SINCAS: un conversor texto-voz en castellano

Josep Martí Roca

Daniel Niñerola Chifoni

Departamento de Acústica

Escuela de Telecomunicación La Salle Bonanova - Barcelona

1. OBJETIVOS Y LIMITACIONES

Hasta el presente han apareciso numerosos sintetizadores de voz basados en tecnologias diversas. Podriamos citar en primer lugar los "vocoders" de parámetros articulatorios (DUNN, 1950) (STEVENS, 1953) (COKER, 1968) que no tuvieron muchas aplicaciones tecnológicas posteriores. De implantación más reciente han sido los "vocoders" de canales paralelos utilizados en telecomunicación (SCHROEDER, 1966) (HOLMES, 1980) (PETERSEN, 1983). La introducción de la tecnología digital ha hecho avanzar notablemente las posibilidades de los sintetizadores de voz totalmente informatizados. Destacan particularmente dos tecnologías: la síntesis por formantes (KLATT, 1980) (FANT, 1963) y la predicción lineal (ATAL, 1951) (ITAKURA, 1972) (MARKEL, 1982). Entre los trabajos más próximos a nosotros podemos mencionar los del Centro Informático de la Universidad de Valencia (TORRES, 1982, 1983) de la Escuela Superior de Telecomunicación de Madrid y el nuestro propio, para la síntesis del catalán, realizado en la Escuela Universitaria de Telecomunicación La Salle Bonanova de Barcelona (MARTI, 1986 a, 1986 b).

Nuestro sistema SINtetizador de CAStellano (SINCAS) pretende ser un conversor texto-voz que a partir de una entrada de texto escrito ortográficamente ofrece una salida oral inmediata de acuerdo con las reglas fonéticas del idioma y sin restricción de vocabulario.

Se ha estudiado la fonética de las cinco vocales castellanas y de las transiciones consonante-vocal [CV] resultantes de la combinación de 23 consonantes con cada una de las 5 vocales, es decir un total de 115 transiciones o difonemas. La creación de este inventario a partir de voz natural ha sido la base para la síntesis posterior. Pero el mismo sistema de síntesis ha servido para seguir definiendo estas transiciones gracias a la percepción auditiva y a las sucesivas modificaciones.

El estudio acústico de la voz natural se ha hecho por análisis espectral (FFT) limitándose a una dinámica de 45 dB y una banda pasante de frecuencias entre 0 y 4 KHz con una resolución temporal mínima de 5 ms. Los datos se han completado con informaciones de la fonética castellana ya existentes previamente (NAVARRO TOMAS, 1982) (RODRIGUEZ, 1984 a) (MARTINEZ CELDRAN, 1984).

El conversor texto-voz se ha diseñado de tal forma que pueda ser controlado por un microordenador personal de bajo coste, que permita una entrada de texto a ritmo mecanográfico y una audición inmediata. El inventario de vocales y difonemas debidamente codificados ocupa 8,5 Kbyte de memoria y el programa que realiza todo el proceso tiene una extensión de 14,5 Kbyte. Este último está escrito en lenguaje Pascal.

2. EL PROCESO BASICO

El sistema está pensado de tal forma que a partir del texto ortográfico, escrito por teclado, se llegue al sonido sintetizado encadenando difonemas directos [CV] o inversos [VC] extraidos, estos últimos, de sus correspondientes directos por una simple

inversión temporal. La experiencia ha demostrado que el resultado de esta inversión resulta siempre inteligible, aunque en algunos casos se pierde una cierta calidad de voz, como sucede en el caso de las transiciones con consonantes oclusivas.

Según esto las sílabas constituidas simplemente por una pareja [CV] o [VC] se obtienen directamente del inventario. También son inmediatas las sílabas que constan de una sola vocal que forman parte ya del conjunto de unidades fonéticas. Las sílabas que tienen más de una consonante se construyen a partir del núcleo [CV] o [VC] correspondiente, añadiendo las consonantes iniciales o finales necesarias, por simple yuxtaposición de fragmentos consonánticos extraidos de otros difonemas debidamente truncados.

El inventario básico está constituido por las vocales [a, e, i, o, u] y las transiciones de 23 consonantes con estas vocales. Concretamente son los sonidos [p, t, k, b, d, g, f, s, θ, x, c, m, n, ɲ, l, ʎ, r, rr, β, ð, ɣ, j, w]. De momento no se han considerado otros alófonos menos frecuentes del castellano.

3. EL SINTETIZADOR MEA 8000

La síntesis de los sonidos se realiza mediante el circuito integrado MEA 8000 de Philips, que de por si es ya un sistema completo de síntesis de voz por formantes, controlable por un ordenador para una síntesis continua del habla. Consta fundamentalmente (Fig. 1) de dos generadores: uno de tono periódico y otro de ruido aleatorio que no actuan simultáneamente y que excitan un sistema de cuatro filtros de banda pasante en cascada. Estos filtros pueden ser controlados respecto a su frecuencia central (F1, F2, F3) y a su banda pasante (B1, B2, B3, B4) al ritmo de los movimientos articulatorios del habla humana. La cuarta resonancia está fijada siempre a 3500 Hz como corresponde a una voz masculina normal y únicamente se puede controlar su banda (B4). El sistema dispone también de un control de volumen (AMPL) y de una variación del tono (INCREM. DE TONO) para dar la entonación.

Todos los controles del sintetizador se pueden modificar cada 8 ms, de forma que, cuando interesen variaciones no muy bruscas, las fluctuaciones no presenten prácticamente discontinuidad. Después de un proceso de interpolación y un conversor digital-analógico, la señal está en condiciones de ser aplicada a un amplificador y al altavoz.

4. CREACION DEL INVENTARIO DE DIFONEMAS

A partir de voz natural analizada por FFT se han estudiado los rasgos básicos que definen las vocales castellanas y las transiciones [CV]. Con esto y la información disponible se ha preparado una primera versión de los difonemas castellanos, que podemos calificar de provisional. Se ha diseñado un sistema editor de voz a partir del mismo sintetizador MEA 8000 que permite una entrada ágil de los parámetros acústicos cada 8 ms para una audición inmediata de cada transición. El método se basa fundamentalmete en una representación gráfica, sobre la pantalla de un terminal, de la evolución de los formantes (o

resonáncias) a lo largo del tiempo; con unos controles simultáneos de tono y de nivel. Las modificaciones y la audición se pueden realizar de una forma inmediata, lo que permite una revisión (siempre muy enriquecidora) de los datos obtenidos por análisis.

Para cada difonema [CV] se ha marcado el punto divisorio entre la consonante y la vocal mediante un puntero que permitirá el truncado de la consonante cuando convenga. Un segundo puntero marca también la parte estacionaria de la vocal, donde habrá que intercalar los microfonemas (unidades temporales mínimas de 8 ms) necesarios para marcar la acentuación de las sílabas tónicas.

Este inventario de difonemas deberá ser mejorado posteriormente con la aportación de expertos en fonética castellana. Todas las mejoras pueden ser introducidas simplemente por la modificación de un fichero que contiene toda la información de los difonemas utilizados. Este proceso no afecta para nada al programa principal.

5. ALGORITMO DE SINTESIS

En la Fig. 2 se puede ver un esquema del proceso de síntesis SINCAS para la generación de voz por palabras aisladas o yuxtapuestas. Después de la entrada de texto, que puede ser introducido manualmente por teclado o leído de ficheros de texto, préviamente guardados en un disco, se realiza un procesado inicial de los signos ortográficos (acento, diéresis y signos de puntuación) que afectan normalmente a la prosodia y que serán tratados al final del proceso.

Todas las reglas fonéticas y gramaticales pueden tener sus excepciones; por tanto habrá que prever un inventario de todos aquellos casos que no se amoldan a las previsiones del sistema. Se trata simplemente de un fichero que tiene simultáneamente las entradas ortográficas de estas excepciones y la correspondiente codificación en el lenguaje del sintetizador utilizado. La verificación de excepciones ha de ser siempre previa a todo el proceso para ahorrárselo cuando sea el caso. Este fichero siempre puede ser ampliado o modificado sin necesidad de alterar el programa.

El paso siguiente consiste en la obtención de la sílaba tónica, tanto en el caso de las palabras con acento gráfico como en el caso de las palabras que no lo llevan. El problema se resuelve por una sencilla aplicación de las reglas de la acentuación.

Inmediatamente se procede a la transcripción fonética que convierte los caracteres ortográficos en los correspondientes sonidos acústicos representados por símbolos fonéticos. La transcripción no es inmediata y responde a unas reglas determinadas que tienen en cuenta el marco concreto en que se utiliza cada carácter para deducir su transcripción. El proceso se aplica a cada palabra considerando los caracteres de izquierda a derecha y consultando todas las reglas que pudieran afectarlos. Estas reglas están inventariadas en un fichero exterior al programa que puede ser actualizado en cualquier momento.

A partir de la cadena de caracteres fonéticos se procede a la descomposición silábica, mediante unas reglas que se aplican

de forma sistemática y sin excepciones. La obtención de las sílabas resulta indispensable para determinar los difonemas a utilizar en su construcción, ya sean difonemas directos [CV] o invertidos [VC] y las restantes consonantes que se obtendrán de difonemas truncados. Todos estos elementos se encuentran ya ordenados y codificados según el código del MEA 8000 en el fichero externo de difonemas igualmente modificable por sucesivas mejoras.

El último paso consiste en un tratamiento prosódico mínimo que marca la acentuación y las pausas entre palabras. La acentuación se realiza por una prolongación sistemática de la vocal tónica con un incremento de tono durante la misma. Las pausas se marcan por un descenso de la entonación a partir de la última vocal tónica y un silencio más o menos prolongado antes de la palabra siguiente. Todo ello traducido a los códigos adecuados es la información que se envia al sintetizador que construye la señal eléctrica que genera la onda acústica deseada a través de un amplificador y del altavoz.

6. VERIFICACION

Aunque el sistema se encuentra aún en una fase de estudio, se ha realizado ya una primera prueba de inteligibilidad con un grupo de 47 estudiantes de Ingenieria Técnica de Telecomunicación que oían por primera vez el sintetizador de voz en castellano. El test se hizo a nivel de palabras y a nivel de frases. Las correspondientes listas fueron elegidas entre las que ofrece el Dr. Jorge Perelló en su traducción y adaptación del Manual de Logopedia (PIALOUX, 1978). Se trata de una lista de 25 palabras y otra de 10 frases, consideradas por el autor como fonéticamente equilibradas, es decir, representativas del castellano hablado normal.

Los resultados estadísticos obtenidos (Fig. 3 y 4) son los siguientes:

Test de papabras: con una media de 18,6 aciertos (sobre 25) y una desviación típica de 1,6; lo que representa un 74,5 % de inteligibilidad.

Test de frases: con una media de 9,6 aciertos (sobre 10) y una desviación típica de 0,6; lo que representa un 96,0 % de inteligibilidad.

Hay que advertir que la corrección del test se hizo bajo un criterio absolutamente riguroso, según el cual se rechazó como incorrecta cualquier palabra interpretada con alguna diferencia respecto al modelo, por pequeña que fuera. Igualmente se rechazaron las frases incorrectas, aunque fuera por una sola palabra.

7. CONCLUSIONES

Somos muy conscientes del carácter previo de este trabajo con todas sus limitaciones. La definición inicial de las transiciones [CV] queda abierta a nuevas aportaciones que puedan mejorar la calidad de la síntesis.

El sintetizador utilizado presenta también sus limitaciones que impiden una alta calidad de los resultados. En este sentido hay que destacar la falta de resonancias nasales y de antiresonancias, así como también la imposibilidad de simultanear la fuente periódica y la de ruido aleatorio para una correcta generación de consonantes fricativas sonoras. La limitación de la banda pasante a los 4 KHz no permite superar la calidad de una voz telefónica.

Con todo, pensamos que los objetivos del trabajo se han conseguido satisfactiriamente en cuanto al diseño de un conversor texto-voz en castellano sin limitaciones de vocabulario, con un reducido inventario de unidades fonéticas y unas sencillas reglas de concatenación. La velocidad del proceso y la inteligibilidad conseguida, resultan muy aceptables teniendo en cuenta la reducida extensión del programa i de la memória necesaria para su desarrollo. Esta economia facilita la implantación del sistema como una erramienta paralela a cualaquier terminal de ordenador que podria dar todas sus informaciones en forma no solamente visual, sino también acústica.

BIBLIOGRAFIA

ATAL, B.S.; HANAUER, Suzanne L. (1971)
"Speech analysis and synthesis by linear prediction of the speech wave". JASA 50, pp 637-655.

COKER, C.H. (1968)
"Speech synthesis with a parametric articulatori model". Speech Symposium. Kyoto. Paper A-4.

DUNN, H.K. (1950)
"The calculation of vowel resonances, and electrical vocal tract". JASA 22, pp. 740-753.

FANT, G.; RISBERG, A.; STEVENS, K. N. (1963)
"Evaluation of various analysis-synthesis speech systems". JASA 35, p. 804.

HOLMES, J. N. (1980)
"The JSRU chanel vocoder". Proc. Institute of Electrical Engineers, 127 (F1), pp. 53-60.

ITAKURA, F.; SAITO, S. (1972)
"On the optimum-quantization of feature parameters in the parcor synthesizer". Proc. Conf. Speech Commun. Process, pp 434-437.

MARKEL, J. D.; GRAY, A. H. (1982)
"Linear prediction of speech" Ed. Springer-Verlag. Third printing.

MARTI, J. (1986 a)
"Estudi acústic del català i síntesi automàtica per ordinador". Tesi doctoral. Universidad de Valencia. Facultad de Ciencias Físicas.

MARTI, J. (1986 b)
"Sincat, el sintetitzador català de veu". Quaderns Tècnics, 7, pp. 13-19.

MARTINEZ CELDRAN, E. (1984)
"Fonética". Ed. Teide. Barcelona.

NAVARRO TOMAS, T. (1982)
"Manual de pronunciación española". CSIC. Publicaciones de la Revista de Filologia Española.

PETERSEN, T. L.; BOLL, S. F. (1983)
"Critical band analysis-synthesis. ASSP-31, pp. 656-663.

PIALOUX, P.; VALTAT, M.; FREYSS, G.; LEGENT, F. (1978)
"Manual de Logopedia". Traducido y adaptado por PERELLO, J. Ed. Toray - Masson, S. A. Barcelona.

RODRIGEZ, M.; OLABE, J. C.; SANTOS, A.; MUÑOZ, P.; VILLASECA, I.; MUÑOZ, E. (1984 a)
"Visión panorámica de la respuesta oral de máquinas". Mundo Electrónico, 144, pp. 57-66.

RODRIGUEZ, M.; IGLESIAS, E.; MARTINEZ, R.; MUÑOZ, E. (1984 b)
"Alternativas para síntesis de voz". Mundo Electrónico, 144, pp. 67-79.

SCHROEDER, M. R. (1966)
"Vocoders: analysis and synthesis of speech". Proc. IEEE, vol. 54, pp. 720-734.

STEVENS, K. N.; KASOWSKI, S.; FANT, G. (1953)
"An electrical analog of the vocal tract". JASA, 24, pp. 734-742.

TORRES, B.; VIDAL, E. (1982)
"Sistema de síntesis automática del castellano hablado". V Congreso de Informática y Automática. Madrid. pp. 405-409.
TORRES, B.; VIDAL, E. (1983)
"Synthèse par diphonèmes des phrases éspagnoles". 11e' Congrès International d'Acoustique. vol. 4, p. 195.

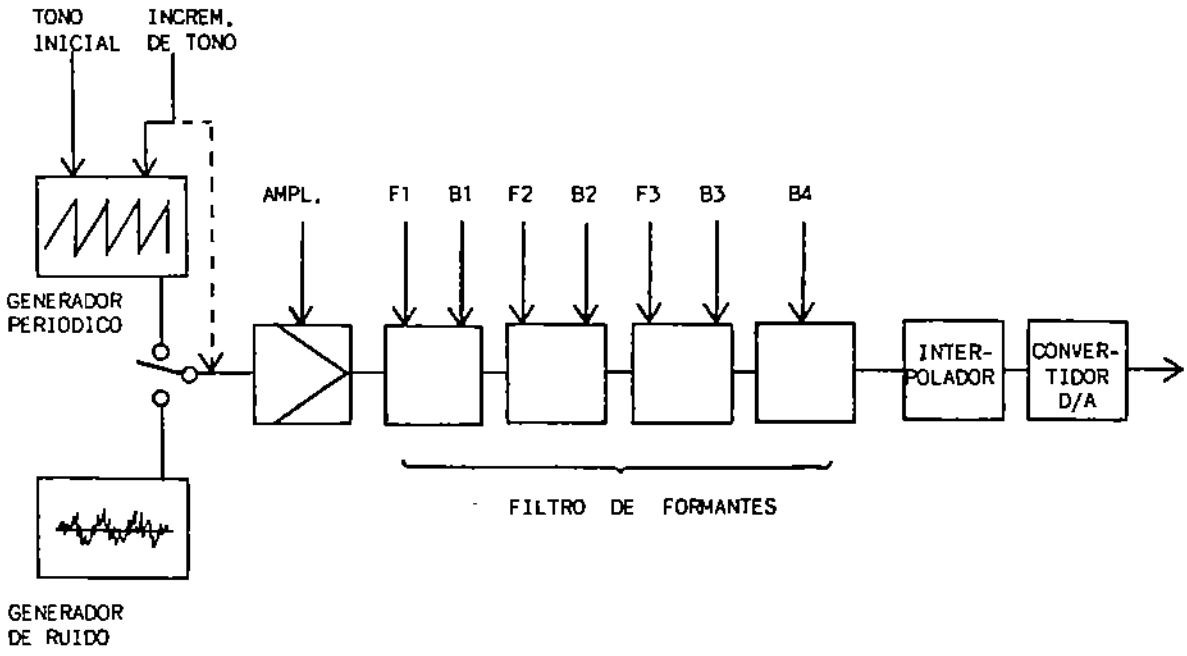

FIG. 1. DIAGRAMA DE BLOQUES DEL CIRCUITO MEA 8000

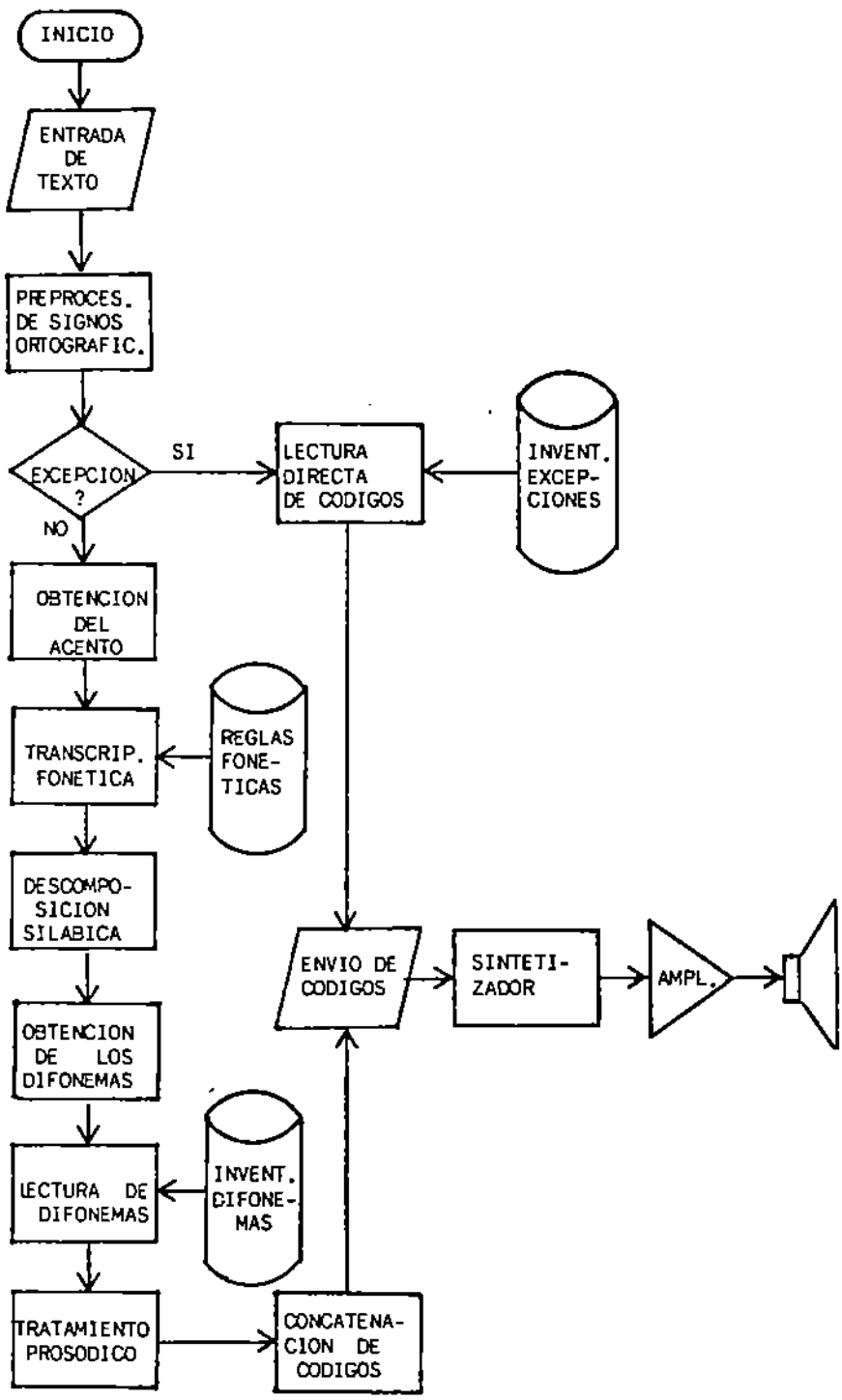

FIG. 2. DIAGRAMA DE BLOQUES DEL ALGORITMO DE SINTESIS.

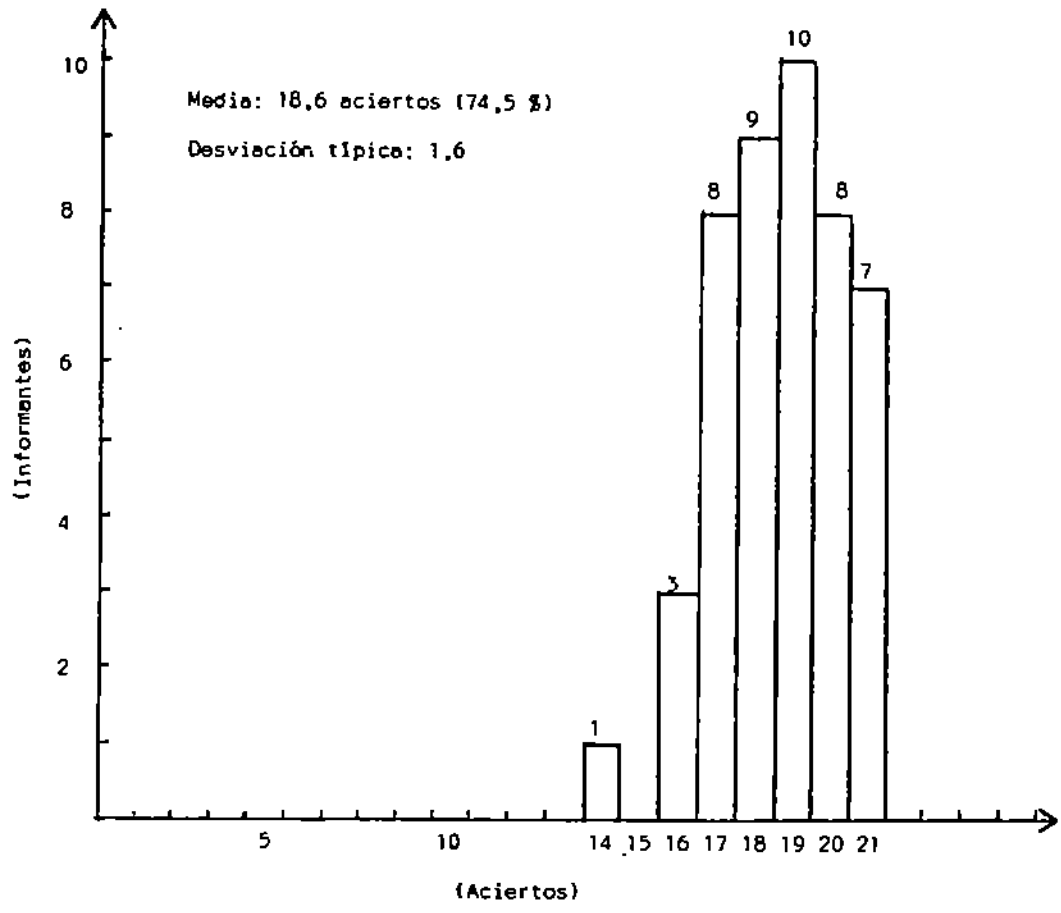

Fig. 3 Resultados del test de percepción de palabras

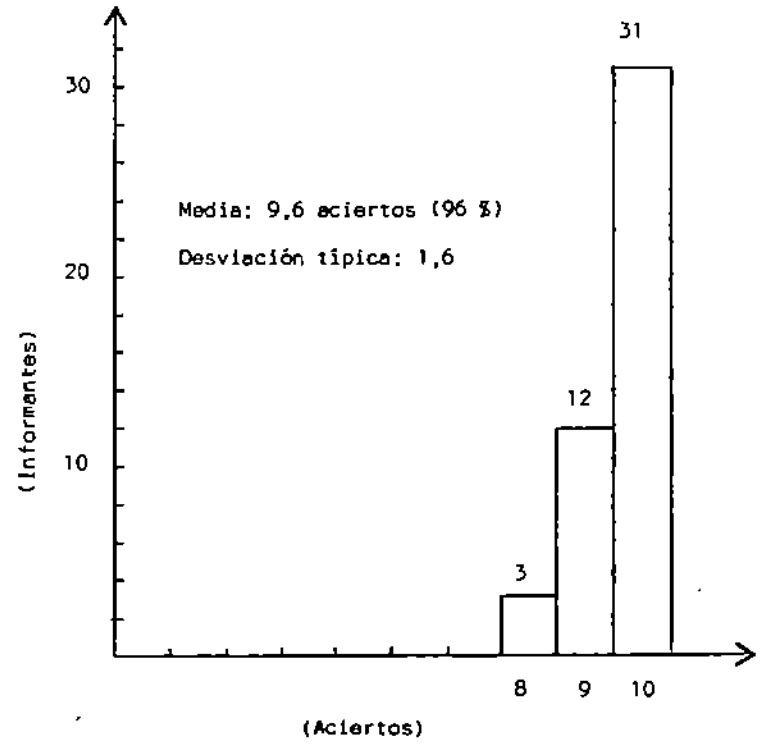

Fig. 4 Resultados del test de percepción de frases