Cálculo y Análisis del Pitch en Señales Sonoras de Voz Humana

Wainschenker R.S., Doorn J.H., Legrottaglie C. F., Castro M.

INTIA, Facultad de Ciencias Exactas, Universidad Nacional del Centro de la Provincia de Buenos Aires, Paraje Arroyo Seco, Campus Universitario (7000) Tandil Argentina.
e-mail: { rfw, jdoorn, clegrott, mcastro }@exa.unicen.edu.ar

Introducción:

El análisis del pitch involucra diferentes tópicos dentro del estudio de señales sonoras aún no explorados completamente. En virtud de la imprecisión en su propia definición se pueden implementar una gran variedad de algoritmos para su adquisición. Históricamente se ha definido al pitch como la frecuencia fundamental de espectro de frecuencias del habla [Casacuberta87] y se lo ha asociado al movimiento que realiza la glotis en la generación del sonido [Husson62]. Desafortunadamente cualquiera sea la forma en la que se lo defina no se ajustará a la realidad, porque la oscilación glotal es una función cuasi-periódica [Klatt87].

Además, esta frecuencia no es fácilmente identificable debido a que en algunas situaciones prácticamente desaparece de la onda sonora. Esto ocurre cuando las articulaciones del tracto vocal hacen que la energía del sonido se concentre en algunos de sus armónicos. No obstante no se lo pierde completamente y se puede utilizar dichos armónicos para su rastreo.

Se ha observado que esta vibración no es constante a lo largo del discurso, detectándose variaciones a lo largo de la frase y también dentro mismo de una palabra. Estas variaciones se deben tanto a la entonación de la frase, como a la acentuación de los fonemas así como al estado emocional del orador [Rocha87] [Klatt87].

El análisis del pitch es de fundamental importancia en el estudio de señales sonoras tanto para musicales como de voz humana. En el caso del análisis de la voz humana se relaciona tanto al reconocimiento de voz en forma computacional como a la síntesis robusta. Hoy en día los mejores sintetizadores de voz humana están basados en la concatenación de demi-fonemas, los cuales poseen buenos resultados [Sproat98]. También existen desarrollos de sintetizadores basados en la simulación del aparato fonador, especialmente, simulando los resonadores internos del tracto vocal. El valor del pitch en esta clase de sintetizadores es fijado en forma arbitraria alrededor de valores típicos según la voz que se quiera generar y las variaciones de este son generadas al azar dentro de un rango alrededor del valor fijado. Por lo tanto, al incorporarse un mayor conocimiento de las variaciones del pitch a esta clase de sintetizadores, se generaría un discurso mucho mas cercano a la verdadera voz humana [Klatt87].

En este trabajo se implementarán un conjunto de herramientas que permitan determinar los valores del pitch y sus variaciones a lo largo de un fonema. Esto va a permitir realizar un seguimiento del mismo tanto en una palabra como en una frase. Para lograr este objetivo se implementaran las técnicas de función de correlación, integración y distancia entre máximos, entre otras.

Marco teórico:

En la generación de voz intervienen la presión de aire que ejercen los pulmones, la apertura y el cierre de la glotis y los resonadores internos al tracto vocal. Dado que el pitch esta asociado al movimiento de la glotis es interesante observar los valores de la presión inmediatamente después de la misma. Esta tiene una forma

aproximadamente sinusoidal conformada por dos exponenciales, una creciente y otra decreciente, producto del pasaje de aire durante la apertura y cierre glotal como se detalla en la Fig.1 [LeHuche93].

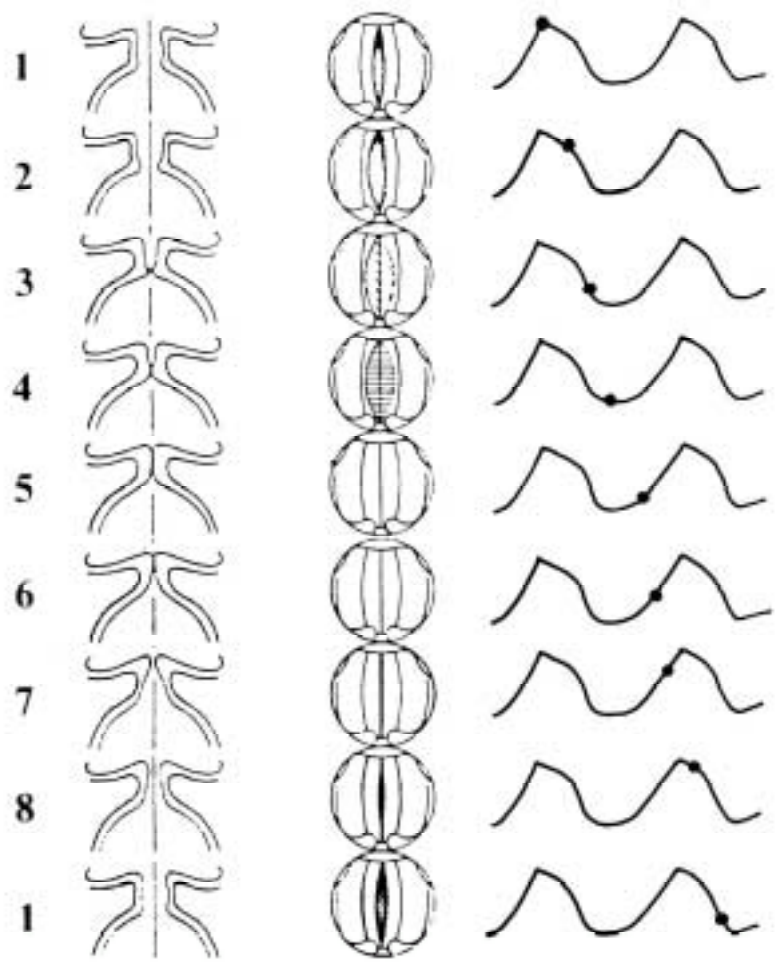

Fig.1 Esquema que representa los movimientos de los pliegues vocales y la ondulación durante cada ciclo vibratorio.

Como los resonadores del tracto vocal solamente refuerzan múltiplos de la frecuencia fundamental, con la que oscila la glotis, es entonces evidente que un cambio en esta ultima, no solo cambia el pitch, sino que también cambia la composición en frecuencias del sonido emitido. Dado que estas variaciones se detectan ciclo a ciclo, se produce un batido entre estas frecuencias ligeramente modificadas generando ondas de muy baja frecuencia cuyo período es tan largo que afecta varios ciclos de la señal vocal

Entre la gran cantidad de algoritmos para la determinación del pitch de ondas sonoras tanto musicales como de voz humana se encuentran métodos que trabajan tanto en el dominio del tiempo como en el dominio de la frecuencia. Entre los primeros podemos mencionar la función de auto-correlación (ACF), función de diferencia de magnitud promedio (AMDF), función de diferencia cuadrada promedio (ASDF), entre otros, entre los segundos se encuentran la transformada de Fourier discreta (DFT) y transformada rápida de Fourier (FFT), transformada de Fourier de corto tiempo discreta (DSTFT), transformada de constante Q, entre otros. En la recopilación realizada por Uppgård [Uppgård01] se muestra que ninguno de ellos a dado extraordinarios

resultados y ni se ha señalado como el mejor y presupone que sería posible mejorar los resultados para el rastreo del pitch con alguna combinación de diversos métodos.

La metodología de trabajo consiste en buscar un valor aproximado del pitch utilizando un procedimiento clásico como las transformadas de Fourier, tanto rápida como por definición [Hsu1970a], cuyos resultados serán utilizados como dato de partida en la aplicación de métodos de rastreo que se pasarán a detallar.

Se usará la función de Correlación en lugar de la función de auto-correlación. Esta última se basa en la correlación cruzada de la onda consigo misma. La modificación propuesta se basa en correlacionar la onda, no consigo misma, sino con funciones seno cuyas frecuencias y fases se toman como variable de trabajo, de esta manera se espera encontrar un valor de pitch para cada ciclo, independientemente de los calculados en los demás ciclos de la onda.

Otro método propuesto corresponde en la aplicación de la integración. Se presupone que la integral a lo largo de un período debería ser nula, debido a la simetría de la onda. En base a esto se procede a buscar los puntos en los que la integral de la misma adopta un mínimo.

Además se aplicará el método de distancias entre máximos, mínimos o ceros equivalentes.

Objetivos:

El objetivo general del trabajo es implementar un conjunto de herramientas que permitan estudiar las variaciones del pitch en ondas sonoras de voz humana.

- Que permitan realizar diversas mediciones que resulten en los valores del pitch de cada uno de los distintos ciclos de la onda a estudiar.
- Que permitan variar parámetros en los cuales basar los cálculos.
- Que permitan comparar la eficiencia de los resultados de los métodos aplicados.

Objetivos Específicos:

El trabajo consiste en implementar un soporte que maneje un conjunto de herramientas orientadas al tratamiento de señales sonoras. Estas herramientas pueden dividirse en los siguientes grupos:

a) Herramientas de pre-procesamiento: De centrado: centrado de señales en 0 voltios.

b) Herramientas de análisis: Las Transformada de Fourier FFT y por definición permite analizar el espectro de frecuencias presentes en la señal. Se permitirá la visualización de las amplitudes de la FFT en función de la frecuencia.

c) Herramientas de rastreo: Permiten obtener las variaciones del pitch de las ondas estudiadas variando parámetros y métodos de rastreo.

d) Herramientas de comparación: Permiten analizar las diferencias entre los resultados del estudio de ondas sonoras tanto de los distintos parámetros como de distintos métodos de rastreo.

A estas herramientas deben sumarse las que surjan durante el desarrollo del trabajo.

Estado de Avance:

Se han implementado las herramientas para realizar el rastreo automático del pitch por los métodos de mínimo de integración, distancia entre puntos equivalentes y correlación con senos de frecuencia y fase variables. Los cálculos se realizan luego de hacer un análisis clásico de frecuencias, de forma tal de orientar la elección de los valores iniciales para realizar el seguimiento y otros parámetros de cálculo.

Para el ajuste de la precisión de cada método se ha procedido de diferentes maneras, para ello se utilizan distintos métodos de integración numérica, interpolación, con la posibilidad de variar el rango de precisión de las distintas variables de trabajo.

Bibliografía:

[Casacuberta87] Casacuberta, F., Vidal, E. "Reconocimiento automático del habla" Biuleraux Editores (1987).

[Hsu70a] Hsu H. P. "Análisis de Fourier"Addison-Wesley Iberoamericana (1970).

[Husson62] Husson, Raoul "Physioligie de la Pnonation" Masson et cie Éditeurs (1962).

[Klatt87] Klatt, D. H. "Review o f test-to-speech conversion for English" J. Acoust. Soc. Am **82** (3) 737-793 (1987).

[Klatt90] Klatt, D. H. "Review of test-to-speech conversion for English" J. Acoust Soc. Am **87** (2) 820-857 (1987).

[LeHuche93] Le Huche F., Allali A. "La voz, Anatomía y fisiología de la voz y del habla" Masson (1993).

[López99] López M. R. "Ingeniería Acústica" Parainfo (1999).

[Rocha87] Rocha. L "Procesamiento de Sonido" Kapeluz (1987).

[Scheid93] Scheid F., Di Constanzo R. E. "Métodos Numéricos" Mc Graw Hill (1993).

[Sproat98] Sproat, R. "Multilingual Text-to-Speech Synthesis: The Bell Labs Approach" Kluwer Academic Publishers, London, (1998).

[Uppgård01] Uppgård, S. "Implementation and Analysis of Pitch Tracking Algorithms" Master of Science Thesis Project at Clavia and KTH S3 KTH, Stockholm, Sweden (2001).