ESTADISTICAS DE ORDEN SUPERIOR EN RECONOCIMIENTO DE VOZ. ESTUDIO COMPARATIVO[1]

Sergio Tórtola, Asunción Moreno y Josep Vidal
Universidad Politécnica de Cataluña
Dep. Teoría de la Señal y Comunicaciones
Apdo 30002
08080 Barcelona
e_mail: amoreno@tsc.upc.es

***Abstract** - This paper studies the application of Higher Order Statistics (HOS) to speech recognition. Most part of the work developed in this fieldis based in autocorrelation and its Fourier transform (Power Spectral Density), even though its performance is considerably worse when noisy signals are considered. The main advantage is that HOS (cumulants) can eliminate Gaussian noise added to the signal. In this paper this property is applied to Linear Predictive (LP) estimation noisy speech. Two HOS based algorythms, the w-slice and the 1-D slice, are used for the first time in speech recognition, and compared with the well known Yule-Walker algorythm, based both in HOS and second order statistics.*

INTRODUCCION

El análisis mediante estimación LP de la señal de voz se ha centrado mayormente en métodos basados en las estadísticas de segundo orden (autocorrelación). Estas estadísticas describen perfectamente procesos gaussianos, pero pierden información cuando el proceso subyacente es no gaussiano o presenta no linearidades. Por el contrario, los cumulantes conservan la información de fase, información sobre no linearidades y sobre desviaciones de la estadística gaussiana. Estas propiedades son la causa de que haya habido un interés creciente en las aplicaciones de HOS al procesado de señal. La aplicación que nos concierne es el análisis LP de la señal de voz en entornos ruidosos, y se deriva de la propiedad por la cual los cumulantes de procesos con ruido coloreado (o no) aditivo pierden la información del ruido, por lo que el análisis da resultados no alterados por éste.

El análisis LP del habla modela el trato vocal mediante un filtro AR de fase mínima y variante con el tiempo, y la generación de la señal de voz mediante la excitación de dicho filtro por un ruido gaussiano blanco. El análisis se realiza dividiendo la señal en tramas donde se puede considerar estacionaria. En cada trama se estiman los coeficientes LP de la señal, y de ellos se derivan los coeficientes cepstrales, que son posteriormente enventanados, y constituyen, junto con la energía de la señal en la trama, los parámetros de partida del sistema de reconocimiento.

En la estimación de los parámetros AR del filtro con HOS se han empleado [4] las ecuaciones de Yule-Walker de orden 3. En el presente artículo se estudiará la ejecución en el reconocimiento del habla del algoritmo de Yule-Walker y de otros dos algoritmos ya empleados en codificación: w-slice [2] y 1-D slice [3]. Comprobaremos cómo puede afectar el método de resolución de los sistemas de ecuaciones correspondientes y el tipo de ventana cepstral, en la robustez de los distintos algoritmos frente al ruido.

ESTIMACION DE LOS COEFICIENTES LP MEDIANTE HOS

Se considera la señal de voz generada por un modelo AR(p) causal y estable, a la que se añade un ruido w(n)

$$y(n) = \sum_{i=1}^{p} a(i)y(n-i) + v(n)$$
$$z(n) = y(n) + w(n) \qquad (1)$$

El proceso de entrada v(n) es una secuencia independiente e idénticamente distribuida (i.i.d.), no gaussiana, de media nula y con cumulante k-ésimo $\gamma_{k,v}$ no nulo. El ruido aditivo w(n) es independiente de v(n), de

1. Este trabajo ha sido financiado por el gobierno Español TIC-92-0800-C05-04

media nula, gaussiano de espectro de potencia desconocido, o no gaussiano si $\gamma_{k,w}=0$. El filtro H(z) es exponencialmente estable. Por las propiedades de los cumulantes [1], las imposiciones sobre w(n) nos garantizan que $C_{k,w}=0$, y podemos trabajar independientemente con $C_{k,z}$ y $C_{k,y}$, ya que son iguales.

Consideraremos tres algoritmos de estimación de la secuencia AR: el algoritmo de Yule-Walker, el w-slice y el 1-D slice.

Algoritmo Yule-Walker de orden superior

La expresión de partida de las ecuaciones de Yule-Walker de orden superior [5] es

$$\sum_{l=0}^{p} a(l)\, C_{k,y}(m-l,k_0,0,\ldots,0) = \gamma_{k,v}\, h^{k-2}(-m)\, h(-m+k_0) \tag{2}$$

Por consiguiente, si consideramos únicamente el caso en que el término de la derecha se anula, se tiene

$$\begin{aligned} &\sum_{l=0}^{p} a(l)\, C_{k,y}(m-l,k_0,0,\ldots,0) = 0 \qquad si\ m>0\ ,\ cualquier\ k_0\ ;\ si\ m\le 0\ ,\ k_0<m \\ &\sum_{l=0}^{p} a(l)\, C_{k,y}(m-l,k_0,0,\ldots,0) = 0 \qquad si\ k_0>0\ ,\ m>0\ ;\ si\ k_0\le 0\ ,\ m>k_0 \end{aligned} \tag{3}$$

A fin de resolver las ecuaciones de Yule-Walker, hay que poder asegurar que el slice unidimensional $C_{k,y}(m,k_0,0,..,0)$, $k\ge 3$, m=1,...,p, tenga rango pleno, p. Se ha demostrado [5] que un slice unidimensional no garantiza rango pleno y es necesario concatenar p+1 slices, k_0=-p,...,0, en (3).

Al utilizar estimaciones de cumulantes, la resolución del sistema mediante LS o TLS, puede conducir a soluciones no estables. Consideraremos dos formas de mejorar la estabilidad: aumentar el número de slices (k_0=-p-M,...,0 y M>0), o aumentar el número de ecuaciones por slice, añadiendo ecuaciones de entre las permitidas por la ecuación (3) para los slices negativos. De la última solución consideraremos tres casos: (S1) el número de ecuaciones por slice es p (número mínimo), m=1,...,q; (S2) los slices negativos ($k_0<0$) incluyen p+1 ecuaciones, m=0,...,p; (S3) los slices negativos incluyen p-k_0 ecuaciones, m=1+k_0,...,p

Algoritmo w-slice

El algoritmo w-slice [2] se basa en la siguiente suma ponderada de slices de cumulantes

$$C_w(i) = w_2 C_{2,y}(i) + \sum_{j=-L}^{N} w_3(j)\, C_{3,y}(i,j) + \sum_{j=-L}^{N}\sum_{k=-L}^{N} w_{4,y}(j,k)\, C_{4,y}(i,j,k) + \cdots \tag{4}$$

y se desarrolla en tres pasos

P1. Escoger los pesos w_2, $w_3(j)$, $w_4(j,l)$,... de forma que se cumpla que

$$\begin{aligned} C_w(i) &= 0 \qquad , i=-P,\ldots,-1 \\ C_w(0) &= 1 \end{aligned} \tag{5}$$

donde P≥p. Se toma N≥0 y L≥p+M en (4), con la misma sobredeterminación (M) que la resolución de P3.

P2. La condición (4) permite estimar los P primeros términos de la respuesta impulsional a partir de (5), ya que $h(i)=C_w(i)$ para i=1,...,P.

P3. A partir de la respuesta impulsional estimada, h(i), se puede construir el sistema de ecuaciones

$$\sum_{l=0}^{p} a(l)\, h(i-l) = 0 \qquad , \ i=1,\ldots,P \tag{6}$$

y su resolución producirá la secuencia AR estimada.

La resolución de P1 se hace con LS. En la resolución de la ecuación matricial de P3 influye mucho la varianza de la h(i) calculada. Una opción es sobredeterminar la matriz, para que LS ó TLS den soluciones estables (P=p+M,M>0), o bien resolver mediante correlación (CORR) el sistema, lo que conduce a

$$\sum_{k=0}^{p} a(k) R_{hh}(k-l) = 0 \qquad , l=1,...,p \tag{7}$$

y se asegura estabilidad de la secuencia AR obtenida. En cualquier caso, para obtener coeficientes AR bien estimados, es suficiente considerar slices bidimensionales en (4).

Algoritmo 1-D slice

El algoritmo 1-D slice es una variante del algoritmo de Yule-Walker, que propone obtener los parámetros AR a partir, únicamente, de la información contenida en un slice, $C_{k,y}(m,k_0,0,...,0)$. Resultados anteriores demuestran que un slice unidimensional no tiene porqué tener rango pleno, por lo que es poco fiable resolver directamente la ecuación (3), ya que la solución obtenida puede no ser estable, e incluso puede no haber solución. Una forma de solventar este problema es emplear la correlación de los cumulantes, a fin de conseguir una matriz Toeplitz, simétrica y diagonal dominante.

La ecuación de partida es (2). Calculando la correlación con el slice unidimensional, se llega a

$$\sum_{l=0}^{p} a(l) R_c(l-l') = \gamma_{k,v} \sum_{m} h^{k-2}(-m) h(-m+k_0) C_{k,y}(m-l',k_0,0,...,0) \tag{8}$$

donde $R_c(i)$ es la correlación de los cumulantes. Como no se conoce la respuesta impulsional, proponemos despreciar el término a la derecha de la igualdad, y resolver el sistema de ecuaciones

$$\sum_{l=0}^{p} a(l) R_c(l-l') \equiv 0 \qquad l=0,...,p \tag{9}$$

lo cual dará una secuencia AR estimada diferente de la real. La ventaja es que este sistema garantiza estabilidad. Para reducir la varianza de la correlación se emplean 150 muestras del cumulante y $k_0=0$.

Consideraciones sobre la estabilidad

En el caso de que la resolución de las ecuaciones mediante LS ó TLS produzcan soluciones inestables hemos considerado tres posibles vías de actuación: (E1) eliminar las tramas con coeficientes LPC inestables; (E2) invertir los polos de fase máxima; (E3) tratar por igual tramas estables e inestables. En el apartado de resultados se utilizarán los tres métodos, y se escogerá el que de mayores tasas de reconocimiento, salvo si se obtienen resultados similares, en cuyo caso se escogerá E3 por su menor coste computacional.

RESULTADOS

La base de datos empleada es la de los diez dígitos catalanes. La base se ha creado a partir de diez locutores (tres mujeres y siete hombres), con diez repeticiones de cada dígito por locutor. Las señales se han filtrado a 3,4KHz y muestreado a 8KHz. Las señales con ruido se generan añadiendo ruido gaussiano blanco sintético a toda la señal limpia. Posteriormente se eliminan las muestras de silencio en la señal limpia. El análisis LPC (AR(8)) se realiza en tramas de 300 muestras (37.5ms) para algoritmos de orden superior, y de 240 muestras (30ms) para Yule-Walker de orden 2, con solapamiento del 50%. A continuación se calculan los coeficientes cepstrales, los coeficientes cepstrales diferenciales y la energía diferencial, y se utilizan observaciones en el sistema de reconocimiento basado en HMM [6]. A este fin, se toman las cinco primeras realizaciones de todos los locutores para entrenar los HMM asociados a cada dígito, y se reconocen las restantes.

Estas pruebas se realizan en primer término con señales limpias y con señales ruidosas con SNR=10dB, con todos los algoritmos y combinaciones posibles de variaciones consideradas de la parametrización. En función

		∞dB		20dB		10dB		0dB	
orden	M	seno	rampa	seno	rampa	seno	rampa	seno	ramp
YW3	0	99.2	99.6	95.6	96.2	67.6	76.8	21.2	25.6
	16	99.4	99.2	94.8	96.4	65.4	72.4	28.2	24.8
YW4	0	99.2	100	95.4	96.6	68.2	69.0	19.4	24.4
	16	99.4	99.0	96.4	96.4	64.8	79.4	19.0	25.4

Tabla 1. YW3 y YW4, LS, S2 y E3

		∞dB		20dB		10dB		0dB	
orden	método	seno	ramp	seno	ramp	seno	ramp	seno	ramp
WS3	LS	98.6	98.6	96.4	95.6	80.0	82.6	40.0	44.8
	CORR	99.2	99.0	94.6	96.0	81.2	82.8	29.8	48.8
WS4	LS	98.8	97.8	95.6	93.6	82.2	83.8	39.8	46.0
	CORR	98.4	99.4	96.6	97.2	85.0	88.0	40.0	51.6

Tabla 2. WS, con sobredeterminación

de las tasas de reconocimiento obtenidas, se prueban los algoritmos que mejor resultado han dado, junto con las mejores combinaciones consideradas, con señales de test con SNR=0db y SNR=20dB. Los resultados se compararán con los de Yule-Walker de segundo orden (YW2), utilizando ventana de rampa y sin preénfasis.

Yule-Walker de orden superior

Los resultados utilizando LS, con cumulantes de ordenes 3 (YW3) y 4 (YW4), no difieren apreciablemente en ausencia de ruido, pero a 10dB S2 mejora el reconocimiento frente a S1 y S3. Los mejores reconocimientos en todas las pruebas se logran sin considerar la estabilidad de las tramas (E3). Considerando sólo S2, la resolución del sistema mediante TLS produce parámetros LPC muy inestables, y la mejor opción es invertir los polos de fase máxima (E2), lo que aumenta el número de operaciones. Además los mejores resultados se obtienen mediante LS, de forma que de aquí en adelante éste será el método que se emplee.

Las diferencias entre tasas de reconocimiento con señales limpias no son significativas, pero la ventana de rampa mejora el reconocimiento cuando aumenta el ruido, por lo que será la ventana empleada en adelante.

En las parametrizaciones con cumulantes de orden 3, prescindiendo de los resultados a 0dB por su invalidez como tasas de reconocimiento, la sobredeterminación (M=16) parece no afectar en condiciones de bajo ruido, pero empeora el resultado cuando SNR=10dB, por lo que el sistema de tercer orden no se sobredterminará. En las parametrizaciones con cumulantes de orden 4, la sobredeterminación mejora el reconocimiento en condiciones ruidosas, siendo la mejora tanto mayor cuanto mayor lo es el ruido, prescindiendo de nuevo de los resultado a 0dB, de forma que con cuarto orden se tomará M=16.

Algoritmo w-slice

La inestabilidad de las tramas puede provenir del cálculo de los pesos y del cálculo de los coeficientes LPC (sólo LS y TLS), de forma que el número de tramas inestables es mayor para TLS y menor para correlación,

y mayor cuando no se emplea sobredeterminación (M=0). La resolución por TLS se descarta por dar peores resultados. Cuando se emplea correlación los mejores reconocimientos se consiguen eliminando tramas inestables (E1), y con LS la mayor estabilidad obliga a invertir los polos de fase máxima (E2). Aun así, hay

ventana	∞dB	20dB	10dB	0dB
seno	98.4	97.2	90.4	61.6
rampa	98.4	97.0	91.0	59.8

Tabla 3. 1-D slice de orden 3

método	∞dB	20dB	10dB	0dB
YW2	99.6	98.6	82.2	34.4
YW3	99.6	96.2	76.8	21.2
YW4	99.0	98.6	79.4	34.4
WS3	99.0	96.0	82.8	48.8
WS4	99.4	97.2	88.0	51.6
1D3	98.4	97.0	91.0	59.8

Tabla 4. Comparación con Yule-Walker de segundo orden

un número remanente de tramas inestables (entorno al 2%) debido a la inestabilidad en la obtención de los pesos, que hay que eliminar.

Cuando se utiliza LS, la ventana de seno alzado es mejor en condiciones de bajo ruido, pero a partir de 10dB la ventana de rampa mejora las tasas de reconocimiento. Por el contrario, cuando se emplea correlación, aun en condiciones de bajo ruido da mejores resultados la ventana de rampa. Por consiguiente, se empleará la ventana cepstral de rampa. Además el método de correlación mejora el reconocimiento en condiciones de ruido elevado respecto a LS.

Algoritmo 1-D slice

La ejecución del algoritmo es virtualmente independiente de la ventana cepstral empleada. Tomando como criterio elegir la que produzca una mayor tasa de reconocimiento a 10dB, elegimos la ventana de rampa.

CONCLUSIONES

Los mejores reconocimientos en presencia de ruido se obtienen con ventana cepstral de rampa, resolviendo las ecuaciones de Yule-Walker mediante LS, y las de w-slice mediante correlación. Por consiguiente, el mejor reconocimiento va asociado a una reducción de la varianza en la estimación de la secuencia AR. En concreto, el algoritmo 1-D slice supera a todos los demás en condiciones de ruido elevado, y no empeora excesivamente en ausencia de ruido, ya que, si bien su estimación es peor en media, es mucho mejor en varianza.

BIBLIOGRAFIA

1. J.M.Mendel, Tutorial on Higher Order Statistics (Spectra) in Signal Processing ans System Theory: Theoretical Results and Some Applications, IEEE Proceedings Vol.79 No.3, March 1991
2. J.Vidal and J.A.R.Fonollosa, Causal AR Modeling Using a Linear Combination of Cumulant Slices, Elsevier Science: Signal Processing 36, Aug. 1992
3. A.Moreno,J.A.R.Fonollosa and J.Vidal, HOS Analysis of Speech. A Vocoder Application, Proceedings of ICSP'93 pp.282-285
4. K.K.Paliwal and M.M.Sondhi, Recognition of Noisys Speech Using Cumulant-Based Linear Prediction Analysis, Proceedings of the ICASP'91, Toronto (Canada), May 1991
5 A.Swami and J.M.Mendel, AR Identifiability Using Cumulant Slices, Proceedings of the Workshop in Higher Order Spectral Analysis, CO pp.13-18, June 1989
6. L.R.Rabiner, A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, IEEE Proceedings Vol. 77 No.2, Febr. 1989