Codificación bidimensional de patrones vocales mediante un esquema de reducción dimensional basado en redes neuronales

Alejandro Bassi A.

Universidad de Chile
Facultad de Ciencias Físicas y Matemáticas
Departamento de Ciencias de la Computación
Av. Blanco Encalada 2120, Santiago, Chile.
e-mail: abassi@dcc.uchile.cl

Resumen
Una de las maneras más aceptadas de describir las vocales del habla humana es a través del "triángulo" articulatorio, que define un espacio abstracto donde cada vocal queda claramente diferenciada por su localización en función de dos rasgos: apertura y frontalidad. Una representación similar al "triángulo" articulatorio se puede obtener a partir del espectro en frecuencia de las vocales, tomando como coordenadas las posiciones de los dos primeros formantes. Sin embargo, la realización automática de una proyección robusta de esta naturaleza es difícil de lograr.

En el presente artículo se propone una técnica no supervisada basada en redes neuronales con entrenamiento autoasociativo que contribuye a resolver este problema, para el reconocimiento y visualización de vocales y diptongos.

Palabras claves: inteligencia artificial, interfaces hombre-máquina.

1 Introducción

Un patrón puede definirse como un vector de características que describe una realidad perceptible. En situaciones prácticas, las ocurrencias de un patrón presentan una gran variabilidad, y el principal problema para su reconocimiento automático es focalizar el análisis, determinando qué diferencias son relevantes para separar las distintas clases perceptuales. Es conveniente entonces elegir una representación que simplifique esta tarea.

En el caso del habla humana, existe una representación abstracta muy económica para describir las vocales según su articulación, es decir, según la manera de pronunciarlas (Crystal, 1987). Ésta consiste en proyectar cada vocal dentro de un espacio bidimensional, donde una coordenada corresponde a la apertura (que tan cerca o lejos del paladar se halla la lengua), y

la otra especifica la frontalidad (que tan cerca o lejos del extremo delantero del tracto vocal se halla la constricción). Esta proyección se suele graficar como se muestra en la figura 1.

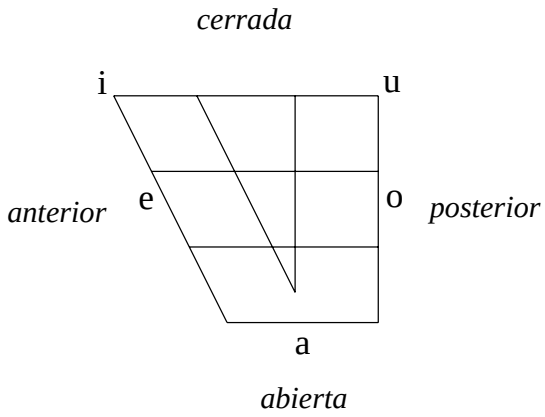

Figura 1: Las vocales del castellano en el "triángulo" articulatorio

Para el procesamiento automático de la voz, no se puede utilizar directamente una representación fonética articulatoria, ya que sólo se tiene acceso al sonido y no a los parámetros de la articulación. Sin embargo, existe una correlación directa entre las posiciones de los formantes de una vocal y su ubicación en el espacio articulatorio (Rabiner & Juang, 1993). Los formantes son zonas de alta energía en el espectro en frecuencia de la señal, que corresponden a resonancias del tracto vocal. El primer formante (F1) es alto para vocales abiertas y bajo para vocales cerradas, el segundo formante (F2) es alto para vocales anteriores y bajo para vocales posteriores (figura 2).

Existen métodos conocidos para determinar las posiciones de los formantes (por ejemplo, a partir de los polos de un modelo LPC). Sin embargo presentan problemas de robustez (en particular, la confusión de formantes). Una manera de enfrentar esta dificultad es a través de un esquema de reducción dimensional que, a partir de una representación del espectro en frecuencia de la señal analizada, sea capaz de producir una proyección bidimensional comparable al "triángulo" articulatorio. Para este propósito, se propone una red neuronal codificadora-decodificadora con entrenamiento autoasociativo que produce una codificación no supervisada.

El artículo está organizado como sigue. En la sección **2** se entrega una breve descripción del proceso de generación de patrones a partir de la señal de voz. En la sección **3** se discuten distintos esquemas de reducción dimensional y sus problemas. La sección **4** describe la topología de la red neuronal propuesta como solución. La sección **5** precisa la manera de entrenar la red. La sección **6** presenta algunos resultados aplicando la técnica propuesta. La sección **7** corresponde a las conclusiones del trabajo.

2 Generación de patrones

El objetivo del procesamiento automático de la voz es reproducir las capacidades perceptuales de un oyente humano. Para garantizar categorizaciones de buena calidad, la representación de los patrones debe ajustarse a este requerimiento. Las más aceptadas consisten en representa-

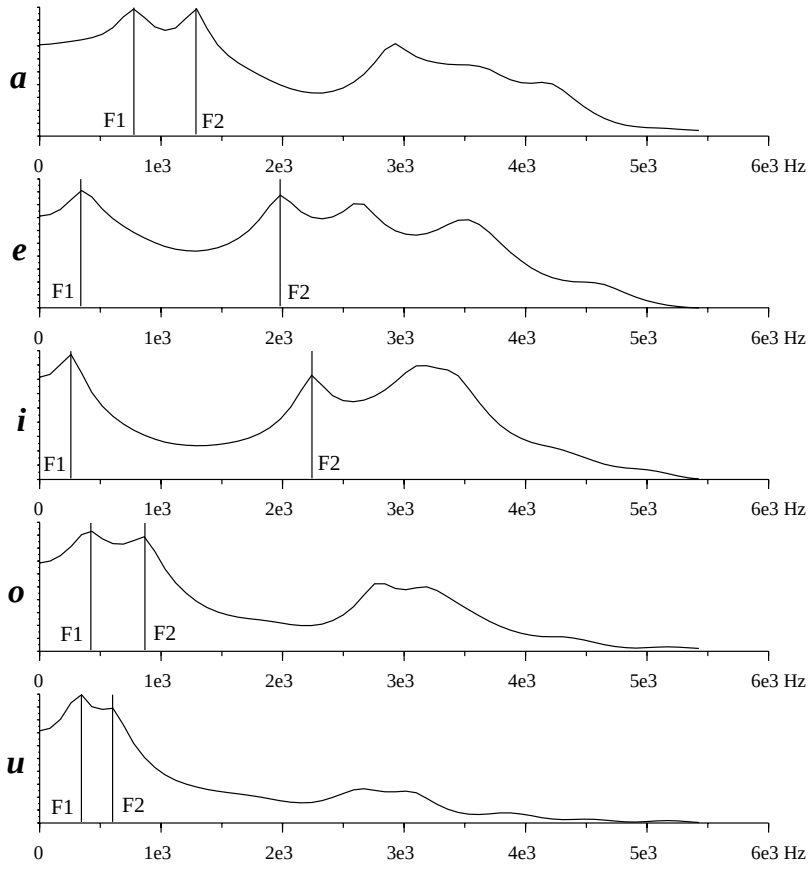

Figura 2: Posición de los formantes de las vocales del castellano para un hablante

ciones del espectro en frecuencia de la señal de voz, preferiblemente en una escala no lineal que corresponde a la respuesta del oído interno humano (Rabiner & Juang, 1993).

En el presente estudio, se trabajó con señales muestreadas a 11.025 Hz, y los patrones se generaron siguiendo la siguiente secuencia de pasos:

- Aplicación de un preénfasis $s_n = 1/2 * (s_{n+1} - s_{n-1})$ a la señal para destacar un rango de frecuencias medio (los métodos tradicionales enfatizan las frecuencias altas, sin embargo los formantes más altos aportan poca información para caracterizar los sonidos vocálicos).
- Segmentación de la señal en ventanas de Hamming de 128 muestras separadas por 64 muestras.

Y, para cada ventana:

- Cálculo de los coeficientes de un LPC de 9 parámetros utilizando el método de autocorrelación (no es necesario que las ventanas de análisis sean de un tamaño potencia de dos, pero el que lo sean permite comparar más fácilmente con otros métodos que sí lo requieren).
- Cálculo del espectro logarítmico de la respuesta del LPC evaluando la función de transferencia del LPC en 64 puntos equiespaciados de la escala mel (la escala mel se define de la siguiente manera: $mel = 1125 * ln(0.0016 * hertz + 1)$).

- Filtraje del espectro (*liftering*) para eliminar variaciones muy suaves o muy abruptas, lo contribuye a reducir la influencia de las condiciones de captura de la voz (en particular del micrófono).
- Normalización de la varianza, lo que otorga un mejor desempeño ante ruido aditivo.

La figura 3 muestra algunos pasos de este proceso para un segmento de voz en particular que corresponde a la vocal /e/.

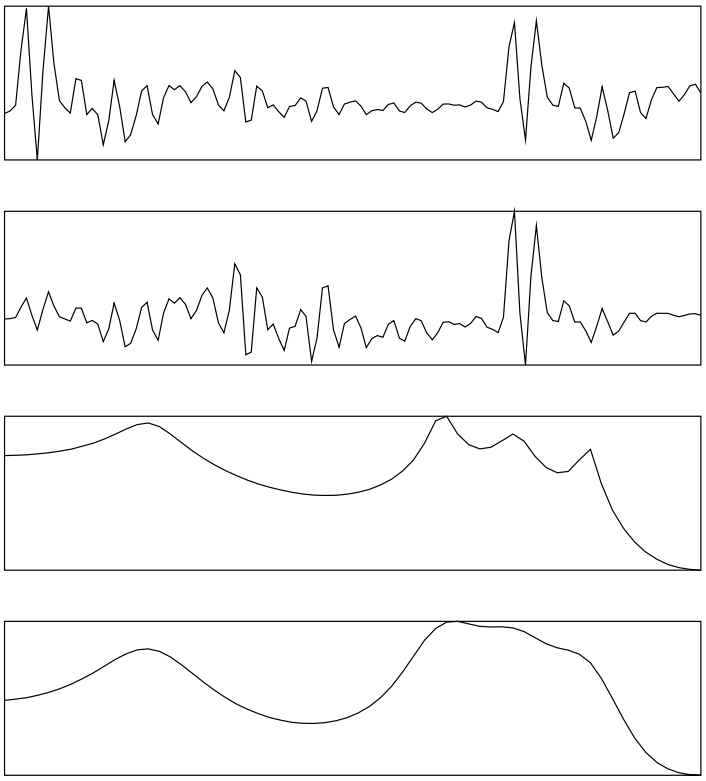

Figura 3: Procesamiento de la señal: señal preenfatizada, ventana de Hamming, espectro logarítmico en escala mel y espectro filtrado

Los patrones obtenidos son vectores de 64 componentes que constituyen una representación, en la escala mel, de la envolvente del espectro logarítmico del segmento de señal analizado. Puede observarse que esta representación está sobremuestreada. En la literatura se recomiendan generalmente representaciones más resumidas e indirectas del espectro, y las preferidas se basan en los primeros coeficientes del cepstrum (Rabiner & Juang, 1993). Sin embargo, la representación propuesta aquí no está destinada a ser utilizada tal cual, sino sólo después de haber sido sometida a un proceso de reducción dimensional. El sobremuestreo ayuda a obtener una reducción dimensional de mejor calidad.

3 Reducción dimensional lineal y no lineal

Una de las técnicas básicas del reconocimiento de patrones consiste en aplicar transformaciones con el objetivo que las características de los vectores transformados no estén correlacionadas, lo que elimina redundancias (Theodiridis & Koutroumpas, 1999). Además, si se descartan aquellas características menos significativas, se obtiene una representación de menor dimensión que la original, pero conservando la mayor parte de la información necesaria para discriminar. A pesar de ser muy conocidas, no existe mucha información sobre el uso de estas técnicas para procesamiento de voz (Pijpers, Alder, & Togneri, 1993).

Entre las transformaciones lineales que cumplen estos requerimientos, se destaca la de Karhunen-Loève (KL). La transformación KL toma como base a los vectores propios de la matriz de correlación de los patrones analizados y, como índice de relevancia, a los valores propios. Para obtener una reducción dimensional, se ordenan las características según los valores propios y se toman aquellas que tengan los valores más altos. Esto constituye exactamente un análisis de componentes principales. Como ejemplo, la figura 4 muestra una proyección en dos dimensiones para el conjunto de patrones de las cinco vocales del castellano de un hablante particular.

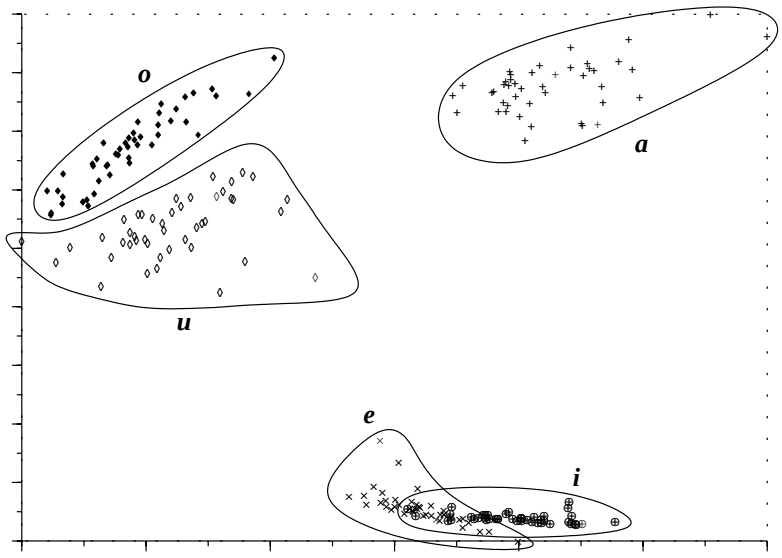

Figura 4: Reducción dimensional lineal para las vocales del castellano

Puede observarse que la proyección lineal no permite diferenciar bien los fonemas /i/ y /e/. Se puede demostrar que este tipo de proyección es equivalente al que se obtendría de una red neuronal de topología n-2-n con función de activación lineal, entrenada en modo autoasociativo (la salida reproduce la entrada). Esto sugiere que es posible mejorar los resultados utilizando redes más complejas con activación no lineal (Bishop, 1995). La figura 5 muestra el resultado aplicando una red de topología n-16-2-12-n con activación sigmoide, excepto la última capa lineal. Una red para reducción dimensional se puede ver como un codificador (en el ejemplo n-16-2) seguido de un decodificador (2-16-n). Entrenar la red consiste en reducir el error de reproducción, lo que implica optimizar la codificación.

Se observa que la no linealidad mejora la codificación. Sin embargo, se verificó que ninguno de estos métodos es capaz de producir buenos resultados cuando se trata de predecir la codificación

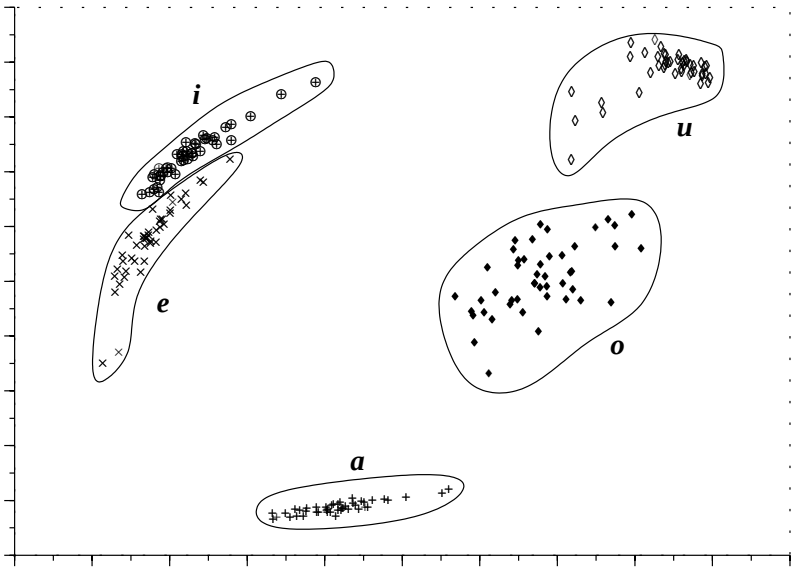

Figura 5: Reducción dimensional no lineal para las vocales del castellano

de un patrón intermedio, lo cual es necesario para diptongos. El problema se produce porque una posición intermedia entre dos patrones se obtiene normalmente dando valores intermedios a cada componente, cuando lo que conviene hacer aquí es colocar los formantes en posiciones intermedias (figura 6).

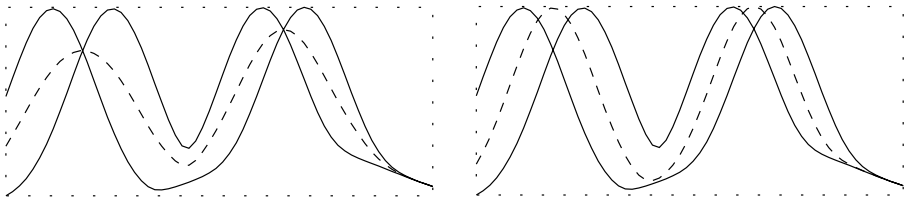

Figura 6: Dos intermedios posibles, el de la derecha es el más adecuado

4 Topología de la red utilizada

La diferencia esencial entre una red con unidades de activación sigmoide y la red propuesta aquí radica en el decodificador. Para mejorar la capacidad de interpolación, se optó por una codificación basada en una composición de cuatro gaussianas y un ajuste de nivel. Cada gaussiana está definida por su posición en el eje horizontal (que corresponde a la posición de un formante), su desviación estándar y su amplitud. El ajuste de nivel está definido por dos parámetros, uno para cada extremo del espectro. En total hay 14 parámetros y_k, que se generan a partir de los dos parámetros (x_1 y x_2) de la proyección bidimensional. Esta generación se hace de la siguiente manera:

$$y_k = w_{k,1}x_1 + w_{k,2}x_2 + w_{k,3}x_1^2 + w_{k,4}x_1x_2 + w_{k,5}x_2^2$$

El codificador mantiene una estructura clásica n-16-2. Sin embargo, la nueva red se puede utilizar de manera de optimizar la codificación bidimensional. Esto consiste en no tomar directamente las predicciones producidas por el codificador, sino aplicar antes un proceso de corrección mediante un descenso del gradiente del error de la decodificación. La figura 7 ilustra el grado de aproximación al cual se puede llegar con este método.

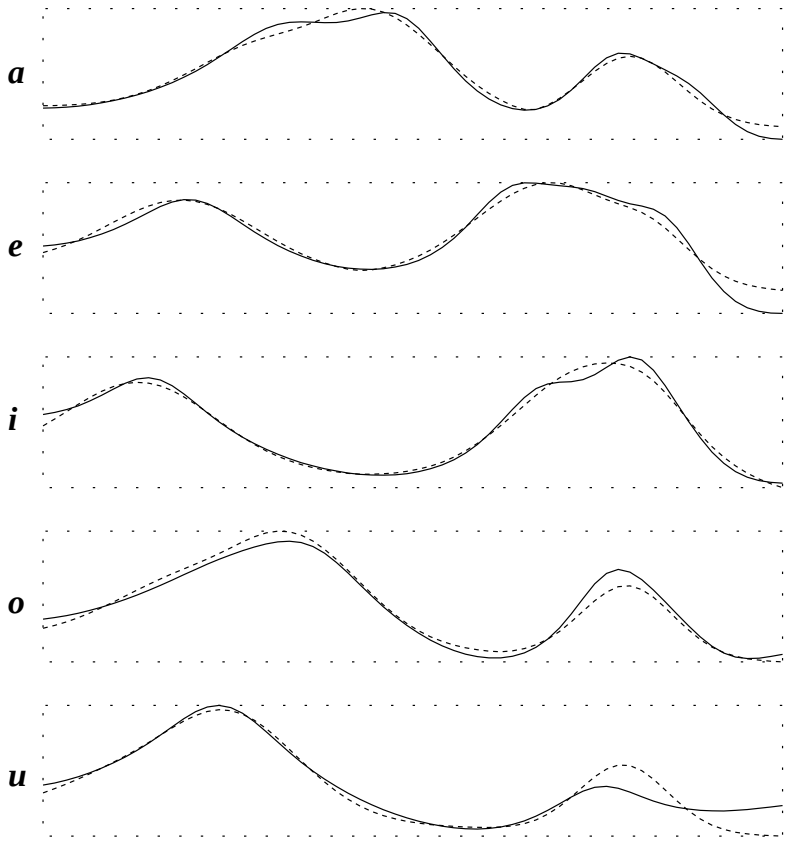

Figura 7: Patrón original y su aproximación gaussiana para las vocales

5 Método de entrenamiento de la red

Las redes multicapas, como la propuesta en este trabajo, tienen la reputación de ser difíciles de entrenar, lo que puede explicar que no se las utilice más a menudo. Sin embargo, cuando se tiene suficiente conocimiento del problema abordado, como en este caso, se pueden inicializar los parámetros de la red de manera que el entrenamiento corresponda más un ajuste que a un aprendizaje a partir de cero. La inicialización consiste en hacer que a cada entrada del decodificador le corresponda como salida una versión idealizada del espectro que se debería obtener para un sonido ubicado en esa región del "triángulo" articulatorio. Para esto se controla la posición de las dos primeras gaussianas, que representan a los dos primeros formantes. Además, se dejan fijos los parámetros que definen estas posiciones (no se entrenan). Bajo estas condiciones, el entrenamiento a probado ser estable y sin problemas de convergencia.

El proceso de entrenamiento es como sigue:

- Dado un conjunto de patrones, se identifica, para cada patrón, qué codificación entrega el resultado más parecido. Para ello, se toma un muestreo discreto de las posibles codificaciones y se comparan los espectros utilizando la suma de los valores absolutos de las diferencias entre componentes como medida de distorsión.
- Se entrena el decodificador, tomando como entrada a la codificación de la fase previa, y, como salida, a los patrones.
- Se entrena el codificador, tomando como entrada a los patrones, y, como salida, a la codificación aproximada de la primera fase.
- Se junta el codificador y el decodificador, ambos entrenados, para formar una red completa.
- Se entrena la red completa, tomando como entrada a los patrones, y, como salida, a los mismos patrones (entrenamiento autoasociativo).

El entrenamiento previo del codificador y del decodificador permite una convergencia más rápida y segura. El primer paso del entrenamiento aprovecha la inicialización del decodificador. A pesar de que en estricto rigor el entrenamiento autoasociativo es supervisado, la codificación queda definida sin supervisión directa (se puede considerar que la inicialización del decodificador es una especie de supervisión indirecta).

6 Resultados experimentales

La técnica no se ha probado aún de manera extensiva. Sin embargo se han hecho varios experimentos con voces muy distintas (hombres, mujeres y niños) para verificar la robustez del método. A continuación, se muestran las proyecciones bidimensionales de las vocales y las transiciones de formantes de los diptongos /ai/, /au/, /ui/, para dos voces (un hombre adulto y un niño). Además se presenta el caso de las transiciones de formantes de la vocal /a/ en el contexto de las consonantes oclusivas sonoras /b/, /d/ y /g/. Se observa la caída del primer formante producto de la oclusión y las diferentes trayectorias del segundo formante según el punto de articulación de la consonante correspondiente (cae para /b/, se mantiene para /d/ y sube para /g/). El entrenamiento se hizo con una muestra de las vocales puras (alrededor de 200 patrones para cada caso). Las señales de voz se capturaron sin precauciones particulares en un computador personal.

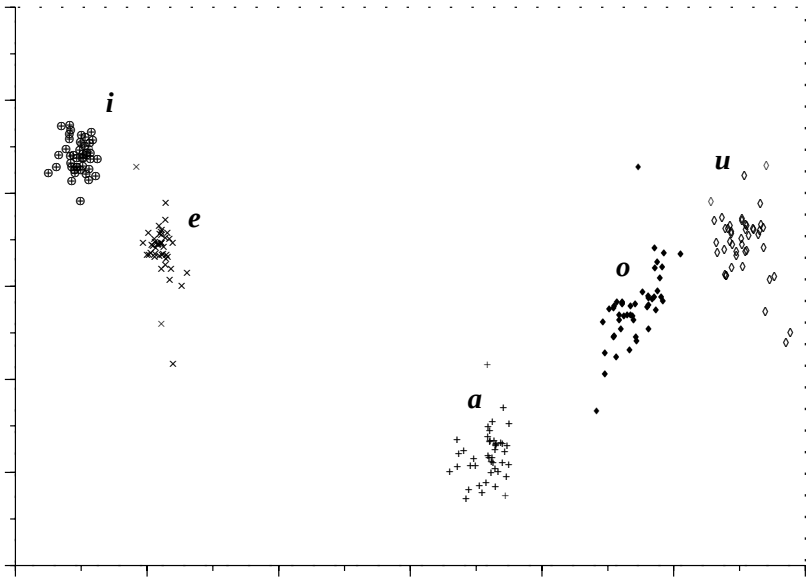

Figura 8: Proyección bidimensional de las vocales de un hablante adulto masculino

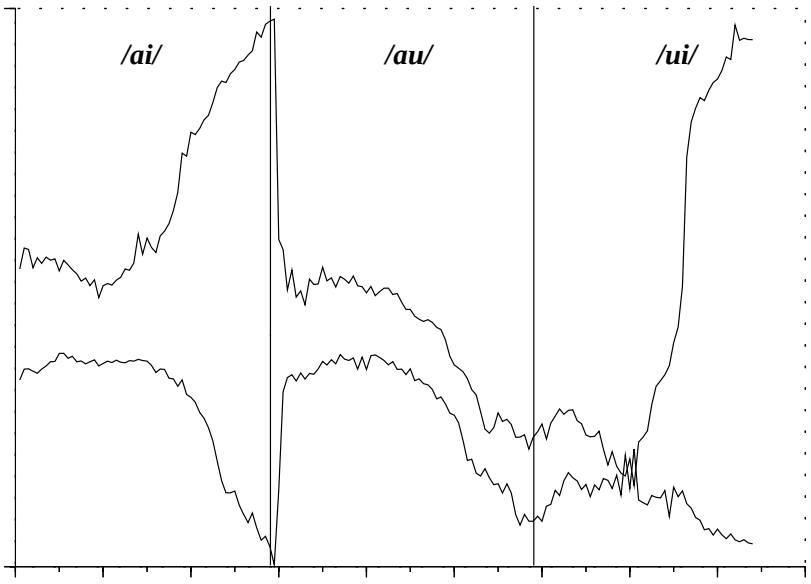

Figura 9: Transiciones de los formantes para los diptongos /ai/, /au/, /ui/ de un hablante adulto masculino (red entrenada sólo con vocales)

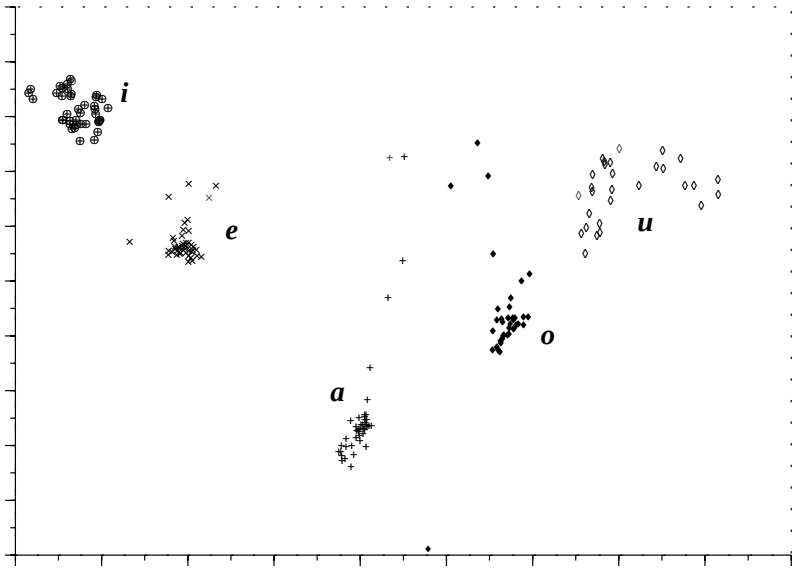

Figura 10: Proyección bidimensional de las vocales de un niño

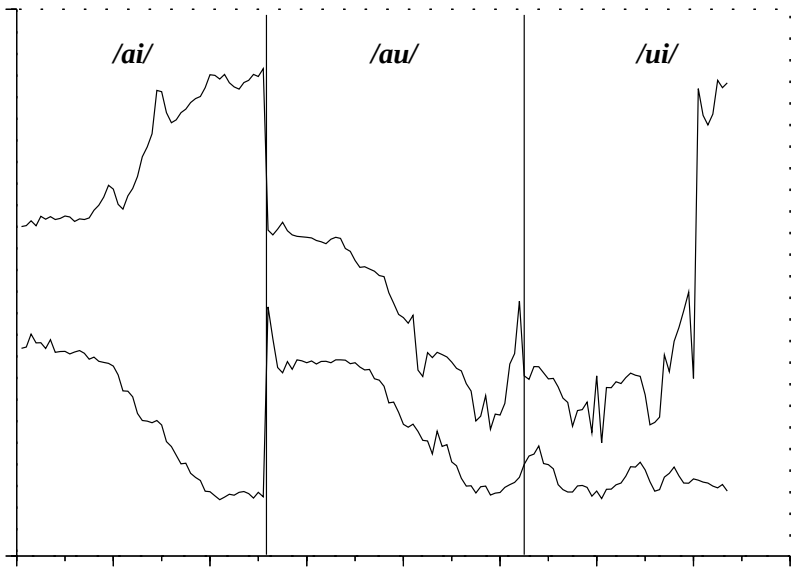

Figura 11: Transiciones de los formantes para los diptongos /ai/, /au/, /ui/ de un niño (red entrenada sólo con vocales)

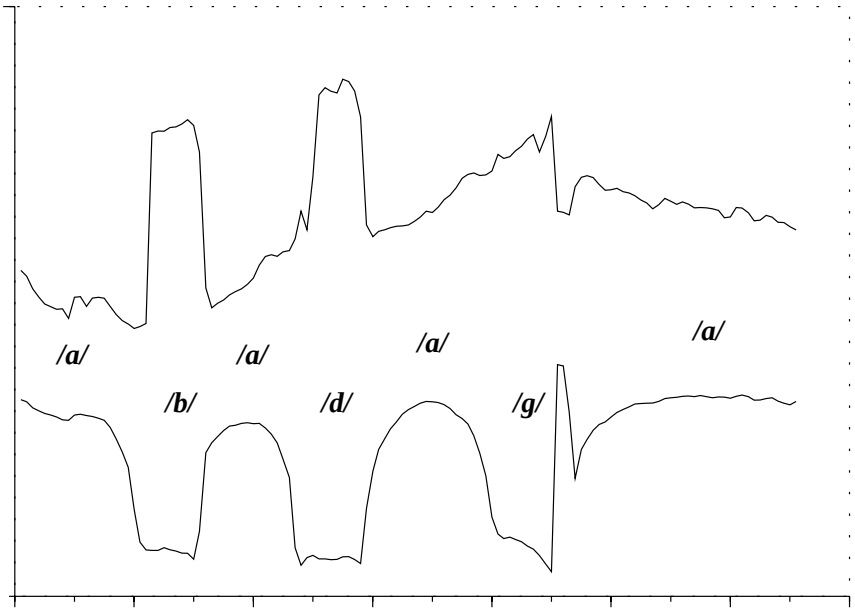

Figura 12: Transiciones de los formantes para la vocal /a/ de un hablante adulto masculino en el contexto de las consonantes /b/, /d/ y /g/ (red entrenada sólo con vocales)

7 Conclusiones

La técnica propuesta para reducción dimensional permite una caracterización abstracta de la señal de voz procesada. Los experimentos preliminares llevados a cabo muestran que el método de entrenamiento de la red codificadora-decodificadora es robusto y que sus resultados son mejores que los de reducción dimensional lineal y redes neuronales de activación sigmoide, especialmente para modelar transiciones. El entrenamiento se puede hacer con un conjunto reducido de patrones de ejemplo, lográndose buenas generalizaciones.

Esta técnica de codificación se puede utilizar como una primera fase de un proceso de reconocimiento del habla. También existe la posibilidad de aplicarla como una herramienta de visualización, especialmente como retroacción visual para el aprendizaje del lenguaje hablado en personas con mala audición. Se ha construido un prototipo que demuestra la factibilidad de la visualización en tiempo real (Vera, 1999).

Referencias

Bishop, C. M. (1995). *Neural networks for pattern recognition.* Oxford.

Crystal, D. (1987). *Enciclopedia del lenguaje.* edición española dirigida por J.C. Moreno Cabrera, Taurus.

Pijpers, M., Alder, M., & Togneri, R. (1993). Finding structure in the vowel space. In *Proceedings of the First Australian and New Zealand Conference on Intelligent Information Processing Systems, Perth, Western Australia, 1-3 December.*

Rabiner, L., & Juang, B.-H. (1993). *Fundamentals of speech recognition.* Prentice Hall.

Theodiridis, S., & Koutroumpas, K. (1999). *Pattern recognition.* Academic Press.

Vera, A. (1999). *Visualización de la fonética del lenguaje hablado.* memoria de ingeniero, Universidad de Chile.