Revista Ingeniería Biomédica
ISSN 1909-9762. Volumen 6, número 11, enero-junio 2012, 53-60
Escuela de Ingeniería de Antioquia-Universidad CES, Medellín, Colombia

Importancia de las frecuencias de resonancia del tracto vocal en la estimación de posiciones articulatorias

Alexander Sepúlveda[1, Ψ], Diana Margarita Casas Gómez[1], Germán Castellanos[2]

[1]*INDETECA, Escuela Colombiana de Carreras Industriales, Bogotá.*
[2]*Signal Processing and Recognition Group, Universidad Nacional de Colombia-Manizales.*

Recibido 13 de abril de 2012. Aceptado 11 de julio de 2012

IMPORTANCE OF THE RESONANCE FREQUENCIES OF THE VOCAL TRACT IN ESTIMATING ARTICULATORY POSITIONS

Resumen —La inversión articulatoria, cuyo objetivo es estimar la posición de los órganos articuladores a partir de la información contenida en la señal de voz, ofrece una variedad de potenciales aplicaciones en el campo de la voz; sin embargo, este es un problema aún por resolver. En este sentido, buscar representaciones con la capacidad de incrementar el desempeño de los sistemas de inversión articulatoria es una tarea importante. El presente trabajo analiza la relevancia de los formantes como entrada para los sistemas de inversión articulatoria. Para ello se implementa un análisis analítico y estadístico. En el caso analítico se utiliza un sintetizador articulario, el cual simula la ecuación de tubos concatenados que modelan el tracto vocal. Para el análisis estadístico se estudian datos reales provenientes de un articulógrafo electromagnético para los cuales se estima la asociación entre las características acústicas y los movimientos de los órganos articuladores. A modo de medida de asociación estadística se utiliza la medida de información . Los resultados entregados por el análisis son corroborados en un sistema de inversión articulatoria basado en redes neuronales. Se observa una mejora en el valor de error cuadrático medio del 2,2% y para el caso de la medida de desempeño de la correlación, una mejora del 2,8%.

Palabras clave — Inversión articulatoria, Información mutua, Resonancias del tracto vocal, Redes neuronales, Sintetizador articulatorio.

Abstract — Acoustic-to-Articulatory inversion, which seeks to estimate an articulator position using the acoustic information in the speech signal, offers several potential applications in the field of speech processing. In this context, it is important to use acoustic parameters with the ability to increase the performance of acoustic-to-articulatory inversion systems. This paper analyzes the importance of formants as inputs to such inversion systems from an analytical and a statistical perspective. The former is based on an articulatory synthesizer that simulates the voice signal from the vocal tract. The statistical analysis is based on real data provided by an electromagnetic articulograph, for which we estimate the statistical association between acoustic features and articulator movement. As a measure of statistical association, the information measure is utilized. The results are tested on a neural-network-based Acoustic-to-Articulatory inversion system. The use of formants as inputs led to an improvement of 2.2% and 2.8% in the root-mean-square error and correlation values, respectively.

Keywords — Acoustic-to-Articulatory inversion, Articulatory synthesizer, Mutual information, Neural network, Resonances of the vocal tract.

Ψ Dirección para correspondencia: alexander.sepulveda.sepulveda@gmail.com

I. Introducción

La inversión acústica-articulatoria, cuyo objetivo es obtener información articulatoria a partir de la información acústica contenida en las señales de la voz, ofrece nuevas perspectivas y aplicaciones interesantes en el campo del procesamiento de la voz [1]. Sin embargo, esta área está en investigación [2] y aún no ha sido desarrollado un sistema que entregue una solución al problema propuesto. Un sistema adecuado para la reconstrucción de las configuraciones articulatorias podría ser utilizado en varias aplicaciones, tales como: a) aplicaciones basadas en cabezas parlantes; un ejemplo de esto son los programas de computador que guían el aprendizaje de una segunda lengua y las ayudas visuales para las terapias articulatorias en los niños con dificultades auditivas y del habla; b) en codificación, dado que los movimientos de articulación son relativamente lentos; y c) para complementar la representación en los sistemas de reconocimiento del habla con el fin de mejorar su rendimiento debido a su capacidad para representar de una mejor manera los fenómenos co-articulatorios.

Varios métodos de mapeo acústico a articulatorio usan datos reales para el entrenamiento de los correspondientes modelos, los cuales han incrementado en popularidad debido al desarrollo reciente de sistemas que permiten la recolección de información articulatoria. Entre estos métodos se mencionan: métodos basados en redes neuronales [3], modelos de mezclas gaussianas [4-5] y modelos basados en cadenas de Markov [6]. En estos sistemas la representación más comúnmente usada para las señales de voz corresponde a parámetros basados en bancos de filtros en la escala *Mel*, los cuales pueden ser optimizados con el fin de obtener un mejor rendimiento [7]. Sin embargo, parámetros acústicos de notable importancia en el campo de la voz como lo son los formantes son muy poco utilizados en sistemas como los mencionados anteriormente.

Los formantes corresponden a las frecuencias más relevantes de resonancia del tracto vocal. A medida que los articuladores se mueven, la forma del tracto vocal cambia, por lo tanto las frecuencias de resonancia también cambian. Por consiguiente hay una relación estrecha entre la posición de los articuladores y las frecuencias de resonancia. Parámetros de voz que poseen esta misma propiedad pueden ser usados para mejorar el desempeño de los sistemas basados en aprendizaje de máquinas. Al respecto, en [8] se obtuvieron los mejores resultados al utilizar LSF (*Line Spectral Frequencies*) y PLP (*Perceptual Linear Predictive*), los cuales están mejor relacionados con el tracto vocal que otras características usadas en el mencionado trabajo. Sin embargo, las resonancias del tracto vocal no fueron incluidas a modo de parámetros a estudiar en [8]. Por otro lado, en [9] se reporta una mejoría en el rendimiento cuando se adicionan los formantes a un sistema basado en MFCC (*Mel-Frequency Cepstral Coefficients*) y GMM (*Gaussian Mixture Models*).

La presente publicación analiza la contribución de las resonancias del tracto vocal en la inferencia de las posiciones de los articuladores. Para este fin, analizaremos la relación entre las trayectorias articulatorias y los formantes desde un enfoque analítico y estadístico. Desde el punto de vista analítico hacemos uso del sintetizador articulatorio de Maeda. Desde el punto de vista estadístico, estimamos los valores de asociación estadística entre el movimiento de los articuladores y las frecuencias de resonancia del tracto vocal mediante la estimación de la información mutua; seguido de una evaluación del desempeño de un sistema de inversión articulatoria al agregar las frecuencias de resonancia del tracto vocal a modo de entradas. Esto con el fin de mostrar la utilidad del fenómeno expuesto en un sistema de inversión articulatoria basado en redes neuronales.

El análisis estadístico se basa en datos articulatorios reales provenientes de un articulógrafo electromagnético (EMA, *Electromagnetic Articulograph*). El desarrollo de este dispositivo es reciente y permite realizar mediciones de la actividad mecánica del habla. Además, por su principio de funcionamiento, hace posible la adquisición de cantidades relativamente grandes de datos articulatorios reales, y por tanto, permite el análisis de la relación estadística entre los fenómenos articulatorio y acústico de una manera más fiable.

Este trabajo muestra que la relación estadística es mayor para el caso de los formantes que para el caso de los MFCC; además, esta mayor asociación estadística con los movimientos articulatorios se traduce en una mejor estimación de la forma del tracto vocal.

II. Materiales y Métodos

En la presente sección se aborda el sintetizador de Maeda, la base de datos y el concepto de información mutua X^2 utilizada para medir la asociación estadística. Finalmente, se muestra la forma de evaluación de la contribución de los formantes en los sistemas de inversión articulatoria.

a. Sintetizador articulatorio de Maeda

El sintetizador articulatorio de Maeda y su respectiva implementación se expone en [10]. Los parámetros usados para gobernar la forma del tracto vocal se muestran en la Fig. 1. La información obtenida es organizada en un *Codebook;* sin embargo, la construcción de un *codebook*

que cubra una parte lo suficientemente representativa en los dominios articulatorio y acústico es una tarea difícil. La implementación del *codebook* se realiza como se muestra en [2]. Las entradas articulatorias son los parámetros articulatorios de la Fig. 1, en donde cada configuración articulatoria genera un conjunto particular de valores de formantes, constituyendo de esta forma la contraparte acústica. El *Codebook* está diseñado de tal forma que el patrón de distribución en dos dimensiones de F1-F2 cubre un amplio intervalo, en donde F1 corresponde al primer formante y F2 al segundo. La configuración de los parámetros de calibración del sintetizador es obtenido a partir de imágenes de RM (Resonancia Magnética) del tracto vocal de un hablante femenino utilizando el método desarrollado en [11].

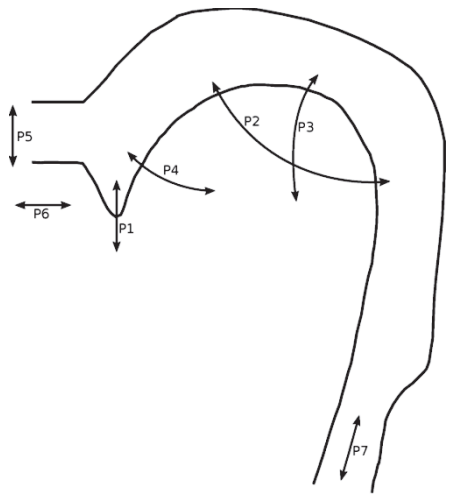

Fig. 1. Parámetros de Maeda.

b. Base de datos.

Los datos articulatorios del presente trabajo provienen de la base de datos MOCHA-TIMIT. Las frases usadas proporcionan señales de voz fonéticamente diversas. Se usan los datos correspondientes al hablante *fseω0*. El conjunto de entrenamiento está conformado por 368 frases, mientras que el conjunto de prueba esta compuesto por 46 frases. La base de datos *MOCHA* incluye cuatro secuencias de datos grabadas simultáneamente: la señal acústica con una frecuencia de muestreo de 16 Khz, laringografía, electropalatografía y datos EMA.

El sistema EMA se basa en el hecho de que cuando una bobina es introducida en un campo magnético, el cual varía de forma sinusoidal a una frecuencia en particular, una corriente con la misma frecuencia se induce en la bobina. El voltaje de la corriente inducida es inversamente proporcional, aunque de forma aproximada, al cubo de la distancia entre el transmisor y las bobinas. Por lo tanto, al medir el voltaje de la corriente inducida se puede inferir la distancia respecto a algún punto de referencia en particular.

El movimiento de las bobinas receptoras en el sistema EMA, las cuales se adhieren a los articuladores, son muestreadas a 500 Hz. Las bobinas se fijan a los incisivos inferiores (li), el labio superior (up), el labio inferior (ll), la punta de la lengua (tt), cuerpo de la lengua (tb), el dorso de la lengua (td), y el paladar blando (v). Las dos bobinas en el puente de la nariz y los incisivos superiores proporcionan puntos de referencia que permiten corregir los errores producidos por los movimientos de la cabeza. Las etiquetas que acompañan la base de datos MOCHA-TIMIT fueron usadas con el fin de descartar segmentos de silencio presentes al inicio y fin de las frases [3]. Las trayectorias de EMA se submuestrean a 100Hz, después de un proceso de filtrado que remueve las frecuencias altas.

Dado que los movimientos de los articuladores son causados por las contracciones musculares en el tracto vocal, que generalmente tienen anchos de banda por debajo de los 15Hz [12], las trayectorias EMA se suavizan con un filtro pasa-bajas con el fin de atenuar aquellas frecuencias por encima de los 20 Hz. El proceso de filtrado de las señales EMA se realizó tanto de manera directa como inversa con el fin de evitar posibles distorsiones de fase sobre la señal.

Se realiza el proceso de normalización sugerido en [13]. El proceso de normalización convencional calcula los valores promedio y desviaciones estándar globales para luego aplicarlos a las trayectorias EMA, pero ello podría generar dificultades debido a que los valores medios varían de una frase a otra durante el proceso de grabación. Mientras que los cambios rápidos de los valores medios son atribuidos al contenido fonético de cada frase, los cambios lentos son causados principalmente por la adaptación articulatoria del sujeto durante la sesión de grabación. Es útil eliminar la segunda clase de variación mientras se mantiene la segunda; lo que se consigue restando una versión de valores medios obtenidos al pasar el vector de medias por un filtro pasabajas. Se trabaja a una rata de muestreo de 100 Hz.

c. Medida de información X^2 mutua para detectar la relación estadística.

Se utiliza el concepto de información mutua medida mediante el criterio X^2, denotado por $I(\cdot,\cdot) \in \mathbb{R}$, con el fin de medir la asociación estadística entre la posición de los articuladores y los parámetros acústicos. Es importante señalar que la medida $I(\cdot,\cdot) \in \mathbb{R}$ está en capacidad detectar tanto relaciones lineales como no lineales, lo que hace a este concepto muy útil.

$I(x(\cdot), y(\cdot)) \in \mathbb{R}$ contiene la información de la trayectoria del articulador $y^m(t) \in y^i$ correspondiente a cada característica acústica $x(t, f_k)$ en el tiempo t y componente f_k. En términos generales, la medida de información X^2 es considerada como una distancia entre una distribución de probabilidad conjunta $P_{xy}(\cdot,\cdot)$ y el producto de las distribuciones marginales, $P_x(\cdot)$ y $P_y(\cdot)$. En lugar de la medida de información mutua convencional, en el presente estudio se utiliza la medida de la información X^2, la cual puede ser implementada sin necesidad de llevar a cabo una estimación explicita de la función de densidad de probabilidad [13]. La medida de información X^2 se describe según (1).

$$I\left(x(t,f_x), y^m(t)\right) = \iint_{\mathbb{R}} \frac{\left(P_{xy}(x(t,f_k),y^m(t)) - P_x(x(t,f_k),y^m(t))P_y(x(t,f_k),y^m(t))\right)^2}{P_x(x(t,f_k))P_y(y^m(t))} dxdy \quad (1)$$

La expresión dada en (1) puede estimarse basándose en el concepto de la razón de densidades, denotado como $r_k^m = r(x(t,f_k), y^m(t))$, para las variables aleatorias $x(t,f_k)$ y $y^m(t)$, como se sugiere en [14] y según se muestran en (2):

$$I_k^m = I\left(x(t,f_k), y^m(t)\right) = \iint_{\mathbb{R}} (r_k^m - 1)^2 P_x\left(x(t,f_k)\right) P_y\left(y^m(t)\right) dxdy \quad (2)$$

donde el termino $r_k^m \in \mathbb{R}$ se define de la siguiente manera:

$$r_k^m = \frac{P_{xy}\left(x(t,f_k), y^m(t)\right)}{P_x\left(x(t,f_k)\right) P_y\left(y^m(t)\right)} \quad (3)$$

Los detalles adicionales acerca de la estimación de la función de relación de densidad (3) se reportan en [15].

d. Evaluación de la contribución de formantes a un sistema de inversión articulatoria.

La presente sección tiene por objetivo mostrar la utilidad de los formantes en un sistema de inversión articulatoria basado en redes neuronales. Los coeficientes MFCC son ampliamente usados en el procesamiento de señales de voz. En el presente trabajo se tienen dos sistemas de mapeo del dominio acústico al articulatorio; el primero utiliza los coeficientes MFCC para representar la señal de voz, y el segundo, usa los formantes junto a los coeficientes MFCC. Luego se compara el desempeño de estos dos sistemas. La mejora en el desempeño puede ser usada a modo de indicación de la cantidad de información relevante que será adicionada al sistema. Se construyen dos conjuntos de entrenamiento, el primero está compuesto por 20 coeficientes MFCC, y el segundo corresponde a la unión de los MFCC y los formantes, lo que da un total de 24 entradas. Los parámetros estimados están en sincronía con la señal articulatoria. En trabajos recientes comúnmente se incluye información desde -160 ms a 160 ms y en algunos casos desde -200 ms hasta 200 ms. Si se toman todos los parámetros dentro de esa ventana de contexto, el error cuadrático medio (MSE) podría llegar a ser afectado por el número de parámetros. Si se deja que la cantidad de muestras sea muy superior a la cantidad de entradas la estimación del MSE será notablemente más fiable. Aún más importante, el interés no es superar el rendimiento de las técnicas desarrolladas recientemente; en contraste, es analizar la pertinencia y la contribución de los formantes.

La red neuronal implementada está compuesta por tres capas. La segunda capa utiliza la función *tansig* a modo de función de activación, y la capa de salida utiliza una función de activación lineal. El proceso de entrenamiento de la red neuronal se realizó utilizando el algoritmo de optimización SCG (*Scaled Conjugate Gradient)*. Los algoritmos de optimización de primer orden, tales como el de gradiente descendente, usan la primera derivada de la función de error. En contraste, los de gradiente conjugado utilizan la segunda derivada de la función de error, lo que los hace considerablemente más rápidos que el algoritmo estándar de gradiente descendente [16]. Sin embargo, el requerimiento de memoria para el algoritmo de optimización SCG es menor que en el caso de algoritmos convencionales de segundo orden. Una ventaja adicional es que el SCG no requiere que el usuario busque los parámetros de ajuste tales como la rata de entrenamiento. Por esta razón, seleccionamos SCG como el método para estimar los parámetros de las redes neuronales. El modelo neuronal se entrenó durante 100 épocas; este último valor se seleccionó empíricamente. Luego se calculó el rendimiento de las redes neuronales.

III. Resultados

En esta sección se muestran y analizan los resultados de la evidencia proveniente del sintetizador articulatorio, la evidencia proveniente de datos articulatorios reales, y las pruebas de los formantes en un sistema de inversión articulatoria basado en redes neuronales.

a. Evidencia proveniente del sintetizador articulatorio.

Los diagramas de dispersión entre los formantes y algunos parámetros articulatorios se muestran en la Fig. 2, donde se observa que el valor del primer formante (*F1*) tiende a incrementarse con el valor del parámetro correspondiente al dorso de la lengua (*P2*). Sin embargo, el valor de *F1* no sólo depende de *P2,* ya que muchas configuraciones pueden generar cambios notables en *F1*. En la Fig. 2C se observa que *F3* decrece a medida que la posición del parámetro articulatorio *P4* incrementa. Similar a *F1* muchas configuraciones articulatorias pueden producir grandes cambios al formante. La Fig. 2B muestra el diagrama de dispersión del segundo formante *versus P3* (forma del dorso de la lengua).

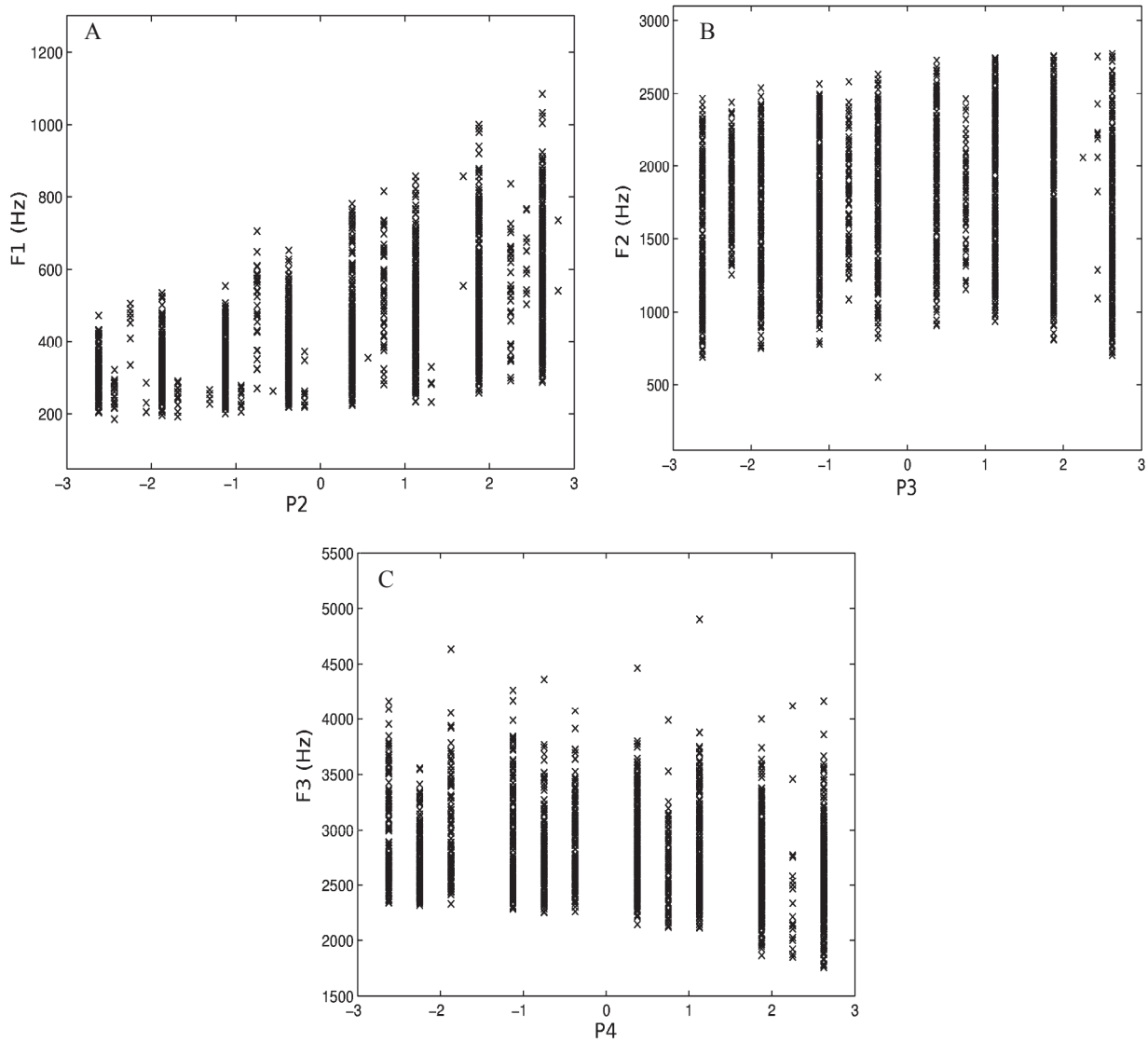

Fig. 2. Diagramas de dispersión de los formantes contra los parámetros articulatorios: A) Primer formante contra la posición del dorso de la lengua (P2). B) Segundo formante contra la forma del dorso de la lengua (P3). C) El tercer formante contra el ápice de la posición (P4).

b. Evidencia proveniente de datos articulatorios reales.

Con el fin de estimar la relevancia de los coeficientes MFCC y los formantes, creamos una medida que consiste en tomar el promedio de los valores de la información mutua de las características de entrada. Esto es, el promedio de los formantes es calculado según (4).

$$\bar{I}(f, y^m(t)) = \sum_{i=1}^{4} I(f_i, y^m(t)) \quad (4)$$

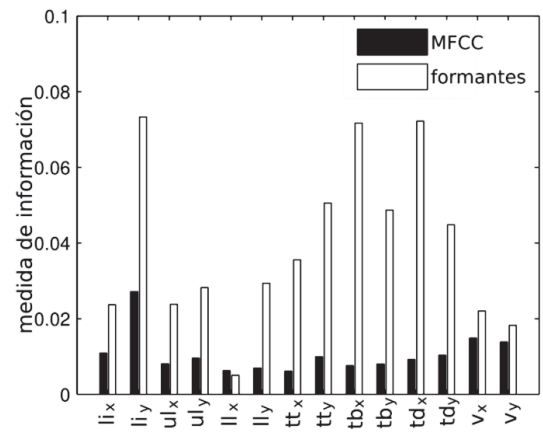

Fig. 3. Medidas de información mutua entre la posición de los articuladores y los diferentes tipos de entradas, MFCC ($\bar{I}(\zeta, y^m)$, en negro) y formantes (($\bar{I}(f, y^m)$, en blanco).

Donde f_i son los formantes y $f = [f_1 \ldots \ldots f_4]^T$ y y^m son los canales EMA m = 1, …14.. El promedio $\bar{I}$ para las características MFCC son calculadas como se muestra en (5).

$$\bar{I}(\zeta, y^m(t)) = \sum_{i=1}^{13} I(\zeta_i, y^m(t)) \quad (5)$$

Donde ζ es el vector de MFCC coeficientes. El número 13 corresponde a la cantidad de coeficientes MFCC, el cual es comúnmente usado en aplicaciones de procesado de la voz. La Fig. 3 muestra los valores de las medidas de información $\bar{I}(f, y^m)$ y $\bar{I}(\zeta, y^m)$ para los 14 canales EMA.

c. Formantes en un sistema de inversión articulatoria basado en redes neuronales.

La presente sección muestra la mejora en el desempeño medido mediante la raíz del error cuadrático medio (*RMSE*, *Root Mean Square Error*) y correlación de Pearson cuando los formantes son adicionados al conjunto de valores MFCC para cada uno de las trayectorias articulatorias. El RMSE se calculó según (6).

$$RMSE_i = \sqrt{\frac{1}{N}\sum_{t=1}^{N}\left(y^m(t) - \hat{y}^m(t)\right)^2} \quad (6)$$

Donde N es el número de vectores de entrada-salida, o muestras, en el conjunto de entrenamiento; $\hat{y}^m$ es la secuencia estimada por la red neuronal m; y y^m es el vector de referencia o trayectoria de datos EMA. Además se estima el valor de la correlación entre las trayectorias EMA y los valores estimados por las redes neuronales de la forma convencional.

Los resultados de los experimentos usando coeficientes MFCC y MFCC-formantes se observan en la Fig. 4. El valor RMSE y correlación se calculan para el caso en el que se utilizan 6, 10, 14, 18 y 22 neuronas en la capa intermedia. Se puede observar que los resultados son mejores para el conjunto de entrada MFCC-formantes que el caso en el que se utilizan sólo coeficientes MFCC, tanto para el RMSE como para la correlación. Incluso al modificar la cantidad de parámetros en la red neuronal los resultados se mantienen.

La mejora en el desempeño para cada uno de los canales, cuando se usan 14 neuronas en la capa intermedia, se muestran en la Fig. 5. Si se compara la Fig. 5 con la Fig. 3, se puede observar similitud entre estas dos gráficas. En particular, valores de asociación estadística mayores implican mayores desempeños en los valores de RMSE y correlación. Para el caso de 14 neuronas en la capa intermedia, al incluir los formantes se logra mejorar el desempeño en términos de RMSE en un 2,5%, y 2,9% para el caso de la medida de correlación.

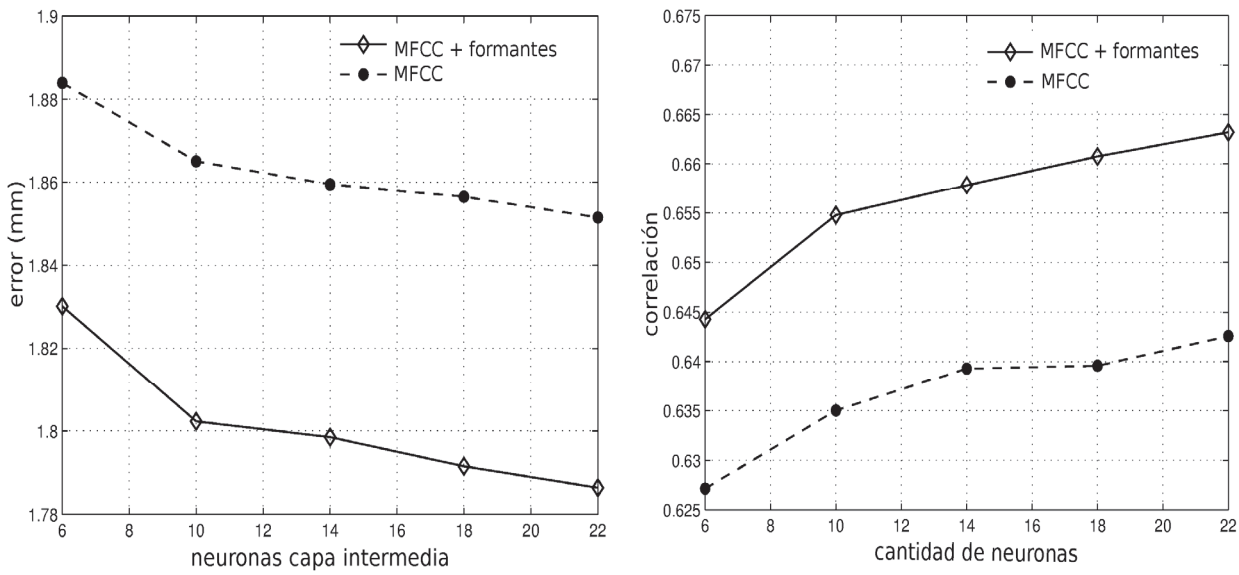

Fig. 4. Gráfica de desempeño utilizando los conjuntos de entrada MFCC y MFCC+formantes en términos del RMSE (izquierda) y de la correlación de Pearson (derecha) para diferentes cantidades de elementos en la capa intermedia de la red neuronal.

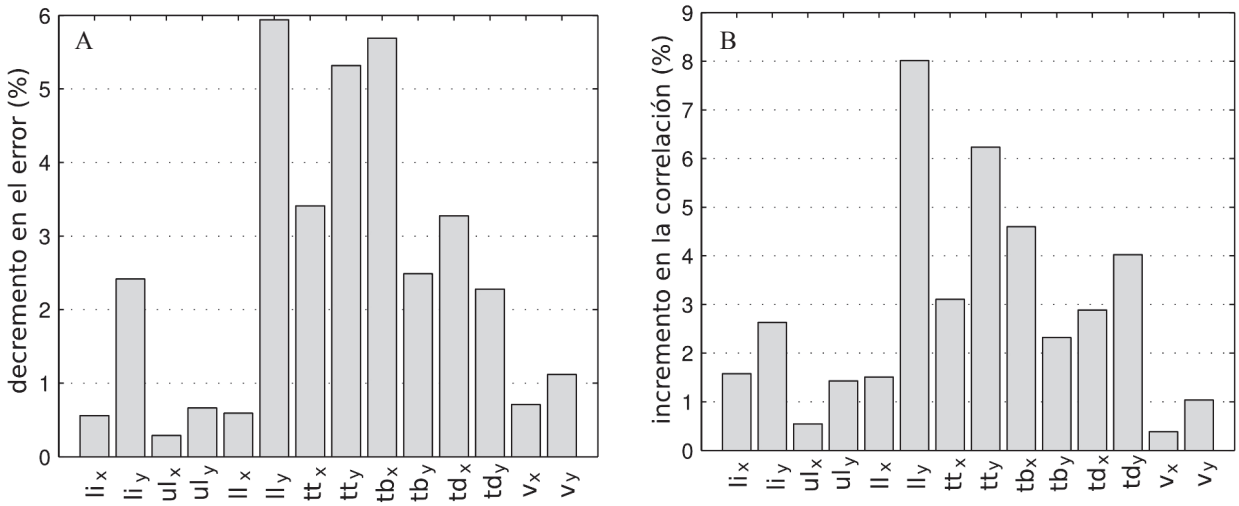

Fig. 5. Medición de la mejora en el desempeño con respecto al RMSE y la correlación para las diferentes posiciones de los articuladores. A) Porcentaje de decremento en el error. B) Porcentaje de incremento en la correlación.

IV. Discusión

Una de las formas de analizar la relación entre el movimiento de los articuladores y los formantes es mediante el uso de sintetizadores articulatorios [17]. No obstante, es preferible realizar el análisis basado en datos reales. Dentro de los métodos de medición de los movimientos articulatorios se tienen rayos X y rayos X focalizados. En el primer caso, por problemas asociados a la radiación, la cantidad de datos que se pueden tomar son escasos. Lo mismo ocurre con los rayos X focalizados, pero por motivos de costos. A diferencia de los métodos basados en Rayos X, la tecnología EMA sí permite la recolección de una buena cantidad de datos a precios moderados, lo cual conlleva a poder realizar análisis más confiables. En nuestro trabajo se utilizó la ventaja ofrecida por la disponibilidad de los mencionados datos para realizar un análisis desde el punto de vista estadístico de la relación entre los formantes y la posición de los articuladores.

En [9] se muestra que al adicionar los formantes a un sistema de inversión articulatoria basado en mezclas gaussianas, la tasa de error disminuye en 3,4% y la correlación aumenta en 2,7%. En el presente trabajo, el cual está basado en redes neuronales artificiales, el error se decrementa en 2,2% aproximadamente y la correlación se incrementa en 2,8%; lo cual es comparable con lo encontrado en [9]. De otra parte, la medida de información mutua permite observar que los formantes son en buena medida influidos por la lengua, el principal órgano articulador. El experimento de regresión basado en redes neuronales del presente trabajo permite confirmar este hallazgo, ya que al agregar los formantes su efecto positivo se ve especialmente reflejado en la inferencia del movimiento del ápice, el cuerpo y el dorso de la lengua.

Se muestra desde el punto de vista estadístico que existe una relación intrínseca entre los formantes y la posición de la lengua (ver Fig. 3), lo que por ende provoca una mejora en el desempeño de los sistemas de inversión articulatoria, como también queda demostrado (ver Fig. 4). Este hallazgo permite establecer que los sistemas de inversión articulatoria encontrados en el estado del arte tendrían un mejor desempeño si incluyeran a los formantes dentro del grupo de parámetros de representación de la señal de voz.

V. Conclusión

La búsqueda de representaciones que permitan inferir de forma más confiable la forma del tracto vocal es un problema abierto y de interés para comunidad científica de la voz. El presente trabajo mostró que la relación estadística entre los formantes y la forma del tracto vocal es mayor que para el caso el caso de los MFCC. Adicionalmente, se encontró que esta mayor relación estadística se puede utilizar para estimar con mayor precisión la posición de los órganos articuladores.

Agradecimiento

El presente trabajo fue financiado principalmente por COLCIENCIAS (apoyo a doctorados nacionales). Adicionalmente, los autores agradecen al profesor Yves Laprie del instituto LORIA-INRIA (Nancy, Francia) por facilitar el software requerido para el desarrollo de la sección II.A y III.A.

Referencias

[1]. Schroeter J., Sondhi M. M. Techniques for estimating vocal-tract shapes from the speech signal. IEEE *Transactions on Speech and Audio Processing*, 2(1), 133–150, 1994.

[2]. Potard B., Laprie Y., Ouni S.. Incorporation of phonetic constraints in acoustic-to-articulatory inversion. *Journal of Acoustical Society of America*, 2008.

[3]. Richmond K., King S., Taylor P.. Modelling the uncertainty in recovering articulation from acoustics. *Computer, Speech & Language*, 17, 153–172, 2003.

[4]. T., Black A., Tokuda K. Statistical mapping between articulatory movements and acoustic spectrum using Gaussian mixture models. *Speech Communication*, 2008.

[5]. Ozbek Y., Hasegawa-Johnson M., Demirekler M. Estimation of articulatory trajectories based on gaussian mixture model (GMM) with audio-visual information fusion and dynamic Kalman smoothing. *IEEE Transactions on Audio, Speech, and Language Processing*, 2011.

[6]. Hiroya S., Mochida T. Multi-speaker articulatory trajectory formation based on speaker-independent articulatory HMM. *Speech Communication*, 48, 1677–1690, 2006.

[7]. Kumar P., Goldstein L., Narayanan S. Processing speech signal using auditory-like filterbank provides least uncertainty about articulatory gestures. *Journal of Acoustical Society of America*, 129(6), 4014–4022, 2011.

[8]. Qin C., Carreira-Perpiņán M. A comparison of acoustic features for articulatory inversion. *In InterSpeech* 2007.

[9]. Ozbek Y., Hasegawa-Jhonson M., Demirekler M. Formant trajectories for acoustic–to–articulatory inversion. *In InterSpeech* 2010.

[10]. S. A digital simulation method of the vocal-tract system. *Speech Communication*, 1(3-4), 199–229, December 1982.

[11]. Fontecave J., Berthommier F. A semi-automatic method for extracting vocal tract movements from X-ray films. *Speech Communication*, 2009.

[12]. Hogden J., Lofqvist A., Gracco V., Zlokarnik I., Rubin P., Saltzman E. Accurate recovery of articulator positions from acoustics: new conclusions based on human data. *Journal of Acoustical Society of America*, 1996.

[13]. Richmond K. Articulatory feature recognition from the acoustic speech signal. PhD thesis, University of Edinburgh, Edinburgh, 2001.

[14]. Suzuki T., Sugiyama M., Kanamori T., Jun Sese J. Mutual information estimation reveals global associations between stimuli and biological processes. *BMC Bioinformatics*, 10(1), 2009.

[15]. P. f-information measures for efficient selection of discriminative genes from microarray data. *IEEE Transactions on Biomedical Engineering*, 56(4), 1063–1069, April 2009.

[16]. Kostopoulos A. E., Grapsa T. N. Self-scaled conjugate gradient training algorithms. *Neurocomputing*, 72, 3000–3019, 2009.

[17]. Lindblom B. E., Sundberg J. E. Acoustical consequences of lip, tongue, jaw, and larynx movement. J. Acoust. Soc. Am. Oct.; 50(4), 1166-79, 1971.