Comparación de algoritmos de reconocimiento de vocales

Pablo Aguado, *ELO 23724*

***Resumen*—En este trabajo se presenta un sistema para la evaluación sistemática de algoritmos de reconocimiento de vocales. Se presentan dos algoritmos habituales para la extracción de formantes y se evalúan con el sistema elaborado, sobre una pequeña base de datos de vocales.**

I. Objetivos y actividades

LAS vocales constituyen una parte muy importante en la mayoría de los idiomas, conformando los núcleos silábicos. Cada idioma tiene su propio conjunto vocálico, y el sonido de cada elemento del conjunto varía según el dialecto, la región y características del hablante. Entre otras razones, se las analiza para describirlas fonológicamente (cómo funcionan dentro del lenguaje), fonéticamente (estudiando su sonido y generación), o para su síntesis digital. En este trabajo se utilizan ciertas características físicas del sonido para identificar fonemas vocálicos de un hablante, con la finalidad de ayudar a mejorar su fonación.

En la Sección II se describe mejor el por qué de este trabajo y se presentan algunos conceptos base referidos a las vocales, su modelado y el reconocimiento. Luego en la Sección III se hace una breve reseña del sistema de evaluación desarrollado en Matlab para poder comparar los sistemas de reconocimiento. Posteriormente se presenta la base de datos utilizada. Las Secciones V y VI describen los dos algoritmos de extracción de formantes que se compararon, LPC y Burg. Por último, se muestran los resultados de las comparaciones y se proponen tareas futuras.

Este es el informe del trabajo práctico integrador realizado para la materia Complementos de Electrónica Analógica II: Sonido y Mediciones Acústicas, dictada por los profesores Ernesto Accolti, Arnoldo Fernández y Eduardo Zavalla, durante el primer semestre de 2016, en la carrera de Ingeniería Electrónica de la Facultad de Ingeniería de la Universidad Nacional de San Juan.
El código fuente del sistema, la bases de datos, los scripts de evaluación y tal vez versiones futuras de este trabajo pueden encontrarse en https://github.com/aguadopd/avi_evaluacion_reconocimiento_formantes .

II. Introducción

II-A. Motivación

El interés sobre las formas de reconocer una vocal surge de la propuesta de desarrollo de un dispositivo cuyo objetivo es estimular y corregir la vocalización llevada a cabo por niños con problemas de aprendizaje, particularmente aquellos con hipoacusia congénita. Este dispositivo tendrá distintos elementos de realimentación que responderán ante el habla del niño y lo orientarán a mejorarla.

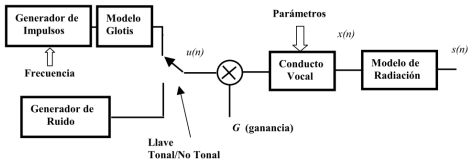


Figura 1. Modelo fuente-filtro de producción de voz. Tomado de [2].

Uno de los factores a intentar corregir es la fonación de vocales. Muchas veces sus métodos de aprendizaje están limitados a la imitación (valga la redundancia, limitada) de la posición de la lengua y los labios, y los resultados obtenidos no son los que deberían; los sonidos emitidos no son los que se corresponden a la vocal deseada y para la zona geográfica en cuestión. Esto es un perjuicio para la comunicación oral con el resto de la sociedad. Entonces, el dispositivo (con ayuda de profesionales fonoaudiólogos) debe informar al usuario cuándo su fonación es correcta para una determinada vocal.

II-B. Vocales

Antonio Quilis (en la Sección 2.6.3 de [1]) define fonéticamente a las vocales (o más ampliamente, a los vocoides) como aquellos sonidos que se articulan con el aparato fonador sin oclusiones. Al emitirlos, "la apertura de los órganos articulatorios es completa, el paso del aire es libre y las cavidades supraglóticas se limitan a modificar el timbre laringeo". A fines prácticos, en este trabajo se busca reconocer a los siguientes sonidos definidos como vocálicos por la *International Phonetics Association*: /a/, /e/, /i/, /o/, /u/, y que corresponden a las vocales del Castellano.

II-B1. Modelo fuente-filtro: Existen diversos modelos utilizados para el análisis y la síntesis vocales. Uno de los más utilizados es el modelo fuente-filtro (*source-filter*, ver Figura 1), que supone que la producción de voz está compuesta por una señal de excitación y un sistema o filtro lineal que la modula para obtener los diversos sonidos del habla humana. Es simple pero útil para representar segmentos (cuasi)uniformes de sonido. Algunas características:

- La señal de excitación es un tren de pulsos para los sonidos tonales y ruido para los no tonales.El tren de pulsos es causado por los ciclos repetidos de apertura

Versión 1.0 - Entregada el 24/06/2016

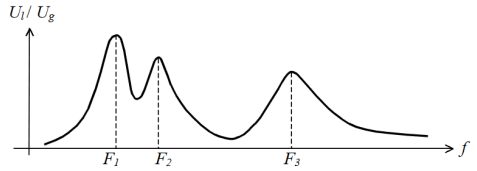


Figura 2. Formantes del tracto vocal para una determinada posición de los órganos articulatorios. Tomado de [3].

y cierre de la glotis. Psicoacústicamente, la frecuencia del tren de pulsos está muy relacionada con la altura de la voz. Esta frecuencia puede ir desde los 80 Hz para varones adultos hasta los 600 Hz en niños.

- El filtro tiene una respuesta frecuencial que se ajusta a la del tracto vocal para la fonación de un sonido determinado. Esta respuesta es la que configura el timbre del sonido.

Un desarrollo más completo puede encontrarse en las Secciones 2.2 y 2.3 de [4].

II-B2. Formantes: La forma más difundida para caracterizar físicamente a los sonidos vocálicos es a través de los formantes, que son los máximos de resonancia del aparato fonador (Figura 2). Debe tenerse en cuenta que para cada vocal el rango de variación de los formantes es amplio, y función de la edad del hablante, su género, el idioma y dialecto que habla e incluso su zona geográfica. En [7] se presentan valores de los formantes F_1 y F_2 para hablantes argentinos de edad adulta.

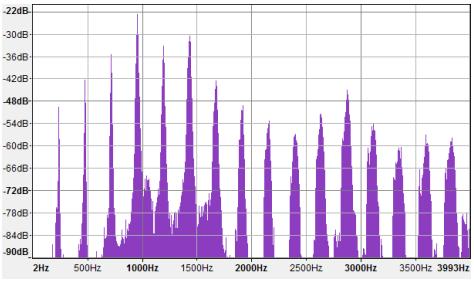
En la Figura 3a se observa el espectro frecuencial correspondiente al sonido de una /a/ emitido por una hablante durante medio segundo. Debido al espaciamiento entre los armónicos del pulso glotal, no es fácil definir qué frecuencias corresponden a las resonancias del tracto vocal. Sin embargo, se puede intuir que los 3 primeros formantes se encuentran alrededor de 1000 Hz, 1500 Hz y 3000 Hz respectivamente. El espectrograma correspondiente está en la Figura 3b; los formantes estimados en el programa Praat están marcados en la Figura 3c y sus valores promedio son 970 Hz, 1473 Hz y 2859 Hz. Nótese que la extracción automática de Formantes no puede reducirse a la búsqueda de máximos en el espectro del sonido; es necesario antes separar los espectros del tren de pulsos glotales y del filtro del tracto vocal.

II-C. Reconocimiento

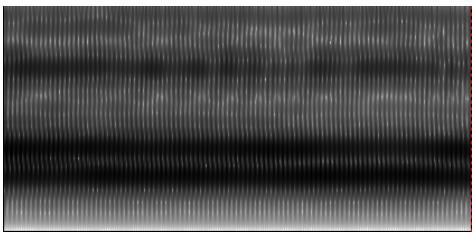
El reconocimiento de un sonido como una vocal comprende la clasificación del sonido en cuestión a partir de un conjunto de características codificadas en él. En este trabajo nos limitaremos a utilizar como características a los dos primeros formantes; de aquí en adelante F_1 y F_2.

Los algoritmos de extracción probados están basados en estimación de modelos lineales y son detallados en las Secciones V y VI; ambos intentan determinar la función de transferencia del modelo lineal del tracto vocal, mencionado en las secciones anteriores.

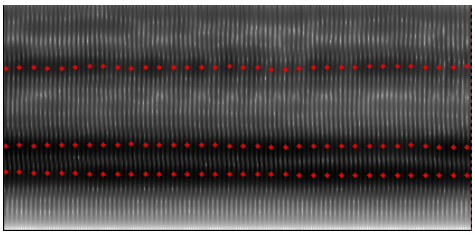
No se implementaron algoritmos clasificadores.



(a) Espectro frecuencial



(b) Espectrograma



(c) Formantes

Figura 3. Información frecuencial correspondiente al sonido de una /a/ femenina

III. Sistema de evaluación

Se desarrolló un sistema de evaluación en Matlab siguiendo el paradigma de programación orientada a objetos, lo que facilita la adición de nuevos algoritmos o bases de datos sobre cual probarlos. Algunas de las clases son:

- Fonema
- BaseDeDatos
- Extractor
- Clasificador
- Resultad
- Evaluación

III-A. Métricas

El sistema permite comparar:

- Conjuntos extractores-clasificadores (de aquí en adelante, algoritmos de reconocimiento) entre sí, al medir la precisión con la que clasifican un conjunto de vocales de la base de datos;
- Algoritmos extractores de formantes entre sí, al medir la distancia entre los formantes estimados y los extraídos manualmente. Los algoritmos extractores son una parte esencial de los algoritmos de reconocimiento, y entonces se hace necesario evaluarlos de forma independiente para una más clara optimización.

Para comparar extractores de formantes usaremos la media cuadrática de las distancias euclídeas entre los vectores de formantes estimados y los marcados manualmente. Para m formantes F de un fonema, se calcula

$$drms = \sqrt{\frac{\sum_{i=1}^{n} {d_i}^2}{n}}$$

donde

$$d_i = \sqrt{\sum_{j=1}^{m} {F_j}^2}$$

Para comparar algoritmos de clasificación la métrica será la proporción del total de fonemas analizados que fue correctamente clasificada.

IV. Bases de datos

No se encontraron disponibles bases de datos de fonemas vocálicos con sus correspondientes formantes estimados manualmente. Se procedió entonces a la caracterización manual del set de sonidos compartido por Juan Carlos Gómez en [6] para la detección de vocales como parte del material de un curso de procesamiento de señales de voz. Le denominaremos base de datos Prodivoz.

Este set de datos está comprendido por 8 archivos PCM *wav* para cada vocal, resultando en 40 archivos. Todos están grabados con una frecuencia de muestreo de 8 kHz y 16 *bits* de profundidad. Para cada vocal hay:

- 3 sonidos largos (> 500 ms), emitidos como sonido único — *Ej: A_Agu_nr, A_Bruno_nr, A_Mari_nr*
- 5 sonidos cortos (< 500 ms), extraídos de una palabra — *Ej: A_Agu_Hablada_nr, A_Bruno_Hablada_nr, A_flaco_nr, a_g, a_p*

Para encontrar los formantes se utilizó el programa Praat [8], que los estima de forma similar a lo detallado en VI. Los parámetros tuvieron que ajustarse manualmente para cada vocal, ya que lo óptimo para las vocales abiertas no era bueno para las vocales cerradas. Este ajuste se basó en inspección visual, asegurándose la coincidencia entre los picos del espectrograma y los formantes marcados sobre él. El trabajo manual es habitual en la determinación de los formantes por parte de quienes realizan estudios fonéticos —ver 4.4.1 en [5].

V. LPC

LPC son las siglas de *Linear Predictive Coding* —Codificación Predictiva Lineal—, un método para codificar señales de voz en función de los parámetros de un filtro lineal cuya respuesta en frecuencia se aproxima a la de cada segmento de sonido. En realidad para el análisis de formantes no es necesario codificar nada, pero se analiza el mismo filtro lineal utilizado en LPC y habitualmente este análisis se denomina de la misma manera o, de manera más correcta, Análisis Lineal Predictivo.

La predicción lineal surge del modelado de sistemas y señales. Consiste en suponer que para un instante determinado una señal, la respuesta de un sistema LTI discreto ante una entrada, puede ser aproximada como una combinación lineal de la entrada actual y las salidas anteriores (a esto también se lo denomina modelo autoregresivo o AR). Se supone además que la señal de entrada es ruido blanco con una varianza determinada. Los parámetros a optimizar en el modelado, entonces, son los coeficientes del filtro y la varianza del ruido de entrada; la optimización consiste en la minimización del error, que es la diferencia entre la señal real y la señal obtenida como salida del filtro.

Uno de los posibles métodos matemáticos de minimización del error de predicción resulta en un conjunto de ecuaciones denominadas ecuaciones de Yule-Walker. Una forma popular y eficiente de resolverlas es a través del algoritmo recursivo de Levinson-Durbin —Resolverlas implica determinar el valor de los coeficientes del filtro.

El orden del filtro es una variable de diseño en el análisis lineal predictivo. Mientras mayor el orden, menor será el error de predicción y la respuesta frecuencial del filtro se aproximará más a la de la señal que se pretende modelar. La señal de voz disponible, y de acuerdo al modelo fuente-filtro (ver Sección II-B), contiene información espectral de la fuente y del tracto vocal; sólo esta última es necesaria para extraer los formantes. Por tanto, un orden muy elevado del filtro predictor tendrá máximos correspondientes a los armónicos de la frecuencia glotal y que interfieren con la estimación de las frecuencias formantes. Lo usual es que el orden del filtro predictor se aproxime al doble (porque tiene polos resonantes de segundo orden) de la cantidad de formantes que se desean encontrar en el rango de frecuencias de la señal.

Una vez encontrado el sistema, se encuentran las raíces de su polinomio característico, que son los polos del sistema. Se eliminan aquellos polos cuyo ancho de banda sea muy chico o muy grande como para representar un formante. Las frecuencias de resonancia de los polos resultantes son las frecuencias formantes (estimadas) del tracto vocal. El procedimiento completo se describe en el Algoritmo 1.

VI. Burg

El método de Burg es otra forma de optimizar los coeficientes predictores del filtro lineal mencionado anteriormente, haciendo diferentes suposiciones para solucionar las ecuaciones de Yule-Walker. Pruebas de otros autores sugieren que es un mejor método de predicción lineal que el tradicional LPC resuelto con el algoritmo de Levinson-Durbin. El método de Burg es el que usa el programa Praat en su estimación de formantes, según se aclara en la sección *Sound: To Formant (burg)* de la ayuda del programa.

La estimación de los formantes prosigue de la misma manera que para LPC.

El capítulo 6 de [4] presenta detalladamente el análisis de predicción lineal.

VII. Resultados

Se evaluaron diversas combinaciones de parámetros para los extractores LPC y Burg sobre la base de datos

Algoritmo 1 Análisis de predicción lineal para la estimación de los formantes F_1 y F_2

Entrada: Sonido digitalizado X, con una frecuencia de muestreo Fs. Orden P del filtro predictor deseado.

Salida: F_1 y F_2

1: [Opcional] Remuestrear X. Praat remuestrea al doble de la frecuencia máxima a la que se espera encontrar un formante.
$Fs \leftarrow Fs$

2: [Opcional] Ventaneo con una ventana W del tipo deseado. Es habitual usar una ventana tipo Hamming; Praat utiliza una gaussiana truncada.
$X \leftarrow X * W$

3: Predicción del filtro lineal. Se obtiene un vector V de longitud P correspondientes a los coeficientes del polinomio mónico característico del filtro LIT discreto.
$V \leftarrow LPC(X, P)$ o $V \leftarrow BURG(X, P)$

4: Se calculan las raíces R del polinomio V, que corresponden a los polos del sistema.
$R \leftarrow RAICES(V)$

5: Se calculan las frecuencias de resonancia FR de cada polo y sus anchos de banda B.
$FR \leftarrow FR(R)$
$B \leftarrow B(R)$

6: Se eliminan los polos de muy baja frecuencia ($FR < 50\ Hz$) y de muy alta frecuencia ($FR > F_{Nyquist} - 50\ Hz$), ya que probablemente son *artifacts* del algoritmo de predicción.
$R \leftarrow filtrado(R, FR)$

7: Se eliminan aquellos polos cuyo ancho de banda sea muy grande como para representar un formante vocálico ($> 400\ Hz$).
$R \leftarrow filtrado(R, B)$

8: Se ordenan los polos según su frecuencia de resonancia, de menor a mayor. Los dos primeros (si existen) corresponderán a F_1 y F_2.
$F \leftarrow ordenar(F, FR)$
$F_1 \leftarrow F[1]$, $F_2 \leftarrow F[2]$

9: **return** F_1, F_2

Cuadro I
MEJORES RESULTADOS EN PRODIVOZ

Extractor	drms	Orden
Burg	120 Hz	9
LPC	123 Hz	6

Prodivoz, utilizando muestras de 50 ms.[1]. Los mejores resultados se observan en la Tabla I.

Si bien difieren en el orden del filtro, los mejores resultados coinciden en que en ambos se aplicó un filtro de preénfasis antes de la predicción lineal. La función de este preénfasis es contrarrestar la pendiente descendente del espectro del pulso glotal excitante, buscando una respuesta plana en amplitud. Recordemos que en el análisis de predicción lineal se utiliza como fuente un generador de ruido blanco cuya respuesta frecuencial es plana en el rango de interés.

No se construyó ningún clasificador y por tanto no hay resultados referidos a la capacidad clasificadora.

VIII. Conclusiones

Los mejores resultados obtenidos indican que el error de estimación del vector formante es de aproximadamente $120Hz$. Creo que este número debería ser aún menor, aunque se debe tener en cuenta que incluso los formantes verdaderos son estimaciones susceptibles a error. Al medir dichos formantes de referencia en Praat, obtuvimos valores con diferencias del mismo orden, según la medición se realizara en distintos tiempos del fonema, a pesar de observarse formantes muy estables en el espectrograma. También había cambios ante modificaciones de los parámetros del estimador de Praat.

El sistema de evaluación elaborado es extensible y servirá de base para estudios más completos y con un mayor conjunto de datos. Algunas tareas propuestas como continuación:

1. Programar diversos clasificadores que utilicen clases definidas manualmente (por ejemplo, a partir de los datos de [7]) o que sean aprendidas de la base de datos.
2. Construir una base de datos más extensa. Esto puede realizarse a partir de *corpus* de habla existentes o se puede optar por la creación desde cero de una base representativa de la zona y rango etario objetivos.
3. Evaluar la incidencia de cada parámetro de los extractores y clasificadores, con el fin de obtener la mejor combinación;
   a) considerándolos independientes entre sí;
   b) considerándolos dependendientes —para esto se puede recurrir a técnicas de optimización como los algoritmos genéticos.
4. Añadir algoritmos de extracción, ya sea de formantes o de otras características. *Hice pruebas preliminares de filtrado cepstral para obtención de formantes, pero los resultados fueron malos y el algoritmo merece revisión.*

Considero importante mejorar y documentar el sistema de evaluación, que será compartido con la intención de reducir la barrera de entrada al estudio de reconocimiento de vocales. Los algoritmos de reconocimiento, las bases de datos y formantes "verdaderos" extraídos manualmente también serán publicados[2] para facilitar la comparación de nuevos enfoques.

Apéndice A
Formantes de la base de datos Prodivoz

Obtenidos manualmente en Praat, variando los parámetros para obtener valores confiables en base a inspección visual del espectrograma.

[1]Estas evaluaciones están en los scripts `prodivoz1_LPC.m` (extractor6) y `prodivoz1_BURG.m` (extractor3) respectivamente, de la rama v1.0 del repositorio

[2]El repositorio se encuentra en https://github.com/aguadopd/avi_evaluacion_reconocimiento_formantes

Cuadro II
FORMANTES PRODIVOZ

Archivo	**Fonema**	F_1	F_2
A_Agu_Hablada_nr.wav	a	1010	1536
A_Agu_nr.wav	a	973	1474
A_Bruno_Hablada_nr.wav	a	820	1266
A_Bruno_nr.wav	a	788	1217
A_flaco_nr.wav	a	765	1234
a_g.wav	a	770	1378
A_Mari_nr.wav	a	985	1506
a_p.wav	a	752	1270
E_Agu_Hablada_nr.wav	e	468	2419
E_Agu_nr.wav	e	586	2204
E_Bruno_Hablada_nr.wav	e	438	2032
E_Bruno_nr.wav	e	453	1893
E_flaco_nr.wav	e	459	1889
e_g.wav	e	421	2031
E_Mari_nr.wav	e	563	2321
e_p.wav	e	467	2017
I_Agu_Hablada_nr.wav	i	392	2653
I_Agu_nr.wav	i	318	2509
I_Bruno_Hablada_nr.wav	i	364	2295
I_Bruno_nr.wav	i	316	2184
I_flaco_nr.wav	i	318	2071
i_g.wav	i	284	2379
I_Mari_nr.wav	i	438	2568
i_p.wav	i	273	2192
O_Agu_Hablada_nr.wav	o	575	909
O_Agu_nr.wav	o	560	978
O_Bruno_Hablada_nr.wav	o	567	918
O_Bruno_nr.wav	o	554	924
O_flaco_nr.wav	o	576	896
o_g.wav	o	510	889
O_Mari_nr.wav	o	716	997
o_p.wav	o	544	902
U_Agu_Hablada_nr.wav	u	379	678
U_Agu_nr.wav	u	315	764
U_Bruno_Hablada_nr.wav	u	327	775
U_Bruno_nr.wav	u	323	783
U_flaco_nr.wav	u	344	694
u_g.wav	u	281	689
U_Mari_nr.wav	u	363	732
u_p.wav	u	338	692

AGRADECIMIENTOS

El autor agradece a Germán Más por su ayuda en el estudio del reconocimiento de vocales. También a él y a Ariadna Areche, Nicolás Icard y Edwin Barragán por su trabajo como miembros del equipo desarrollador del Asistente Vocal Interactivo, el dispositivo que originó este análisis.

REFERENCIAS

[1] Antonio Quilis. *Tratado de fonología y fonética españolas*. Ed. por Gredos. Editorial Gredos, 1993. Cap. 2.

[2] Juan Carlos Gómez. «Modelos de producción de voz». 1 de oct. de 2001. URL: http://www.fceia.unr.edu.ar/prodivoz/apuntes_index.html (visitado 24-06-2016).

[3] Federico Miyara. *Acústica del tracto vocal*. 2001. URL: http://www.fceia.unr.edu.ar/prodivoz/apuntes_index.html (visitado 24-06-2016).

[4] Lawrence R. Rabiner y Ronald W. Schafer. «Introduction to Digital Speech Processing». En: *FNT in Signal Processing* 1.1–2 (2007), págs. 1-194. DOI: 10.1561/2000000001.

[5] Keelan Evanini. «The permeability of dialect boundaries: A case study of the region surrounding Erie, Pennsylvania». Tesis doct. University of Pennysilvania, 22 de dic. de 2009. Cap. 4.

[6] Juan Carlos Gómez. *Procesamiento Digital de Señales de Voz*. 2012. URL: http://www.fceia.unr.edu.ar/prodivoz/home_index.html (visitado 24-06-2016).

[7] Edgardo Bonzi y col. «Study of the characteristic parameters of the normal voices of Argentinian speakers». En: *Papers in Physics* 6 (jul. de 2014). DOI: 10.4279/pip.060002.

[8] Paul Boersma y David Weenink. *Praat: doing phonetics by computer*. Ver. 6.0.19. 2016. URL: http://www.praat.org.