Caracterización del habla mediante métodos de análisis tiempo-frecuencia

PACS: 43.60.Gk

Cerdá S. Romero J.

Laboratorio de Acústica.Universidad de Valencia

Resumen

We apply new time-frequency techniques to observe speech features. We compare the new representations with conventional methods used. Rapidly time-variying formants and harmonic structure is revealed. Examples for synthetic and real speech are provided.

1 Introducción

La trasformada general de ajuste mínimo de ventana (GTAMV) [4], es un mecanismo de análisis de señales no estacionarias que se adapta a las características espectrales de la señal. Este nuevo método de análisis es una generalización inmediata de la transformada de ajuste mínimo de ventana (TAMV) [3]. La adaptación se realiza mediante el análisis del espectro global de la señal. A partir de dicho espectro, se determina dónde aparecen frecuencias bien definidas. El análisis se realiza utilizando diferentes tamaños en la ventana de localización temporal para cada frecuencia, segun sea la contribución que hay en la señal para dicha frecuencia. La transformada GTAMV permite seleccionar la resolución temporal (consecuentemente en frecuencia), en cada instante y para cada frecuencia analizada. En su forma automatizada, utiliza el principio de incertidumbre para señales. Para definirla utilizamos los conceptos que ya presentamos en [3]. Dado un intervalo $[\nu-\epsilon, \nu+\epsilon]$, definimos la **media local de frecuencias**

$$\mathcal{W}(\nu,\epsilon) = \frac{\int_{\nu-\epsilon}^{\nu+\epsilon} \omega\gamma(\omega)d\omega}{\int_{\nu-\epsilon}^{\nu+\epsilon} \gamma(\omega)d\omega} \tag{1.1}$$

La **dispersión local en frecuencias** será

$$\mathcal{S}^2(\nu,\epsilon) = \frac{\int_{\nu-\epsilon}^{\nu+\epsilon} (\omega-\mathcal{W}(\nu,\epsilon))^2\gamma(\omega)d\omega}{\int_{\nu-\epsilon}^{\nu+\epsilon} \gamma(\omega)d\omega} \tag{1.2}$$

Nuestro punto de partida es la dispersión local en frecuencias $\mathcal{S}(\omega)$ (1.2). Ahora vamos a interpretar que si para una frecuencia ω, tenemos una dispersión $\mathcal{S}(\omega)$, la señal analizada va a estar compuesta por un átomo tiempo-frecuencia, de frecuencia ω y dispersión temporal [1]

$$\sigma_t(\omega) = \frac{1}{\mathcal{S}(\omega)} \tag{1.3}$$

Hemos incluido la dependencia con ω para hacerla más explícita. De esta forma, podemos escojer el siguiente valor para el número de ciclos que deseamos entren en el tamaño de la ventana de análisis:

$$K(\omega) = \lambda\frac{\sigma_t(\omega)\omega}{\sqrt{2}\pi} \tag{1.4}$$

[1] El caso $\mathcal{S}(\omega)=0$, sólo puede ocurrir si $\gamma(\omega)=\delta(\omega-\nu_0)$. Este caso para señales físicas reales es imposible. De todos modos, nos proporciona el resultado bien conocido que una señal perfectamente localizada en frecuencias tiene que estar totalmente deslocalizada en el dominio temporal.

en donde hemos incluido una constante de proporcionalidad λ que nos permite controlar la resolución global del método. Se incluye esto pues para frecuencias no presentes en una señal, obtenemos valores de $\mathcal{S}(\omega)$ grandes que nos proporcionan valores de $\sigma_t(\omega)$ muy pequeños. Esto significa que el número de ciclos a observar puede ser muy pequeño lo que conlleva una resolución muy baja en frecuencias. El factor λ evita esta pérdida de resolución que puede ser contraproducente. Con esta elección las funciones de referencia son

$$\varphi_\omega(t) = e^{-\left(\frac{t}{\sqrt{2}\lambda\sigma_t(\omega)}\right)^2} \cos(\omega t + \pi/2) \tag{1.5}$$

En este método nos queda exclusivamente un parámetro libre. Al igual que el TAMV es un método comparativo adaptativo, necesitando de un preanálisis de Fourier. Como vemos el número de ciclos que analizamos depende de la frecuencia de análisis. Formulado de esta forma este método correspondería a una versión real del *GSTFT* que aparece en [2]. Por último, la expresión genérica del GTAMV viene dada por

$$\mathcal{G}f(t,\omega) = \frac{1}{||\varphi_\omega||} \int_{-\infty}^{+\infty} f(\tau)\varphi_\omega(\tau - t)d\tau \tag{1.6}$$

Para un estudio más detallado de la transformada GTAMV el lector puede dirijirse a [5].

2 Caracterización del "Habla"

La caracterización del habla ha sido posible gracias al desarrollo de los analizadores de habla y los sintetizadores de voz. El análisis tiempo-frecuencia permite extraer las características acústicas contenidas en la señal y presentarlas en el plano Tiempo-Frecuencia (plano TF), respecto a las dimensiones de frecuencia, intensidad y tiempo. De estas representaciones se puede extraer las propiedades genéricas que permiten una caracterización de la señal: dimensiones físicas, frecuencia, intensidad y tiempo [1]. Desde este punto de vista, las vocales aisladas se pueden caracterizar mediante la localización de los dos primeros formantes. El tercer formante permanece relativamente estable a lo largo de todas las vocales [1]. Sin embargo, parece ser que el oyente no debe tomar la información proporcionada por F1 y F2 en términos absolutos, sino de forma relativa. En general parece evidente la necesidad de un proceso de normalización en la percepción del lenguaje para poder ajustar las variaciones producidas por los diferentes hablantes. En las diferentes opciones que intentan resolver este problema, entran a formar parte la frecuencia fundametal F0, y los formantes más altos F3 y F4. Notese que estos resultados son válidos para vocales aisladas.

En un contexto consonántico, nos encontramos que el tiempo en que se producen las palabras suele ser muy corto, imposibilitanto el desarrollo de las estructuras resonantes de una forma bien definida. Sin embargo, parece ser que el reconocimiento es favorecido por dicho contexto. Esto implica la necesidad de incluir además de factores acústicos, factores fonológicos. En general se admite la importancia de la información dinámica espectral dada por las transiciones de los formantes , como de la información acerca de la duración en la identificación de la vocal [1]. Los problemas perceptivos en sí, se alejan del objetivo de este trabajo, el lector interesado puede encontrar más información en [1].

Desde la perspectiva del análisis tiempo-frecuencia, los parámetros que nos interesa determinar en la caracterización de las vocales, son:

1. *Duración.*
2. *Frecuencia fundamental.*
3. *Formantes* F1, F2, F3 y F4.

En cuanto a las consonantes existen también una serie de características generalmente admitidas [1]:

1. *Oclusivas*: Intervalo de silencio hasta el comienzo de la vocal (VOT, *voice onset time*) y la transición del segundo formante hacia la vocal siguiente.

2. *Fricatrivas*: presencia de ruido aperiódico en el espectro, baja intensidad y concentración de la energía entre los 3000 y 5000 Hz.

3. *Nasales*: presencia de energía en las frecuencias bajas del espectro. Un primer formante en torno a los 250 Hz, un segundo formante muy débil y el tercero situado en torno a los 2200 Hz.

4. *Líquidas*: Concentración de energía en bandas estables y de larga duración (semejantes a las vocales).

La distribución de energía en el espectro y su amplitud, la forma de las transiciones o conexiones de la energía entre consonantes y vocales, y la duración son los rasgos acústicos que aparecen como dominantes, en la representación espectrográfica de las consonantes [1].

3 Métodos tiempo-frecuencia y análisis del habla

Desde hace mucho tiempo, los métodos para la determinación de los formantes, hacen uso de las representaciones TF. Consecuentemente, los problemas inherentes a estos métodos se trasladan a la determinación de estas características. La representación más utilizada es el *espectrograma* (módulo del Short-Time Fourier Transform, STFT)[2]. Bien conocidas son las deficiencias de este método a causa del principio de incertidumbre para señales [5]. Otras representaciones TF son las denominadas *distribuciones bilineales*, de las que la distribución de Wigner-Ville (WVD), es el elemento central. La bilinealidad de estas representaciones produce términos de interferencia con la consecuente pérdida de definición. Para solucionar los problemas del STFT o de la WVD se han ideado diversas estrategias. El lector interesado puede encontrar información abundante en [5].

Para observar las cualidades o deficiencias del GTAMV, comparamos este método con el STFT de *banda ancha* y de *banda estrecha*, y con la distribución *pseudo-suavizada de Wigner* (SPWVD), variante de la WVD donde las interferencias están reducidas. Para nuestro propósito hemos escojido la frase *"la locura del niño explotado"*, pronunciada por un varon de 30 años de edad. La señal se ha grabado a 16 bits y con una frecuencia de muestreo de 11 KHz. En las siguientes Figuras 1-4, presentamos los resultados obtenidos mediante el STFT con una ventana de análisis de 64 muestras, el STFT con una ventana de análisis de 256 muestras, la distribución SPWVD y la GTAMV.

En la Figura 1 hemos incluido la localización de cada fonema. En esta misma Figura podemos apreciar que el STFT con banda ancha produce una imagen algo "embadurnada". En el STFT de banda estrecha se aprecian muy bien los armónicos presentes. Pero no está muy claro cómo determinar los formantes. La Figura 3, permite apreciar con mayor claridad los formantes. Nótese que en esta figura se aprecia también la periodicidad temporal. En la Figura 4, se aprecia también los formantes. Ahora ya no se aprecia bien la periodicidad temporal, sin embargo se puede apreciar algo mejor la estructura de armónicos que en el SPWVD. Desde este punto de vista se pueden considerar complementarias el SPWVD y la GTAMV, dando la primera prioridad a la estructura temporal mientras que la GTAMV da prioridad a la estructura en frecuencias.

[2]Por supuesto que existen métodos que trabajan directamente sobre la señal, bien en el dominio temporal, bien en el dominio de frecuencia, sin pasar a una representación TF

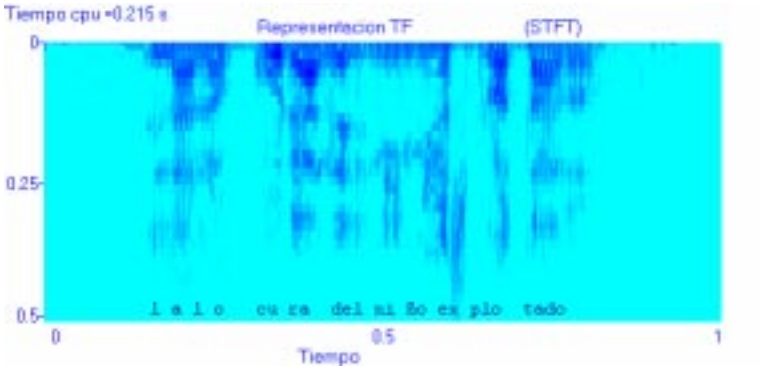

Figura 1: Espectrograma de banda ancha

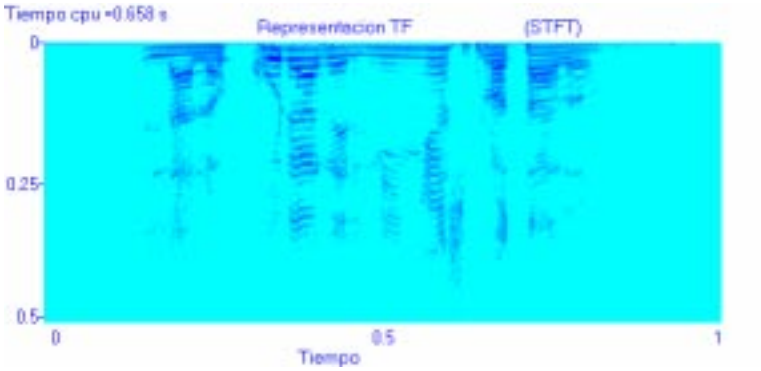

Figura 2: Espectrograma de banda estrecha

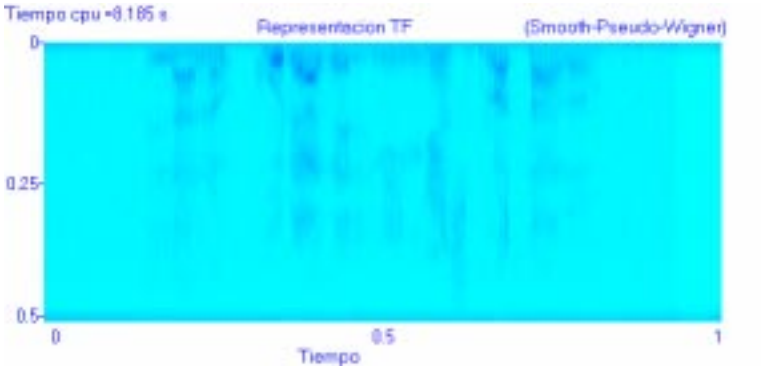

Figura 3: Distribución Smooth-Pseudo Wigner-Ville

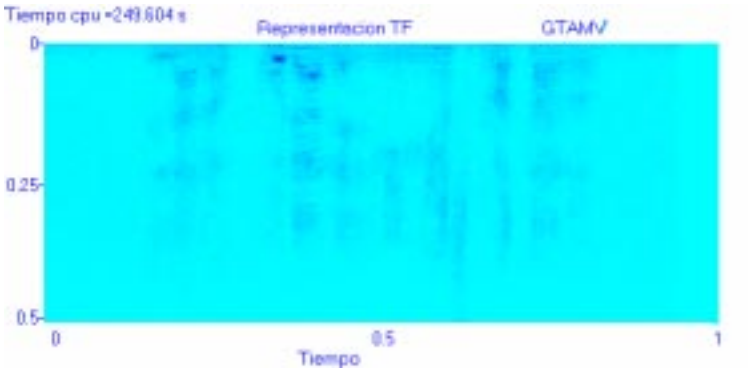

Figura 4: Transformada GTAMV

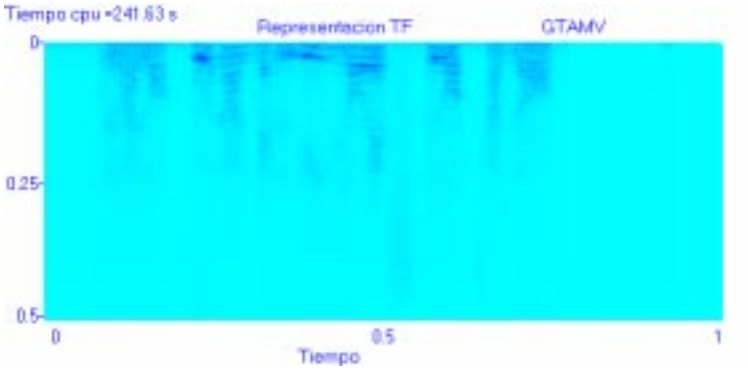

Figura 5: Transformada GTAMV de la señal sintetizada

De las características enunciadas anteriormente, podemos apreciar claramente lo dicho para las consonantes líquidas (ver las palabras "la lo-" al inicio de la frase). Para las nasales, en la palabra "niño", podemos apreciar el contenido energético en bajas frecuencias y el tercer formante. El segundo formante débil no se llega a apreciar. Para el sonido *fricativo* que aparece en la palabra "explotado", se aprecia claramente el contenido energético en la banda alta de frecuencias de 3000-5000 Hz. En estas Figuras no se aprecia bien las características de las oclusivas, para apreciar esto mejor hace falta representar un intervalo temporal más pequeño, figuras que no incluimos por falta de espacio.

Finalmente queremos presentar el análsis mediate la GTAMV de la misma frase ahora sintetizada. Esto aparece en la Figura 5. Como se aprecia en esta Figura, la señal sintetizada muestra una mayor *armonizidad* en frecuencias. Esta es la caracteristica que da la peculiaridad al sonido generado por ordenador (tono metálico). Además se aprecia que los formantes no estan bien definidos.

4 Conclusión

En este trabajo hemos querido mostrar las cualidades y desventajas que tiene la *transformada general de ajuste mínimo de ventana* (GTAMV), frente a otras representaciones tiempo-frecuencia, en el análisis del habla. Esta transformada ha sido desarrollada por los autores en [5] y [4]. Este método de análisis determina localmente las características espectrales y selecciona el tamaño de la ventana de análisis utilizada en cada frecuencia.

Como se ha podido apreciar en los ejemplos presentados la GTAMV es una herramienta eficiente para determinar los formantes de señal hablada. No produce representaciones de tan buena resolución en frecuencias como el STFT de banda estrecha. Sin embargo se puede considerar complementaria a la distribución SPWV mostrando claramente los formantes y mayor estructura armónica que esta distribución. En los ejemplos mostrados se pueden apreciar las características básicas de las estructuras consonánticas.

Referencias

[1] Torregrosa F. *Transmisión y recepción de la información verbal en deficientes auditivos en aulas de integración.* PhD tesis, Universidad de Valencia, 1997.

[2] Anderson J. A wavelet magnitude analysis theorem. *IEEE Trans. Sig. Proc.*, 41:3541–3543, 1993.

[3] Cerdá S. Romero J. Análisis tiempo frecuencia adaptado a las características espectrales de la señal. En *XXVIII JORNADAS NACIONALES DE ACÚSTICA Y ENCUENTRO IBÉRICO DE ACÚSTICA*, pages 177–180. SEA, 1997.

[4] Cerdá S. Romero J. La transformada GTAMV: Transformada General de Ajuste Mínimo de Ventana. *Revista de Acústica*, 30:13–20, 1999.

[5] Cerdá S. *Análsis tiempo-frecuencia adaptado a la señal: síntesis y contribuciones.* PhD tesis, Universidad de Valencia, 1999.