Revista Ingeniería UC, Vol. 17, No. 7, Abril 2010 55 - 60

Uso y aplicaciones de los métodos de calculo de la frecuencia fundamental y de la calidad objetiva de la señal de voz.

Jesús Jiménez[*,a], José Antonio Díaz[b], Carlos Jiménez[c], Miguel Fasanella[c]

[a]*Universidad de Carabobo, Ciclo Básico, Departamento de Matemáticas, Barbula, Carabobo, Venezuela*
[b]*Escuela de Ingeniera en Telecomunicaciones, Barbula, Carabobo, Venezuela*
[c]*Universidad de Carabobo, Escuela de Ingeniería Eléctrica, Barbula, Carabobo, Venezuela*

Resumen.-

La frecuencia fundamental es un parámetro importante para la determinación de la calidad de la voz, los estudios acerca de la frecuencia fundamental en señales de voz han arrojado, en algunos casos, resultados diferentes entre ellos para distintas muestras de señales de voz. Los parámetros propuestos, para determinar la calidad de la voz a través de indicadores han sido diversos. Del análisis de las propuestas y de los métodos de obtención de la frecuencia fundamental se determina que la información relevante se encuentra en la zona de baja frecuencia del espectro.

Palabras clave: frecuencia fundamental, pitch, calidad de voz, evaluación de la voz.

Use and applications of the methods for determining the fundamental frequency and the objective quality of the voice sign

Abstract.-

The fundamental frequency is an important parameter for determining voice quality, some of the studies about the computation of the fundamental frequency in voice signals have yielded, in some cases, different results among them for different samples of voice signals. The proposed parameters, for calculation the voice quality through indicators, have been diverse. The analysis of the proposals and methods for obtaining the fundamental frequency suggests that the relevant information is in the area of low frequency of the spectrum.

Keywords: fundamental frequency, pitch, voice quality, evaluation of the voice.

Recibido: 14 mayo 2009
Aceptado: 24 marzo 2010

1. Introducción

La frecuencia fundamental (F_0) es el número de veces que vibran por segundo las cuerdas vocales. La percepción de los cambios de Fo viene dado por el tono, cuando aumenta F_0 el tono se hace agudo y cuando esta baja el tono se hace más grave. Anatómicamente [1], si las cuerdas vocales tienen mucha masa y están muy vascularizadas, la voz es grave; al contrario, si las cuerdas vocales son pequeñas y poco vascularizadas, la voz es aguda. Existen variaciones de F_0 voluntarias de la voz, por ejemplo, cuando hacemos inflexiones y cuando se canta; éstas se logran utilizando los músculos variando la longitud y tensión de las cuerdas vocales, o aumentando la presión subglotica.

En una situación ideal la F_0 no variaría, sin embargo, esto en la realidad no sucede, y dentro de la normalidad, la frecuencia entre cada ciclo vocal y ciclo siguiente no es exactamente igual.

[*]Autor para correspondencia
Correos-e: jjjimene@uc.edu.ve (Jesús Jiménez), jadiaz@uc.edu.ve (José Antonio Díaz), cjimenez@uc.edu.ve (Carlos Jiménez), miguelfasanella@hotmail.com (Miguel Fasanella)

Adicionalmente, se conoce que la frecuencia fundamental [1] durante la niñez esta alrededor de 240 *Hz*, en la pubertad para los varones desciende a 110 *Hz* y en las mujeres se coloca a 210 *Hz*, y hacia la tercera edad en los hombres aumenta a 140 *Hz*, y en las mujeres disminuye a 190 *Hz* en promedio.

La frecuencia fundamental tiene una importancia esencial en el procesamiento de señales de voz (síntesis, codificación, etc.), y de allí nace la necesidad de realizar el cálculo preciso y confiable de la misma. Sin embargo, existen muchas razones que hacen extremadamente difícil obtener la frecuencia fundamental de forma precisa y confiable [2]: la primera, la onda de excitación glotal no es un perfecto tren de impulsos periódicos, la segunda, la interacción del tracto vocal y la excitación glotal, en algunos casos, puede modificar significativamente la forma de la onda, la tercera, la dificultad de definir el exacto inicio y final del periodo correspondiente a F_0 en un segmento de señal de voz, la cuarta, poder distinguir entre sonido no vocal y vocal de bajo nivel, entre otros.

Existen variaciones involuntarias de F_0 de la voz [1], por falta de control en los músculos vocales (neurológicas), cuando existe un defecto en el cierre glótico que provoca vibraciones irregulares de las cuerdas vocales (aerodinámicas), y por causas mecánicas como asimetrías, o cambios en propiedades biomecánicas de las cuerdas. Las variaciones de frecuencias de ciclo a ciclo son razonablemente más altas para voces que presenten patologías.

2. Cálculo de la frecuencia fundamental

En el cálculo de la frecuencia fundamental se utilizan [3] métodos basados en el dominio del tiempo, en el dominio de la frecuencia, y métodos estadísticos que utilizan la teoría de las probabilidades. Se han desarrollado varios métodos, entre los más conocidos se tienen:

- Autocorrelación modificada usando clipping (AUTOC) [4]: utilizando el método de Center Clipping, que remueve los formantes de la señal, y luego utiliza el método de autocorrelación para obtener el periodo correspondiente a F_0.
- Método del Cepstrum (CEP)[5][6][7][8]: análisis cepstral es una forma de análisis espectral, donde la salida es la transformada inversa de Fourier del logaritmo de la transformada de Fourier de la señal de entrada. Fue diseñado para su utilización en síntesis y análisis de señales de voz, y se basa en el modelo de producción de voz compuesto por la convolución de la secuencia de excitación ($e(n)$) y la respuesta al impulso del tracto vocal ($v(n)$). Se conoce que la contribución del tracto vocal tiende a variar lentamente con la frecuencia, mientras la contribución de la excitación tiende a variar más rápidamente y periódicamente con la frecuencia; como consecuencia, se espera que la contribución de la excitación ocurra en múltiplos del periodo asociado a la Fo, y la contribución del tracto vocal ocurra cercano al origen de la gráfica del cepstrum.
- Técnica de Filtrado Inverso Simplificado (SIFT): en esta técnica la señal de voz es muestreada a 10 *kHz*, procesada por un filtro pasa-baja de 0,8 *kHz* [8] (0,9 *kHz* en [2]), se le aplica un proceso de decimación para reducir la rata de muestreo a 2 *kHz*, se obtienen los coeficientes de cuarto orden de filtro inverso LPC, se filtra la señal, su salida es autocorrelacionada, y es estimado el periodo asociado a Fo donde se presenta el mayor pico, se realiza la interpolación [8] en la vecindad del pico para realizar la clasificación de la señal en vocal o no vocal, y si es vocal queda definida la F_0.
- Método de reducción de datos (DARD): este método [2] consiste en hacer pasar la señal por un filtro pasa bajas de 900 *Hz*, se detecta el recorrido de los ciclos utilizando cruce por ceros (zero-crossing), se aísla e identifica el principal recorrido de ciclos usando la energía y limites silábicos del periodo asociado a F_0 , y un corrector de errores para proveer una medida razonable del periodo asociado a F_0.
- Método de procesamiento en paralelo (PPROC): este método [2] consiste en hacer pasar la señal de voz por un filtro pasa bajas de 900 *Hz*, son tomadas las medidas con diferentes referencias [9] correspondientes a m1, m2, m3, m4, m5 y m6 de la figura 1, cada una de esas

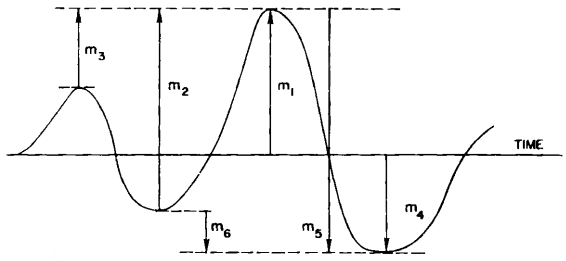

Figura 1: Valores de m1, m2, m3, m4, m5 y m6.

medidas es tomada para procesarla en un elemental estimador del período asociado a la F_0, y los seis valores obtenidos son evaluados en un algoritmo de decisión que determina el valor del periodo asociado a la F_0.

- Método de la media de la diferencia de la magnitudes (AMDF): este método [10] se basa en que dada una secuencia $d(n)$ periódica de periodo P , la secuencia con retardo k definida por $d(n) = s(n) - s(n-k)$ debería ser cero para valores de k múltiplos de P, y utilizando la autocorrelación de d(n), buscaría el valor de k distinto de cero, donde se obtiene el mínimo de la autocorrelacion. Existen diferentes versiones de ésta [2], que radican en la realización de un prefiltrado con filtro pasa baja de 900 *Hz*, y la utilización de cruce por cero, y la energía para clasificar la señal en vocal o no vocal.
- El estimador YIN [11]: este método se basa en la diferencia de funciones que intenta minimizar la diferencia entre la señal de voz y su duplicado retardado:

$$d_t\tau = \sum_{j=1}^{W}(x_j - x_{j+\tau})^2 \qquad (1)$$

Para reducir los errores por los subarmónicos se emplea la función de diferencias media acumulativa normalizada, ecuación (2):

$$d\prime_\tau \left\{ \begin{array}{ll} 1, & \text{si } \tau = 0 \\ \dfrac{d_\tau\tau}{\left[1/\tau \sum\limits_{j=1}^{\tau} d_\tau j\right]} & \end{array} \right\} \qquad (2)$$

Se introduce adicionalmente una interpolación parabólica del mínimo local que minimiza aun más los errores.

- Método de codificación de predicción lineal (LPC): es una técnica usada para modelar el tracto vocal utilizando predicción lineal, y resulta en una gráfica que envuelve al espectro de la señal y es utilizada para obtener la F_0 y los formantes.
- Relación de componentes de frecuencias: esta técnica se inicia con la extracción de picos del espectro de la señal, que llamaremos parciales f_m y f_k, y que son proporciones de la frecuencia fundamental. Para cada pareja de las parciales obtenidas, el algoritmo encuentra las frecuencias de los armónicos más pequeños que correspondan a la pareja de parciales, y existe el cociente $\frac{i}{j}$ que establece la relación entre las proporciones con respecto a la frecuencia fundamental. El error definido $e = \frac{f_m}{f_k} - \frac{i}{j}$ correspondiente a la pareja i y j es el error a minimizar. El cociente $\frac{i}{j}$ es sugerido en la versión original [12] y se tiene la hipótesis [3] que la frecuencia fundamental es superior a los 70 *Hz*.

2.1. Evaluación del cálculo de la frecuencia fundamental.

En el año 1976 fué publicado el trabajo desarrollado por Rabiner y otros [2], en el que se establece la comparación de la gráfica del período asociado a la frecuencia fundamental contra el tiempo para una serie de señales de voz de diferentes individuos, y evaluadas con los algoritmos AUTOC, CEP, SIFT, PPROC, DARD, LPC Y AMDF, este trabajo arrojó que no existía una preponderancia en el desempeño de los algoritmos, produciéndose en algunas oportunidades variaciones importantes entre ellos que dependían de la señal de voz a evaluar.

En el año 2003, Gerhard en su reporte técnico [3] de la comparación de tres métodos de estimación de la frecuencia fundamental, YIN, autocorrelación (AC) y CEP, tomando como criterios: la comparación del promedio de los valores absolutos de las diferencias entre los períodos asociados a la frecuencia fundamental, calculados por cada pareja de métodos en voz hablada y cantada, y el criterio de inspección visual en señales de voz escogidas, concluye que la naturaleza de las señales de voz habladas o cantadas es tal, que los algoritmos son buenos para

unas señales de entrada y no son buenos para otras, y en algunos casos, la inspección visual arroja que hay coincidencia de tres métodos en las medidas, y en otros casos solo la de dos de los métodos.

En la actualidad, como se observa de la formas de cálculo de frecuencia mostradas en este trabajo, y haciendo la salvedad de que no se incluyeron métodos estadísticos, nos lleva a afirmar que se requiere de un mayor esfuerzo de investigación para el cálculo de la frecuencia fundamental.

2.2. Jitter y Shimmer

El **jitter** o perturbación de la frecuencia [1], es la variación de la frecuencia fundamental entre cada ciclo vocal y el siguiente, es el tono variante de la voz que causa un sonido áspero. Los valores de jitter crecen al aumentar la edad de las personas, y en una muestra donde la frecuencia fundamental es mayor las perturbaciones son menores.

El **shimmer** [1] es la perturbación de la amplitud de la señal medida ciclo a ciclo.

3. Calidad de la voz

Aronson 1990, citado por [1] afirma que "hay alteración de la voz cuando esta difiere de las voces de otras personas del mismo sexo, similar edad y grupo cultural, en timbre, tono, volumen, flexibilidad y en dicción" (p.62). De la anterior definición se desprende que no existen criterios objetivos para la determinación de la voz normal. Se han realizados esfuerzos en la investigación para establecer objetivamente la calidad de la voz, según [13] es un desafío, y es más probable aprovechar la interacción entre la señal y el escucha para clasificar la voz, que tratar la calidad solamente como una función de la señal de voz, de esto ultimo los autores destacan que se va en camino a la objetivación.

3.1. Propuestas:

En la actualidad existen diversas propuestas para la clasificación objetiva de la calidad de la voz, en este trabajo nos enfocaremos en que métodos se utilizaron, como se escogieron los datos, los parámetros utilizados, y si está reflejado su desempeño en la investigación:

- Michaelis et al [14] proponen el parámetro: Relación de la excitación glotal al ruido (Glottal-to-Noise Excitation Ratio) que denominan GNE, y lo compara con la relación entre la energía del ruido y la energía de la señal con siglas en ingles NNE y con la relación de armónicos a ruido HNR. (NNE y HNR son inversos y se calculan de acuerdo a [15]). Los cálculos de los parámetros se realizan sobre señales sintéticas y se varía el ruido, el jitter y el shimmer. Como conclusión [14], NNE y HNR, no pueden ser usados para evaluar la calidad de la voz, existen relaciones de dependencia entre variación de amplitud o periodicidad y la adición de ruido. GNE es una medida confiable del nivel de ruido relativo (excitación glotal y ruido), en presencia de fuerte variación en amplitud y periodicidad.
- Michaelis et al [16] comprueban con voces sanas y patológicas que GNE puede ser usado para evaluar la calidad de la voz, como complemento del trabajo indicado en el aparte anterior. Las muestras digitalizadas eran de la vocal alemana ϵ sostenida.
- Casado et al [17] realiza estudios con adultos sanos y con disfonía por nódulos y pólipos vocales, utiliza como parámetros jitter, shimmer, NNE, HNR, la media de la f_0 y la relación señal ruido SNR (Signal-to-Noise Ratio), relación de la energía del ruido, y la energía total de la señal. Como resultado de los análisis, se encontró que los que presentaban patologías presentaron menores valores de la media de la F_0 , y más altos valores de jitter, shimmer, NNE, HNR y SNR. Las alteraciones mayores con respecto al grupo sano lo presentaron los pacientes con pólipos vocales. Las muestras digitalizadas eran la vocal /a/ sostenida.
- Del Pino et al [18] proponen los parámetros: PMR cociente del valor pico y del valor medio del espectro. SNRf se le añade un subíndice f para diferenciar la abreviatura utilizada por [17] y se define como la relación de la energía de la señal medida de 50 *Hz* a 2500 *Hz* y la energía de ruido medida de 1000 *Hz* a 2500 *Hz* de la señal de voz. SNRL relación entre la energía e la señal medida de 50 *Hz* y 500 *Hz* a la energía del ruido medida en el mismo rango de frecuencias. SNRM

relación entre la energía e la señal medida de 500 *Hz* y 1500 *Hz* a la energía del ruido medida en el mismo rango de frecuencias. SNRH relación entre la energía e la señal medida de 1500 *Hz* y 2500 *Hz* a la energía del ruido medida en el mismo rango de frecuencias. Las muestras digitalizadas eran voces previamente clasificadas por expertos como normales o patológicas, no se conocen las patologías. Para los cálculos se suavizó la señal con un filtro de mediana. Como resultado se concluye de los análisis estadísticos que los parámetros PMR, SNRM y SNRH resultaron significativos, destacándose como diferenciador PMR.

- Con base a la percepción:

1. Yu et al [19] realizó un estudio con hombres con rango de edades de 23-75 años, grupo de control (sano) y grupo con disfonía. Se digitalizó la vocal sostenida /a/ para el análisis, a excepción del cálculo de la presión subglotica que se utilizó /pa/. Se escogió la evaluación perceptual GRBAS para la clasificación de las disfonías, tomando solamente el grado G de la escala para la comparación univariada. Se obtuvo que en las variables [19] medidas para cada paciente o control del análisis univariado mostrado del grado G, hubo consistencias significativas entre al menos dos grados contiguos de nueve de las once variables escogidas.

2. Kreiman et al [20] escogieron un grupo de oyentes sin problemas auditivos con escasos conocimientos de fonética y otorrinolaringología, seleccionaron aleato-riamente las voces de 10 mujeres y 10 hombres con disfonía de una biblioteca de muestras grabadas, exceptuando la patología bifonación debido que el jitter y shimmer no están definidos para dichas señales, y se le presentaron al grupo de oyentes las voces seleccionadas con sus copia sintetizadas con variaciones en los parámetros jitter, shimmer y SNR, en donde los oyentes debían modificar los parámetros jitter, shimmer y SNR hasta que las voces emparejaran (de manera individual o en conjunto). Los oyentes mostraron muy poca sensibilidad a los cambios realizados en el jitter y el shimmer comparándolos con obtenidos con SNR. También concuerdan los resultados con los obtenidos de trabajos previos realizados por Gerratt et al [21] en donde se utilizaban oyentes expertos y se obtuvo un porcentaje de aciertos mayor evaluando el parámetro SNR.

- Se han realizado clasificaciones de calidad de la voz con modelos estadísticos: vectores de soportes, cadenas de Markov (HMM), redes neuronales, modelos Gaussianos (GMM), etc. Una propuesta muy interesante es la de Dibazar et al [22], que tomando el fonema /ah/ sostenido correspondiente a 700 sujetos de voz normal y de diferentes patologías de la base de datos de Massachusetts Eye and Ear Infirmary (MEEI), en donde utilizó coeficientes cepstrales, banco de filtros con la frecuencia mel, y clasificadores de cadenas de Markov y GMM. El análisis de las muestras dió como resultado la correcta clasificación con los coeficientes cepstrales de la frecuencia mel (MFCC) y la F_0, aplicando GMM de 99,97 % y con HMM la tasa fue de 99,40 % en entrenamiento, con parámetros en el dominio del tiempo los resultados de clasificación correcta se colocaron a un poco más de 50 %.

4. Conclusiones

Al analizar los resultados de esta investigación se tiene que la variación de la frecuencia fundamental es un parámetro que discrimina la calidad de la voz y, a pesar de la cantidad de métodos de cálculo, no se tiene en la actualidad un procedimiento que funcione sin dependencias de la señal de entrada. Se concluye que la frecuencia fundamental requiere un cálculo más robusto y con una precisión mayor.

En una gran parte de los métodos de cálculo de la frecuencia fundamental se realiza un filtrado con un filtro pasabajas de 1000 *Hz* o un poco menos de 1000 *Hz*, para resaltar la zona de baja frecuencia. Los parámetros propuestos para evaluar la calidad de la voz en su mayoría están en el dominio frecuencial. Los parámetros propuestos por Del Pino [18] se basan en la zona del espectro que presenta menos ruido y coincide con la de baja frecuencia. El modelado del comportamiento, como por ejemplo en la referencia [22], con un porcentaje tan elevado de

evaluaciones correctas de las patologías, usa los MFCC, que en su mayoría se encuentran en la zona de baja frecuencia; esto a nuestro juicio nos refuerza la idea de que se debe investigar aun más la zona de baja frecuencia del espectro.

Referencias

[1] Casado, J. y Adrián, J. La evaluación clínica de la voz. (1^a ed.) Ediciones Aljibe S. L. España, 2002

[2] Rabiner, L. Cheng M. Rosemberg A. and Mcgonegal C. A Comparative Performance Study of Several Pitch Detection Algorithms IEEE TRANSACTIONS ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, VOL. ASSP-24, No. 5, pp 399-418 OCTOBER 1976

[3] David Gerhard Pitch Extraction and Fundamental Frequency: History and Current Techniques Technical Report TR-CS 2003-06 November, 2003

[4] Sondhi, M. A New Methods of Pitch Extraction IEEE TRANSACTIONS ON Audio and Electroacoustics, VOL. Au-16, No. 2, pp 262-266 Junio, 1968.

[5] Schafer, W. y Rabiner, L. C. System for automatic formant analysis of voiced speech. J. Acoust. Soc. Amer., vol. 47, pp.634-648 Febrero, 1970.

[6] Noll, A. Cepstrum Pitch Determination. J. Acoust. Soc. Amer., vol. 41, Number 2, pp.293-309 Febrero, 1967.

[7] Oppenheim, A. Speech Analysis-Sinthesys System Based on Homomorphic Filtering. J. Acoust. Soc. Amer., vol. 45 Number 2, pp.458-465 Febrero, 1969.

[8] Marker, J. The SIFT Algorithm for Fundamental Frequency Estimation IEEE TRANSACTIONS ON Audio and electroacoustics VOL. Au-20, No. 5, pp 367-377 Diciembre, 1972

[9] Gold B. y Rabiner, L. C. Parallel Processing Techniques for Estimating Pitch Periods of Speech in the Time Domain. J. Acoust. Soc. Amer., vol. 46 No 2 Parte 2, pp.442-448 Agosto, 1969

[10] Shimamura T. y Kobayashi H. Weighted Autocorrelation for Pitch Extraction of Noisy Speech. IEEE Transactions on Speech and Audio Processing Vol. 9, No. 7, pp 727-730 Octubre, 2001

[11] CheveigneÁ. y Kawahara H. YIN a fundamental frequency estimator for speech and music. J. Acoust. Soc. Am. 111 (4), pp 1917-1930 Abril, 2002.

[12] Piszczalski M. y Galler B. Predicting musical pitch from component frequency ratios. J. Acoust. Soc. Am, 66(3):pp 710–720, Septiembre, 1979.

[13] Kreiman J., Vanlancker-Sidtis D. y Gerratt B. Defining and Measuring Voice Quality. From Sound to Sense: June 11- June 13, MIT, pp C-163 a C-168. 2004.

[14] Michaelis D., Gramss T. y Strube H. Glottal-to-Noise Excitation Ratio – a New Measure for Describing Pathological Voices ACÚSTICA- acta acústica Vol. 83, pp 700 – 706 1997

[15] de Krom. A Cepstrum –Based Technique for Determining a Harmonics-to-NoiseRatio in Speech Signal. Journal of Speech and Hearing Research, Volumen 36, pp 254-266 Abril, 1993.

[16] Michaelis D., Fröhlich M. y Strube H. Selection and combination of acoustic features for the descriptionof pathologic voices. J. Acoust. Soc. Am. 103 (3),:pp 1628–1639 Marzo, 1998.

[17] Casado J., Adrían J., Conde M., Piédrola D., Povedano V., Muñoz E., Cantillo E. y Jurado A. Estudio Objetivo de la Voz en Población Normal y en la Disfonía por Nódulos Y Pólipos Vocales. Acta Otorrinolaringol ; 52: pp 476-482. Esp 2001

[18] Del Pino P., Díaz J., Jiménez C. y Rothman H. Identificación de algunos parámetros espectrales que determinanla calidad de la voz. REVISTA INGENIERÍA UC. Vol. 11, No 3, pp 7-16 2004.

[19] Yu P., Ouaknine M., Revis J., y Giovanni A. Objective Voice Analysis for Dysphonic Patients: A MultiparametricProtocol Including Acoustic and Aerodynamic Measurements Journal of Voice. Vol. 15, No. 4, pp. 529–542. The Voice Foundation. 2001

[20] Kreiman J. y Gerratt B. Perception of aperiodicity in pathological voice Journal of Voice Obtenido en la Red el 01 de abril de 2008: http://repositories.cdlib.org/postprints/1117.

[21] Gerratt, B. R., y Kreiman, J. Measuring voice quality with speech synthesis The Journal of the Acoustical Society of America, 110, pp 2560–2566. 2001

[22] Dibazar A. y Narayanan S. A System for Automatic Detection of Pathological Speech Obtenido en la Red el 01 de abril de 2008: http://sail.usc.edu/publications/ASILOMAR2002-paper-ali.pdf.