Percepción de analogos tonales de palabras en condiciones de ruido blanco y ruido filtrado en diferentes regiones espectrales

Julio González Alvarez * y Teresa Cervera Crespo**

*Universitat Jaume I de Castellón. **Universitat de València

El objetivo del trabajo que aquí se presenta es estudiar el reconocimiento del habla a partir de réplicas sinusoidales dado que no existe, hasta el momento, información al respecto que permita la comparación con las tasas de inteligibilidad encontradas para la lengua inglesa; evaluar la resistencia de las SWS a situaciones diferentes de enmascaramiento con ruido presentado en diferentes bandas espectrales.

1. Introducción

El habla es percibida, en condiciones naturales, en presencia de ruido o de otros sonidos. Sin embargo, el sistema perceptual humano es muy eficiente a la hora de extraer la información relevante necesaria para llevar a cabo el proceso de percepción.

El "Análisis de la Escena Auditiva" (ASA, *Auditoriy Scene Análisis*, Bregman, 1990) proporciona un marco teórico para explicar este hecho. El oyente fusionaría en un unico percepto un conjunto de elementos acústicos que comparten propiedades comunes al proceder de una única fuente. Esto permitiría su separación de otros componentes acústicos sin necesidad de recurrir a procesos de alto nivel.

Un desafío importante a los principios del ASA como explicación de la percepción del habla proviene de los trabajos de Remez y colaboradores (Remez, Rubin, Pisoni & Carrell, 1981; Remez, Rubin, Berns, Pardo & Lang, 1994). Utilizando réplicas sinuosidades de habla SWS (*sine-wave speech)* encuentran que los sujetos son capaces de reconocerlas con bastante exactitud. Estas SWS consisten en la presentación simultánea de tres ondas sinusoidales que varían en su frecuencia y amplitud imitando las resonancias de un tracto vocal. En la práctica se generan con tonos puros cuyos valores proceden de los formantes extraídos mediante análisis LPC, u otro algoritmo, a partir de estímulos naturales del habla.

Estos estímulos son percibidos en ausencia de los principios organizativos del ASA, Las trayectorias que siguen las ondas sinusoidales no permiten su agrupación según principios de:

- relación armónica
- comodulación de amplitud
- comodulación en frecuencia que rigen en los sonidos del habla natural

Por otra parte, estos estímulos retienen únicamente información espectral gruesa variante en el tiempo, mientras que información fina como la frecuencia fundamental, la estructura armónica, el ancho de banda de los formantes o la señal aperiódica correspondiente a fricativas y oclusivas, queda eliminada.

Su única coherencia, de acuerdo con los postulados de Remez y colaboradores, correspondería a un nivel más abstracto. El oyente reconoce SWS al interpretar el estímulo en términos fonéticos, fuera de los principios del ASA. Esta forma de percibir, sostienen, sería específica para el lenguaje.

Sin embargo, no es del todo cierto que las tres réplicas sinusoidales no contengan principios organizativos. De hecho contienen (Barker, 1998):

- un comienzo y final común
- una misma localización espacial
- una alta correlación en frecuencia y amplitud de las trayectorias sinusoidales

Además, la literatura experimental ha demostrado que, cuando a las SWS se le añade una "pista" de agrupamiento como comodulacion en amplitud (Carrell & Opie, 1992), mejora la inteligibilidad. Y cuando se le añade una pista de segregación como la presentación dicótica (Remez *et al*., 1994) u otra SWS simultáneamente (efecto *cocktail party*, Barker & Cooke, 1998), la inteligibilidad empeora. Esto demostraría que la perceptión de SWS es sensible a los principios organizativos del ASA.

2. Método

2.1. Sujetos

En el experimento participaron un total 180 sujetos de ambos sexos, 30 por cada una de las seis condiciones experimentales. Los sujetos eran estudiantes de los últimos cursos de Psicología de las Universidades de Castellón y Valencia, no presentaban problemas de audición y no tenían experiencia previa con las SWS.

2.2. Materiales

Estímulos

Se partió de dos listas de 25 palabras fonéticamente equilibradas (Listas 6 y 7 de Cárdenas & Marrero, 1994). Las palabras fueron pronunciadas por un varón y grabadas digitalmente con un micrófono SHURE SM58 a una frecuencia de muestreo de 10 kHz.

Para cada palabra se obtuvieron los valores de frecuencia y amplitud de las trayectorias de los tres primeros formantes (F1, F2, F3). Esto se llevó a cabo utilizando el programa de análisis CSRE (*Computerized Speech Research Environment*, Avaaz Innovations Inc) en el que la señal es multiplicada por una ventana Hamming y pre-enfatizada para asegurar suficiente energía en los formantes altos:

- se estiman coeficientes de autocorrelación
- se aplica el algoritmo Levinson-Durbin para obtener coeficientes LPC
- se aplica un procedimiento de “peak-picking” para seleccionar los tres primeros valores más altos de cada espectro de densidad de energía (ver Cuadro 1)

El proceso arroja valores de frecuencias y amplitud cada 10 milisegundos que sirven para alimentar tres osciladores generadores de tonos puros. Este proceso se realizó utilizando las rutinas para MATLAB desarrolladas por S. Frost & P. Rubin. en los laboratorios Haskins. De este modo, por cada palabra se crea una réplica de 3 ondas sinusoidales que siguen las trayectorias de los tres primeros formantes. (ver ejemplo en Figura 1)

CREACIÓN DE ESTÍMULOS

ANÁLISIS DE LA SEÑAL NATURAL
(CSRE-Computerized Speech Research Environment)

1. **Señal pre-enfatizada**
2. **Multiplicada por ventana Hamming**
3. **Obtención coeficientes de Autocorrelación**
4. **Algoritmo Levinson-Durbin (LPC)**
5. **Procedimiento "peak-picking"**

↓

Valores de Frecuencias y Amplitudes de F1, F2 y F3

↓

SÍNTESIS DE RÉPLICAS SINUSOIDALES
(MATLAB)

Generación de 3 ondas sinusoides

↓

Fichero WAV

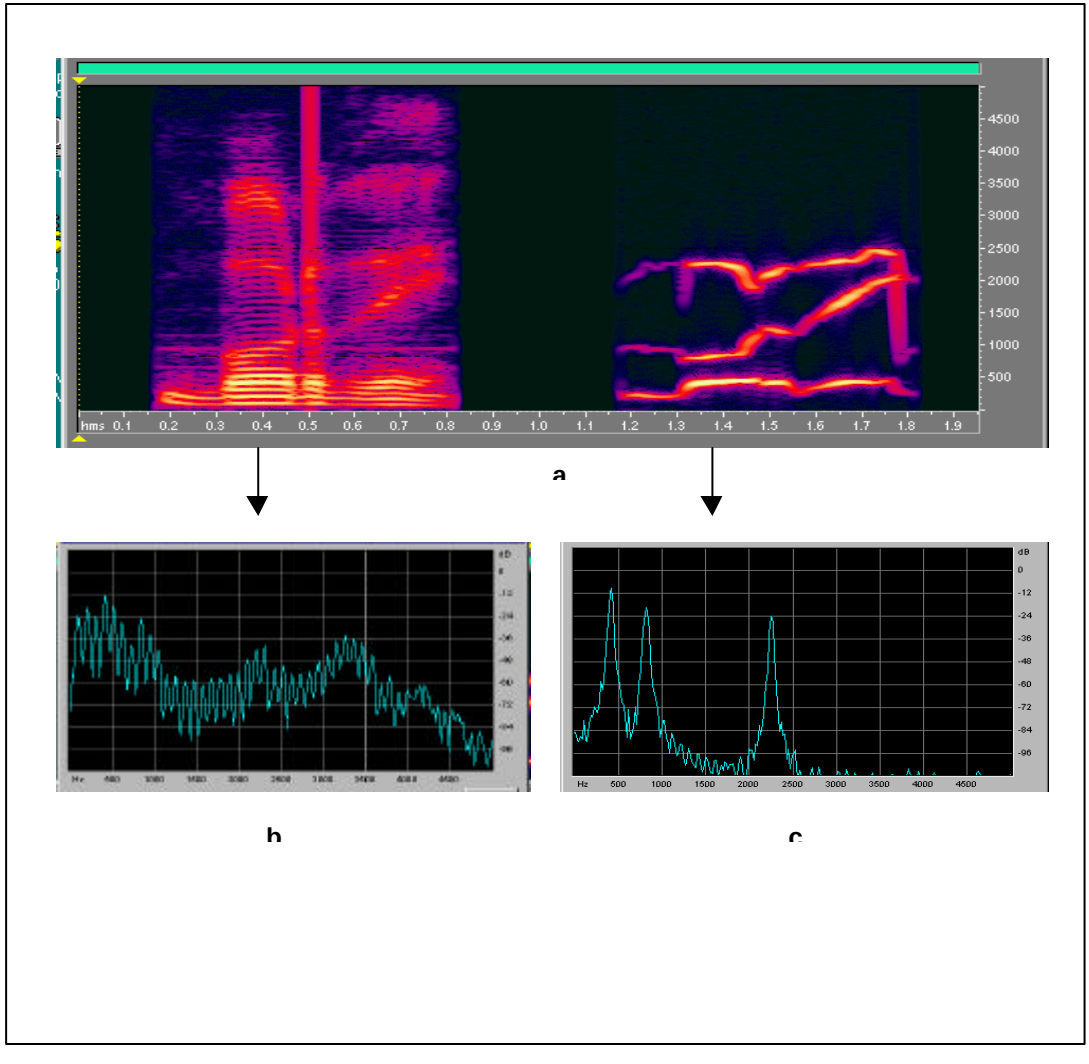

Figura 1: a) Espectrogramas de la palabra natural “borde” y de su réplica SWS. b) Espectro (FFT, 1024 ptos) de la parte central de la vocal “o” natural. c) ídem. de la réplica SWS.

Condiciones de enmascaramiento por ruido

Al conjunto de 50 réplicas SWS de palabras se les añadió ruido blanco a una relación Señal/Ruido de 0 dB. El ruido blanco era previamente filtrado mediante

- Filtro PasoBajo 5 kHz o ruido de banda ancha (LP5)
- Filtro Paso-Bajo 2 kHz (LP2)
- Filtro Paso-Bajo 1 kHz (LP1)
- Filtro Paso-Banda 1-2 kHz (BP1-2)
- Filtro Paso-Banda 2-3 kHz (BP2-3)
- Además se presentó una condición control de habla SWS sin ruido

3. Procedimiento

Cada sujeto fue asignado a una de las seis condiciones. La administración fue individual en dos sesiones separadas por un corto ensayo. La presentación de estímulos se realizó a través de auriculares PHILLIPS SBC HP530 en una sala aislada. Cada ensayo experimental constaba de la presentación de dos veces repetidas de cada estímulo, dejando un silencio de 2 segundos entre ambas. La tarea del sujeto consistía en la trascripción del estímulo utilizando el teclado del ordenador. Las instrucciones enfatizaron el que los sujetos transcribieran cualquier segmento fonético que pudieran identificar del estímulo, sin que necesariamente esperaran palabras del castellano.

Previamente, los sujetos recibian una sesión de práctica escuchando dos veces sucesivas 5 ensayos de réplicas-SWS de frases y 5 de réplicas-SWS de palabras distintas a las presentadas en el experimento.

Las respuestas de los sujetos fueron evaluadas por dos observadores diferentes y se obtuvieron medidas de fiabilidad de las mismas. Se identificó el número de fonemas correctamente identificados por cada oyente para cada palabra. Se consideraron errores tanto las substituciones como las omisiones de fonemas.

4. Resultados

El comportamiento de las réplicas SWS en lo que se refiere a su inteligibilidad en las distintas condiciones de ruido es comparable al reconocimiento del habla natural según las regiones espectrales enmascaradas. Los porcentajes de fonemas correctamente identificados muestran que (Figura 2):

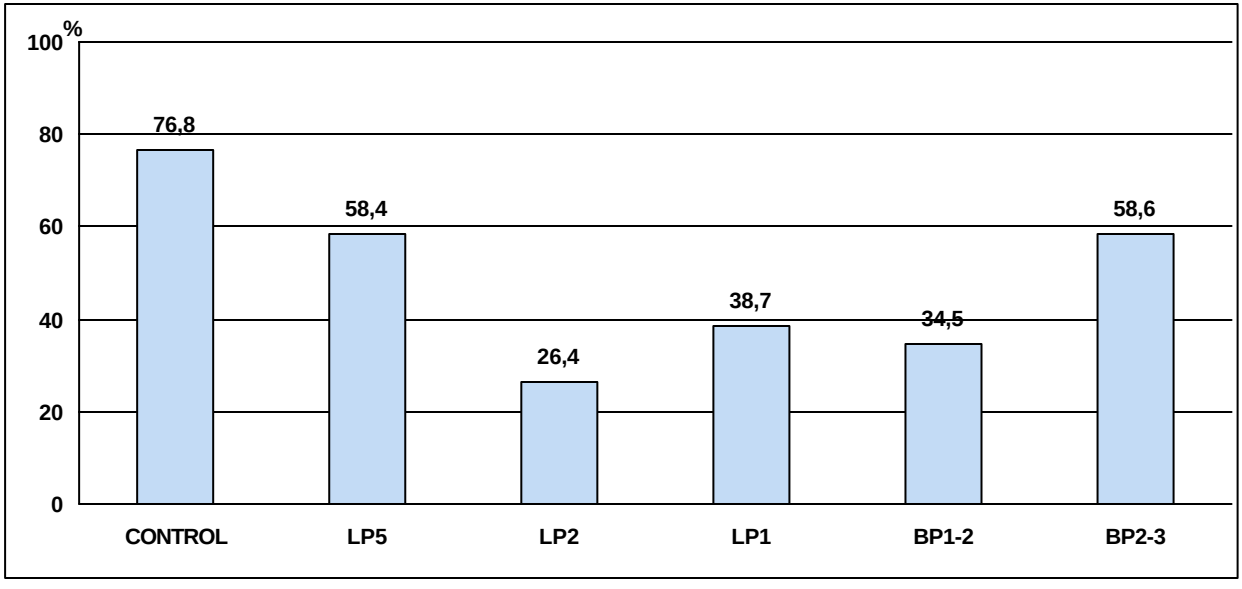

Figura2: Porcentaje de fonemas identificados en las réplicas SWS en las distintas condiciones de enmascaramiento por ruido

El grupo control presenta tasas de aciertos semejantes a las que cabría esperar, tal como aparece en la literatura con las SWS del inglés. El 76.8 % de los fonemas castellanos es identificado correctamente, frente al 69 % de las sílabas identificadas en Remez *et al*. (1994) o el 69 % de los fonemas identificados en Carrell & Opie (1992).

El ruido de banda ancha (LP5) es el que causa menos problemas de inteligibilidad. Esto también ocurre con el habla normal, por cuanto toda la energía enmascarante se distribuye con la misma densidad en todo el ancho de banda de la señal, sin afectar de modo particular regiones espectrales críticas en el reconocimiento del habla.

El ruido de banda estrecha produce más enmascaramiento si afecta frecuencias inferiores a los 2000 Hz. El mayor deterioro en el reconocimiento del habla SWS se da en la condición LP2 porque en ella toda la energía enmascarante se concentra en la banda 0-2000 Hz por la que transcurren fundamentalmente las sinusoides que corresponden a F1 y F2. Le sigue muy de cerca la condición BP1-2, con ruido en 1000-2000 Hz, y LP1 con ruido 0-1000 Hz.

En la condición BP2-3 se conserva un grado importante de inteligibilidad porque el ruido se concentra en la banda 2000-3000 Hz afectando sobre todo al tercer sinusoide (F3) y preservándose en gran parte los dos primeros (F1 y F2).

REFERENCIAS

Barker, J. *The Relationship between Speech Perception and Auditory Organization: Studies with spectrally reduced speech*. Tesis doctoral. Department Computer Science, University of Sheffield, UK, 1998.

Barker, J. & M. P. Cooke. Is the sine wave speech cocktail party worth attending? En *Speech Communication*, 27, 159-174, 1999.

Bregman, A. S. *Auditory Scene Analysis*. Cambridge, MA: MIT Press, 1990.

Cárdenas, M. R. & V. Marrero. *Cuaderno de logoaudiometría*. Madrid: UNED, 1994.

Carrell, T. D. & J. M. Opie. The effect of amplitude comodulation on auditory object formation in sentence perception. En *Perception & Psychophysics*, 52, 437-445, 1992.

Remez, R. E., P. E. Rubin, S. M. Berns, J. S. Pardo & L. M. Lang. On the perceptual organization of speech. En *Psychological Review*, 101, 129-156, 1994.

Remez, R. E., P. E. Rubin, D. B. Pisoni & T. D. Carrell. Speech perception without traditional speech cues. En *Science*, 212, 947-950, 1981.