto no quiere decir que una RTF no negativa deba ser despreciada sin más. Si nuestro interés es extraer información relevante podríamos estar dispuestos a sacrificar la positividad deseada si a cambio se cumplen otras propiedades. Por ejemplo, es deseable un buen seguimiento de la evolución de la señal. En consecuencia, una buena resolución tiempo-frecuencia y buenos soportes resultan de gran interés. Como se puede apreciar en la tabla las representaciones de Stankovic y Núcleo Cónico no son definidas positivas, sin embargo tienen buena resolución y buen soporte lo que las hace potencialmente aplicables a nuestros propósitos.

Otra propiedad importante es la de reducción de términos cruzados o espureos. El hecho de que entre dos componentes frecuenciales pueda aparecer algún espureo conlleva riesgos. El más grave de todos es que no podemos estar seguros de si la información que se obtiene (p.e., visual) está realmente presente o no. A modo de ejemplo, podemos comentar que la RTF de Wigner tiene un gran problema en este hecho y es el principal responsable de su escaso éxito en tratamiento de voz. Por esta razón es razonable estudiar la posibilidad de aplicar otras RTF que reduzcan los espureos a la adquisición de las características de la voz para un sistema de reconocimiento. Sin embargo, los términos cruzados llevan información, p.e., relaciones de fases entre las diferentes componentes de la señal. Este hecho sugiere la posibilidad de que podría resultarnos útil aprovechar esta información o, por lo menos, no despreciarla.

3. SISTEMA DE RECONOCIMIENTO Y BASE DE DATOS.

El esquema básico de reconocimiento que utilizamos en nuestros experimentos es el clásico utilizado en la mayoría de sistemas, p.e., [3] capítulo 3. En él podemos distinguir tres módulos: 1) parametrización, 2) comparación entre las plantillas de referencia y de prueba y 3) regla de decisión a cuya salida se da la unidad fonética reconocida. Nuestro trabajo principal se centra dentro del módulo de parametrización.

Dado que la idea de nuestro trabajo es estudiar nuevas RTF y hacer una comparación con el espectrograma, lo primero que debemos hacer es decidir el método de parametrización. Los métodos de parametrización de más éxito en cuanto a tasa de reconocimiento para el RAH son el banco de filtros, la codificación LPC y el análisis cepstral, bien conjuntamente o por separado. La codificación LPC queda descartado ya que la no positividad de la mayoría de representaciones dificulta una justificación teórica de su uso que está fuera de nuestros objetivos. Entre el banco de filtros y análisis cepstral hemos tomado la opción de empezar por el segundo aplicado sobre $C_x(t,t)$ a intervalos regulares. Las razones pueden ser expresadas en tres puntos: permite llegar a tasas de reconocimiento aceptables, es un buen método de referencia pues ha sido utilizado ampliamente y las herramientas de reconocimiento de que disponemos en nuestro laboratorio (HTK), ade-

más de la base de datos (TI), permiten comparar con experimentos hechos con anterioridad. Como veremos en el apartado 4, también resulta interesante combinar con el banco de filtros, p.e., mel-cepstrum.

El pretratamiento hecho a la señal, previo a la parametrización, es el habitual en la mayoría de sistemas de reconocimiento. Primero, sobre la señal ya discreta, se hace un preénfasis (a = 0'95) y posteriormente un enventanado con ventana de Hamming. El tamaño de las ventanas es de 30 ms (240 muestras) y el desplazamiento entre ventanas sucesivas es de 10 ms (80 muestras). Sobre cada ventana de señal aplicamos las RTF y de ellas obtenemos los parámetros de la forma mencionada en el párrafo anterior.

El método de reconocimiento (módulo 2) es el basado en modelos ocultos de Markov continuos (CHMM) con diez estados por modelo y una gausiana por estado. La idea principal es que la única diferencia entre pruebas sea sólo la RTF aplicada, siendo todo lo demás común. Todas las pruebas se han hecho utilizando la herramienta HTK para reconocimiento automático del habla.

La elección de la base de datos para los experimentos se ha hecho en base a los siguientes criterios: hacer reconocimiento independiente de locutor, debe haber una riqueza fonética suficiente y el número total de realizaciones de las unidades fonéticas debe permitir extraer conclusiones. La base elegida es la de Texas Instruments (TI) donde las señales están muestreadas a 8 KHz. Contiene pronunciaciones de los dígitos aislados ingleses (one, ..., nine, zero, ou) de 112 hombres y 113 mujeres de 21 estados de EEUU con edades comprendidas entre 21 y 70 años para hombres y entre 17 y 59 para mujeres. El número de realizaciones totales utilizadas en la fase de entrenamiento es de 2464 y 2486 en la fase de prueba.

EXPERIMENTOS Y RESULTADOS.

El tratamiento hecho para cada RTF para la obtención de los coeficientes cepstrales es el siguiente,

$$x(t) -> C_x(t_n,k) -> \log(C_x(t_n,k)) -> IDFT\{\log(C_x(t_n,k))\} -> c_l$$

donde t_n es el instante en el que se extrae la información espectral (cada 10 ms), "k" indica frecuencia discreta y "l" es el índice del coeficiente cepstral.

Una de las primeras discusiones que se pueden tener es a cerca del número de coeficientes cepstrales que debemos tomar. En nuestras pruebas hemos tomado 12 coeficientes pues permiten unos buenos resultados de reconocimiento y es un valor generalmente aceptado. Otro aspecto importante es si utilizar sólo la información cepstral o también la delta. En todas las pruebas realizadas los mejores resultados se han obtenido utilizando ambas conjuntamente. De esta forma, los vectores de parámetros que se generan tienen 12 coeficientes cepstrales y 12 δ-cepstrales.

Un aspecto importante es qué es qué hacer con los términos negativos en aquellas RTF que los produzcan. Lógicamente, no se puede hacer un logaritmo sobre ellos. La opción que con más frecuencia se toma es ponerlos a cero. Esta idea está basada en que la mayoría de valores negativos se producen como consecuencia de los términos espureos. De esta forma conseguimos reducir su efecto aún más de lo que ya lo hace una RTF dada mostrando una imagen tiempo-frecuencia más *bonita* de la señal. Esta ha sido la primera opción probada. En la práctica (los resultados se muestran en la tabla 2) hemos dado a los valores negativos un valor positivo muy pequeño para evitar hacer el logaritmo de cero.

Tabla 2: tasas de reconocimiento con los coeficientes negativos puestos a cero

RTF	Espectrograma	Stankovic	Núcleo Cónico	Wigner	Choi-Williams
Error en %	11'50	18'6	14'16	4'10	13'90

En primer lugar cabe destacar un resultado sorprendente. La representación de Wigner, de la que se esperaban los peores resultados por la cantidad de espureos que genera, ofrece los mejores resultados. Esto nos induce a pensar, al menos en principio, que los términos cruzados no son causantes de una degradación excesiva en las tasas de reconocimiento. La causa de esto tiene dos aspectos. De un lado, el ya comentado, los términos cruzados quedan reducidos por la RTF y la puesta a cero de los términos negativos. Y en segundo lugar, si estos términos aparecen con regularidad para un fonema dado, representan información útil para el reconocimiento.

Como ya se ha comentado, los coeficientes negativos también llevan asociada información. Luego, ¿ por qué no utilizarlos ?. En tabla 3 se recogen los resultados de reconocimiento con los términos negativos tomados en valor absoluto ($IDFT\{\log(|C_x(t_n,k)|)\}$).

Tabla 3: tasas de reconocimiento con los coeficientes negativos tomados con valor absoluto

RTF	Espectrograma	Stankovic	Núcleo Cónico	Wigner	Choi-Williams
Error en %	11'50	6,44	5,91	7,77	7,30

Como podemos ver, salvo el espectrograma que no da valores negativos y Wigner, todas las RTF reducen las tasas de error. La justificación está en que, para muchas RTF, los términos negativos están presentes en buena parte de las muestras frecuenciales. Al ponerlos a cero perdemos esa parte de información. Si las frecuencias a las que aparecen fueran, digamos, al azar posiblemente no se habría obtenido los resultados de la tabla 3. Sin embargo, este no es el caso. Los términos negativos aparecen en las mismas zonas para un mismo fonema, de manera que permiten obtener información representativa de los mismos beneficiando el reconocimiento. Por tanto, en adelante las pruebas se hicieron tomando el valor absoluto.

Como vemos, las prestaciones de las RTF van mejorando pero no se han alcanzado unas tasas de reconocimiento equiparables a las mejores. Ya hemos comentado que no estamos restringidos al análisis cepstral. También hay la posibilidad de combinar con un banco de filtros, que es lo habitual por dar las mejores prestaciones en este tipo de sistemas. Esta es la idea que recoge el mel-cepstrum (MFCC +-MFCC en HTK) y de la que mostraremos resultados en la tabla 4, donde además se da el número de filtros con el mejor resultado en cada caso. El margen de filtros probado ha sido de 12 hasta 34.

Tabla 4: tasas de reconocimiento con valor absoluto y MFCC +-MFCC

RTF	Espectrograma	Stankovic	Núcleo Cónico	Wigner	Choi-Williams
Error en %	1,01	1,09	1,37	3,30	2,86
nº de filtros	28	22	20	24	24

En este caso las tasas de error bajan considerablemente en todos los casos y la relación de orden entre tasas de error se mantiene a igualdad de filtros. En el caso del espectrograma era de esperar. Sin embargo, en las demás RTF resulta contradictorio. Es decir, esperábamos que por tener mejor resolución frecuencial que el espectrograma (no podemos decir nada de la resolución temporal pues al tomar información cada 10 ms es muy difícil que sea un factor importante) ciertas representaciones, p.e., Núcleo Cónico, deben dar mejores resultados. Sólo la de Stankovic da valores muy aproximados. ¿ Por qué es así ?. La respuesta más plausible puede ser la siguiente. Hemos comprobado que con voz sorda la mejora en cuanto a resolución y soporte frecuencial es ciertamente mejor con RTF como Núcleo Cónico o Stankovic. Pero no es una mejora abrumadora y las siguientes etapas (filtrado Mel) del tratamiento deshacen esta mejora. En el caso de voz sonora la mejora frecuencial se traduce en un mejor seguimiento de los armónicos del pitch. En reconocimiento independiente de locutor la información de pitch empeora los resultados. De ahí que el banco de filtros Mel contribuya a mejorar las tasas con todas las RTF, pero nuevamente el banco de filtros deshace la (supuesta dependiendo de qué RTF aplicamos) mejora en cuanto a resolución frecuencial.

CONCLUSIONES

La Representaciones Tiempo-Frecuencia de clase Cohen abren una nueva vía para extraer información en un sistema de reconocimiento. Las pruebas hechas hasta ahora permiten albergar la posibilidad de que contribuyan a una mejora en las prestaciones de los sistemas de RAH. De momento comprobar que, con ciertas RTF (Núcleo Cónico, Stankovic, etc.) se consiguen tasas muy cercanas a las conseguidas con el espectrograma es alentador. Hay todavía mucho trabajo futuro por hacer. Entre otras podemos mencionar; confirmar que los espureos no son causantes principales de los (malos?) resultados, confirmar que aprovechar los términos negativos es beneficioso, aprovechar de forma explícita la resolución temporal, y aplicar a otros problemas de reconocimiento (p.e., reconocimiento dependiente de locutor) donde se puedan aprovechar las propiedades de resolución y soporte temporal. Otro aspecto a destacar es de una base de datos con un margen de confianza que permita sacar más conclusiones pues las tasas de error son muy bajas.

REFERENCIAS

[1] J. Jeong, W. J. Williams. "Kernel Design for Reduced Interference Distributions". IEEE Trans. on SP, Vol. 40, Nº 2, February 1992.

[2] L. Stankovic. "A Method for Time-frequency Analysis". IEEE Trans. on SP, Vol. 42, Nº 1, Feb. 1994.

[3] L. Rabiner, B. H. Juang. "Fundamentals of Speech Recognition". Prentice-Hall SP Series, 1993.