CARACTERÍSTICAS ACÚSTICAS DE LOS FONEMAS EN PRESENCIA DE RUIDO: CONSECUENCIAS PARA LA INTELIGIBILIDAD EN RECINTOS

PACS: 43.55.Hy

Feijóo, Sergio; Alvarez, José Manuel; Chisca, Benjamín
Universidad de Santiago de Compostela
Dpto. de Física Aplicada, Fac. de Física
15702 Santiago de Compostela. España
Tel: 981 563100 ext. 14044
Fax: 981 520676
E-mail: fasergio@usc.es

ABSTRACT

The presence of noise alters the acoustical characteristics of phonemes, affecting its perceived intelligibility. The changes produced by background noise in a set of phonemes with Signal-to-Noise ratios of 24,12,6, and 0 dB, and their relationship with the intelligibility estimated by a group of listeners, have been studied. Results show that the acoustic variables affected by the presence of noise are unable to explain listener's results, except in particular instances for certain syllables.

RESUMEN

Los fonemas producidos en presencia de ruido ven alteradas sus características acústicas, afectando a la inteligibilidad del habla. Hemos estudiado los cambios producidos por el ruido de fondo sobre una serie de fonemas con una relación señal-ruido (RSR) de 24, 12, 6 y 0 dB, y hemos relacionado dichos cambios con la inteligibilidad estimada por un grupo de oyentes. Los resultados demuestran que las variables acústicas que se ven afectadas por la presencia de ruido no son capaces de explicar los resultados obtenidos en los experimentos de percepción, salvo de forma particular para algunas sílabas concretas.

INTRODUCCION

El concepto de "*Inteligibilidad del Habla*" se ha venido usando como referencia para determinar la utilidad y validez de una serie de canales de transmisión del sonido (recintos, sistemas de megafonía, sistemas de transmisión de voz, etc.). En ese contexto, la *inteligibilidad* nos permite determinar si un canal es válido para permitir la correcta percepción de la palabra y el mensaje hablado. En la mayoría de canales de transmisión el principal problema que afecta a la percepción del habla es la presencia de "Ruido de Fondo", aunque puede haber otros factores que distorsionen la percepción como la presencia de ecos, reverberación, etc. Estos otros factores, sin embargo, pueden ser considerados como un ruido de fondo efectivo que se añade al ruido de fondo existente en el canal. Esta es la aproximación seguida, por ejemplo, en la determinación de índices acústicos representativos de la *inteligibilidad* tales como el STI [1].

El uso del concepto de Inteligibilidad para un canal determinado se basa en una serie de puntos: a) la *inteligibilidad* puede ser estimada mediante la realización de experimentos de

percepción por una serie de oyentes como un porcentaje de respuestas correctas; b) el porcentaje de respuestas correctas en un determinado test depende fundamentalmente de las características acústico-fonéticas del habla y de las características ambientales propias del canal en cuestión, mientras que otros factores tales como el hablante, la variabilidad intra-fonética o el oyente contribuyen en mucha menor medida; c) el porcentaje de acierto de un test puede relacionarse con medidas acústicas realizadas sobre el canal, bien sobre las magnitudes directas que afectan al habla (p. ej. nivel de ruido de fondo, RSR, tiempo de reverberación), bien sobre magnitudes indirectas (p. ej. el STI, AI, %ALcons, etc.) [2]; d) la *inteligibilidad* medida como porcentaje de aciertos en un test de percepción está directamente relacionada con la *sensación de inteligibilidad* que experimentaría un oyente sin problemas auditivos cuando escucha el habla transmitida por el canal.

En un trabajo previo [3] estudiamos la percepción de una serie de estímulos formados por todas las posibles sílabas que ocurren en posición inicial, pronunciadas por 4 hablantes (2 hombres y 2 mujeres) en 4 condiciones de RSR: 24, 12, 6 y 0 dB. Los oyentes debían identificar la consonante inicial de la sílaba. Se escogieron sílabas como estímulos para minimizar el impacto de factores extra-acústicos que condicionan la identificación fonética en el habla natural. Los resultados de los experimentos mostraron que la consistencia del test era baja en función del factor hablante, cuestionando la suposición de que la identificación fonética depende en su mayor parte de las características acústico-fonéticas de los estímulos, y de la interacción entre éstas y las características ambientales (ruido de fondo, reverberación). Además, observamos que el efecto del ruido no era el mismo sobre la percepción de los diferentes fonemas: mientras algunos fonemas mostraban buenos porcentajes de reconocimiento aún en las condiciones de ruido más desfavorables, otros eran percibidos de forma ambigua aún en las condiciones de ruido más favorables.

En el presente trabajo centramos nuestro interés en la relación entre los cambios que se producen en las características acústicas de los fonemas, debidos a la presencia de ruido, y la percepción fonética. El hecho de que haya fonemas que, desde el punto de vista de la percepción auditiva, son robustos frente al ruido, podría deberse a que sus características acústicas más directamente relacionadas con la percepción son poco afectadas por las características acústicas del ruido de fondo. Por el contrario, los fonemas cuya percepción es más sensible a la presencia de ruido de fondo podrían tener sus principales características acústicas fuertemente afectadas por el ruido de fondo. En parte este hecho podría ser explicado teniendo en cuenta la forma de definir la relación señal-ruido: mientras que el nivel de presión del ruido de fondo es prácticamente constante cuando usamos un ruido de amplio espectro tipo rosa o blanco, el nivel sonoro del habla se obtiene promediando sobre una señal que sufre grandes variaciones a corto plazo, tanto en amplitud como en composición espectral. Como, además, la mayor parte de la energía del habla se concentra en las vocales, el nivel sonoro promedio está básicamente relacionado con la energía de las vocales, cuya percepción presenta generalmente pocos problemas. Sin embargo, los fonemas más confusos del habla son consonantes de baja amplitud y/o corta duración, que apenas contribuyen al nivel sonoro promedio del habla, y para los que la RSR es mucho más desfavorable que para las vocales: para una misma RSR, diferentes consonantes tendrán diferentes relaciones consonante-ruido, debido a las diferencias en amplitud y duración que existen entre los fonemas (Fig.1). Para investigar estos hechos hemos seleccionado una serie de 10 silabas de las utilizadas originalmente en los experimentos citados, y hemos calculado una serie de parámetros acústicos que están directamente relacionados con las modificaciones producidas por el ruido en los fonemas implicados. Por último, estableceremos la correlación existente entre los parámetros acústicos y las respuestas de los oyentes para las diferentes condiciones de RSR.

METODO

Estímulos

Los estímulos utilizados originalmente en el experimente previo consistían en una serie de sílabas CV formadas por la combinación de las consonantes /p,t,k,b,d,g,θ,f,s,ʃ,x, tʃ,m,n,ŋ,l,λ,rr/ con las vocales /a,e,i,o,u/, pronunciadas por 2 hombres y 2 mujeres. Se construyeron estímulos con RSR de 24, 12, 6 y 0 dB añadiendo ruido rosa a los estímulos

originales. Los oyentes debían identificar la consonante inicial en la sílaba. Los detalles del experimento de percepción pueden verse en [3]. Para establecer la relación entre características acústicas de los fonemas y percepción de los oyentes hemos seleccionado una serie de sílabas en base a los siguientes criterios:

a) Incluiremos sílabas de distintas sensibilidades frente al ruido
b) Las sílabas seleccionadas deben formar parte de palabras del lenguaje natural que sean “pares mínimos” en los que el fonema diferenciador sea la consonante inicial de la sílaba

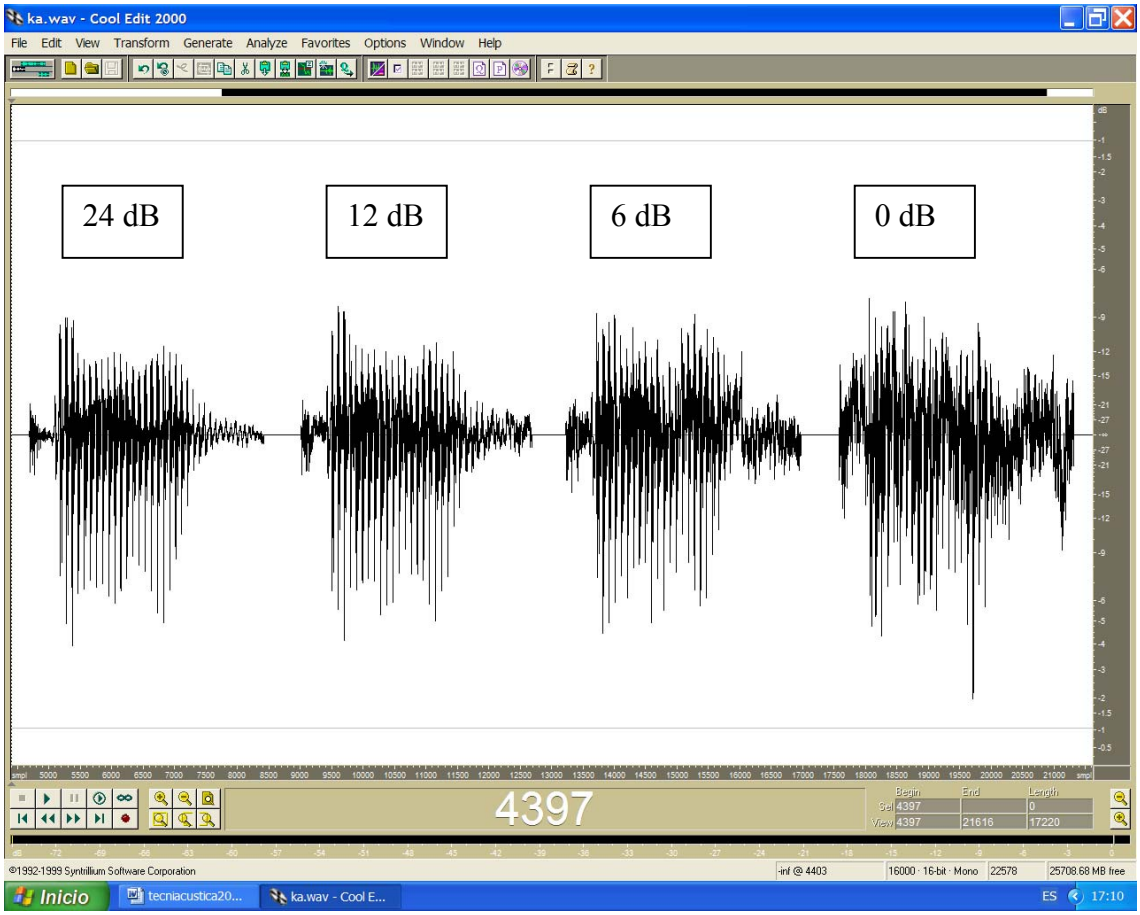

Fig. 1 Formas de onda correspondientes a la sílaba /ka/ de uno de los hablantes en las distintas condiciones de RSR (de izquierda a derecha): 24, 12, 6 y 0 dB.

La primera condición es esencial para poder establecer una relación entre los cambios en los parámetros acústicos y las diferentes respuestas de los oyentes. La segunda condición se debe a que estamos construyendo estímulos para experimentos de percepción basados en palabras y frases formadas por un número controlado de sílabas, tratando de reducir el número de casos para introducir diferentes hablantes y mantener la duración de los experimentos en límites razonables. En este tipo de experimentos de percepción los fallos ocurren cuando hay más de un posible candidato como respuesta (“pares mínimos”) [4]. En base a estas condiciones, las sílabas seleccionadas fueron:

-grupo A: /fe, ga, ti/
-grupo B: /bo, ko, di, po/
-grupo C: /pa, ka, ʃu/

En el grupo A están las sílabas más sensibles frente al ruido; en el grupo B las sílabas sensibles a niveles altos de ruido; y en el grupo C las sílabas menos sensibles al ruido.

Variables acústicas

Las características acústicas escogidas deben tener relación con el efecto del ruido sobre las mismas y ser representativas de la percepción fonética. Las variables consideradas fueron:

- Ec: Energía de la consonante inicial de la sílaba en dB
- Ev: Energía de la vocal en dB
- Er: Energía del ruido en dB
- CF1: Energía del pico principal del espectro de la consonante en dB
- CF2: Energía del 2º pico del espectro de la consonante en dB
- Var1 = Ec (RSR) – Ev (RSR)
- Var2 = Ec (limpia) – Er (RSR)
- Var3 = Ec (RSR) – Ec (limpia)
- Var4 = CF1 (RSR) – CF1 (limpia)
- Var5 = CF2 (RSR) – CF2 (limpia)

La leyenda entre paréntesis se refiere a una determinada relación señal-ruido (RSR), o a la energía de la señal original sin ruido añadido (limpia). *Var1* representa la relación consonante-vocal para una determinada RSR: a medida que el ruido aumente, esta relación disminuye; *Var2* representa la relación consonante-ruido: a medida que el ruido aumenta, esta relación disminuye; *Var3* representa el aumento de energía de la consonante provocado por el ruido; *Var4* y *Var5* representan, respectivamente, el aumento de energía en el primer y segundo pico más prominente de la consonante debido al ruido. Algunas de las variables escogidas fueron empleadas por Dubno & Levitt [5] en su trabajo sobre confusión de consonantes.

RESULTADOS Y DISCUSION

En primer lugar se procedió a determinar si el subconjunto de sílabas escogido era representativo de todo el conjunto. Para ello se obtuvo para cada RSR el porcentaje de error promedio sobre el conjunto de sílabas seleccionadas, y se correlacionó con el porcentaje de error promedio sobre todas las sílabas originales. El coeficiente de correlación entre ambos conjuntos (r = 0.99) permite afirmar que el nuevo subconjunto de sílabas justifica el 98% de la varianza del conjunto original. A continuación ajustamos el porcentaje de error del subconjunto en función de la RSR de los estímulos, obteniéndose un ajuste de tipo exponencial decreciente

$$\%\text{Error} = 2.5 + 40.6\, e^{-\text{RSR}/9.6} \qquad (1)$$

La bondad del ajuste del nuevo subconjunto es superior a la del conjunto total de sílabas (Chi^2 = 0.08 para el nuevo subconjunto vs. Chi^2 = 0.39 para el conjunto total). Estos dos análisis demuestran que el nuevo subconjunto es suficientemente representativo de la percepción de las sílabas en diferentes condiciones de RSR.

Las variables acústicas no pudieron ser completamente determinadas en los casos de las sílabas /po/ y /bo/, ya que apenas mostraban trazas de la consonante, por lo que no fueron incluidas en los análisis posteriores. La Tabla I muestra los coeficientes de correlación (r) obtenidos entre las variables individuales y las respuestas de los oyentes, de forma conjunta y separada para cada una de las 8 silabas.

	Ec	Ev	CF1	CF2	Var1	Var2	Var3	Var4	Var5
Todas	-0.32	-0.33	0.14	-0.2	-0.29	0.53	-0.57	-0.24	-0.46
/di/	-0.48	-0.72	-0.23	-0.65	-0.36	0.60	-0.75	-0.74	-0.81
/ti/	-0.79	-0.74	-0.50	-0.53	-0.78	0.83	-0.87	-0.69	-0.68
/ko/	-0.35	-0.62	0.21	-0.34	-0.25	0.64	-0.69	-0.84	-0.64
/ʃu/	-0.45	-0.36	-0.02	-0.11	-0.41	0.56	-0.73	-0.81	-0.84
/ga/	-0.50	-0.34	-0.03	-0.36	-0.51	0.50	-0.39	-0.21	-0.67
/ka/	-0.37	-0.27	0.09	-0.42	-0.38	0.46	-0.52	-0.04	-0.27
/pa/	-0.59	-0.61	-0.34	-0.62	-0.55	0.46	-0.45	-0.29	-0.40

/fe/	-0.55	-0.56	-0.61	-0.52	-0.52	0.43	-0.43	-0.45	-0.39

Tabla I. Coeficientes de correlación (r) entre las respuestas de los oyentes y las variables acústicas para las 8 silabas juntas y por separado

Los resultados muestran que ninguna de las variables consideradas es capaz de explicar la percepción de los oyentes sobre el conjunto de las 8 silabas. A nivel particular para cada silaba, algunas de las variables (Ev, Ec, Var1, Var2, Var3, Var4 y Var5) obtienen coeficientes de correlación superiores a 0.70 para las sílabas /di/, /ti/, ko/ y /ʃu/.

A continuación, realizamos un análisis de regresión múltiple para determinar si la agrupación de variables podría mejorar la relación con las respuestas de los oyentes. Los resultados obtenidos por los mejores subconjuntos de variables pueden verse en la Tabla II.

	Variables	R^2
Todas	Var3	0.32
/di/	Var5,Var3	0.67
/ti/	Var3,Var2,Ec	0.85
/ko/	Var4,Var3,Var5,Var2,Ev	0.78
/ʃu/	Var5,Var4,Var3,Var2,Ec	0.84
/ga/	Var5,Var1,Var2,Ec	0.50
/ka/	Var3,Var2,CF2,Var1	0.32
/pa/	CF2,Ev,Ec	0.44
/fe/	CF1,Ev,Ec,CF2	0.48

Tabla II. Coeficientes R^2 obtenidos en el análisis de regresión múltiple por los mejores subconjuntos de variables para las 8 silabas juntas y por separado

Vemos que la inclusión de más de una variable acústica ayuda a mejorar la relación con las respuestas de los oyentes, en particular para /ti/, /ko/ y /ʃu/ con varianzas explicadas alrededor del 80%, aunque el resto de las sílabas también se beneficia en menor medida. Cuando se consideran todas las sílabas juntas, sin embargo, la agrupación de variables no supera la varianza explicada por una sola de ellas.

Las variables propuestas están relacionadas con las modificaciones energéticas producidas por el ruido de fondo sobre las consonantes seleccionadas. Estas modificaciones no son uniformes sobre los diferentes tipos de fonemas, ya que algunos de ellos tienen una duración y energía mayor que otros. Por tanto, sería de esperar que los fonemas más afectados por la interacción con el ruido de fondo sufrieran una perdida de inteligibilidad mayor que los menos afectados. Los resultados, sin embargo, nos muestran un panorama diferente.

Si una consonante tiene poca energía, el efecto del ruido de fondo sobre sus características acústicas, y por tanto sobre su percepción, será más acusado que para una consonante de mayor energía. Dado que los estímulos empleados están formados por sílabas CV, la diferencia en energía (dB) entre la consonante y la vocal, para una determinada condición de ruido, representaría la Amplitud de Modulación (AM) de la silaba. Es lógico pensar, entonces, que el efecto del ruido será disminuir la AM y que cuanto más disminuya la amplitud de modulación, más se verá afectada la percepción de la consonante. Este es el concepto en el que se basa, por ejemplo, el índice STI. Podemos definir esa Reducción de la Amplitud de Modulación (en dB) sufrida por la sílaba para una determinada RSR como:

$$\text{RAM (RSR)} = (\text{Ec} - \text{Ev})_{\text{limpia}} - (\text{Ec} - \text{Ev})_{\text{RSR}} \qquad (2)$$

Veamos, como ejemplo, lo que sucede en el caso particular de 2 fonemas que comparten modo y lugar de articulación. La Tabla III nos muestra la energía, duración, RAM y porcentaje de error en el experimento de percepción para los fonemas /k/ (de la sílaba /ka/, hablantes 2, 3 y 4) y /g/ (sílaba /ga/, hablante 1).

Lo primero que se observa es que no existe uniformidad en el efecto del ruido sobre la variable escogida RAM dentro de la misma sílaba /ka/: las características acústicas de la

consonante /k/ del hablante 4 se ven menos afectadas por el ruido que las de los otros dos hablantes. La percepción del fonema /k/ tampoco es uniforme sobre estos tres hablantes. Si existiera una relación directa entre RAM y el porcentaje de error, una disminución de RAM llevaría aparejada una consiguiente reducción en el porcentaje de error. La Tabla III nos muestra que este no es el caso. Una reducción de RAM de -7 dB para el hablante h3 (6 dB de condición RSR) supone un porcentaje de error en le identificación de /k/ del 45.4%, mientras que una mayor reducción de RAM para el hablante h2 (-11.5 dB, condición 0 dB de RSR) mantiene el porcentaje de error en el 0%. Si ahora comparamos lo que pasa con el fonema /g/, vemos que una reducción de RAM de -4.1 dB (condición 6 dB de RSR) implica un porcentaje de error del 95.4%. Teóricamente, los fonemas menos afectados a nivel de percepción por el ruido serían los de mayor energía (/k/ del hablante h4 y /g/ del h1). Sin embargo, los fonemas menos afectados por el ruido son los de mayor duración (/k/ de h2 y /k/ de h4). La duración de un fonema no es afectada directamente por el ruido de fondo, y consonantes de menor duración (por ejemplo /p/ de la sílaba /pa/ del hablante h1, que tiene 12 ms de duración), no sufren de manera tan acusada el descenso en el porcentaje de identificación (4.5 % de error en la condición de RSR de 0 dB).

	Ec(limpia)(dB)	Duración(ms)	RSR-24	RSR-12	RSR-6	RSR-0
/ka/- h2	-30.8	21	-0.7 (0)	-4.4 (0)	-6 (0)	-11.5 (0)
/ka/- h3	-32.7	17	-0.5 (0)	-5.5 (0)	-7 (**45.4**)	-13.3(**68.2**)
/ka/- h4	-26.8	28	-0.4 (0)	-2 (0)	-3.3 (0)	-3.9 (0)
/ga/- h1	-25.2	19	-0.3 (0)	-1.5 (**22.7**)	-4.1 (**95.4**)	-6 (**95.4**)

Tabla III. Energía (dB), duración (ms), RAM (dB) y porcentaje de error en el experimento de percepción (%, entre paréntesis), para los fonemas /k/ (de la sílaba /ka/, hablantes 2, 3 y 4) y /g/ (sílaba /ga/, hablante 1), en las distintas condiciones de RSR.

CONCLUSIONES

- No existe uniformidad en la acción del ruido sobre las características acústicas de un fonema concreto (depende del hablante)
- No existe uniformidad en la acción del ruido sobre la percepción de un fonema concreto (depende del hablante)
- No existe, por tanto, una relación directa para el conjunto de todos los fonemas seleccionados entre el efecto del ruido sobre las variables acústicas y la percepción de los fonemas
- Sí parece existir una relación uniforme entre modificaciones en las características acústicas y la percepción para algunos fonemas concretos por separado, como /di/, /ti/, /ko/, /ʃu/
- A pesar de que la inteligibilidad (definida como un porcentaje de acierto) sobre el conjunto de sílabas seleccionado esté directamente relacionada con la inteligibilidad sobre todas las sílabas originales, y de que dicha inteligibilidad esté directamente relacionada con la RSR, ninguna de las variables acústicas consideradas puede justificar la inteligibilidad de los fonemas.

REFERENCIAS

[1] Steeneken, H.J.M., Houtgast, T. “Basics of the STI measuring method”, en *Past, present and future of the Speech Transmisión Index*, Ed. TNO, 13-43 (2002)
[2] Levitt, H., Webster, J.C. “Effects of noise and reverberation on speech”, en *Handbook of acoustical measurements and noise control*, Ed. Acoustical Society of America, Cap. 16 (1989)
[3] Feijóo, S., Alvarez, J.M. “Acústica de aulas: percepción fonética en presencia de ruido”, Publicación Oficial de Tecniacústica2003, Bilbao (2003)
[4] Feijóo, S., Fernández, S., Barros, N., Balsa, R. “Context effects and acoustic cues for the auditory identification of spanish fricatives /f/ and /θ/”, Acta Acustica united with Acustica, 88, 113-126 (2002)
[5] Dubno, J.R., Levitt, H. “Predicting consonant confusions from acoustic análisis”, Journal of the Acoustical Society of America, 69, 249-261 (1981)