DETECCIÓN DE PITCH EN CONDICIONES ADVERSAS

Ibon Saratxaga, Iker Luengo, Eva Navas, Inmaculada Hernáez, Jon Sánchez, Iñaki Sainz

Aholab – Escuela Técnica Superior de Ingeniería.
Universidad del País Vasco – Euskal Herriko Unibertsitatea
Urkijo zum. z/g 48013 Bilbo
ibon.saratxaga@ehu.es, ikerl@bips.bi.ehu.es, eva.navas@ehu.es, inma.hernaez@ehu.es, jon.sanchez@ehu.es, inaki@bips.bi.ehu.es

RESUMEN

La necesidad de herramientas de detección del pitch lo suficientemente robustas para funcionar en entornos ruidosos se ha visto acrecentada en los últimos tiempos debido a la aparición de nuevos sistemas de codificación de voz, reconocimiento, conversión etc. En este trabajo se presenta un algoritmo de detección basado en los coeficientes cepstrales combinados con un algoritmo de Viterbi, que mantiene cierta robustez en condiciones ruidosas. Este algoritmo ha sido evaluado mediante una base de datos preparada al efecto, y se detallan los resultados obtenidos comparándolos con los de otros algoritmos de uso general.

1. INTRODUCCIÓN

La detección y el marcado de pitch han sido necesidades importantes desde los inicios de la investigación en el área de voz. Si tradicionalmente la detección de pitch había sido esencial para el estudio y modelado de la entonación, más recientemente han surgido nuevas aplicaciones como el reconocimiento de emociones [1], el reconocimiento de lenguajes tonales [2], la conversión de voz [3], el reconocimiento [4] o la verificación de locutor [5] que han reavivado los requerimientos de sistemas de detección y marcado de pitch robustos que sean capaces de trabajar con señales obtenidas fuera de los estudios de grabación, por canales más ruidosos.

Con estos antecedentes, un grupo de trabajo enmarcado en el European Center of Excellence on Speech Synthesis (ECESS) (www.ecess.eu), promovió entre sus miembros, una campaña de evaluación de algoritmos de detección de pitch, tanto para señales de buena calidad, como para aquellas con una relación SNR peor. Para ello se preparó una base de datos de voz de referencia, marcada y revisada manualmente, para su utilización en la evaluación de algoritmos de marcado y detección de pitch [6]. Nuestro grupo de trabajo participó en esta evaluación con una versión mejorada de uno de los algoritmos de detección de pitch de que disponíamos. Los resultados se presentaron en la reunión del ECESS en Maribor (Eslovenia) el 5 de julio de 2006.

Este módulo de detección de pitch es el que se describe en el presente artículo. Los diferentes bloques que componen el mismo se detallan en el siguiente apartado. Después se describen las características de la evaluación y se da cuenta de los resultados obtenidos, incluyendo una somera comparación de éstos con otros resultados publicados. El documento finaliza con unas breves conclusiones y posibles mejoras.

2. MÓDULO DE DETECCIÓN DE PITCH

El módulo de detección de pitch (MDP) se encarga de determinar los valores del pitch de la señal de voz en cada momento, generando así los puntos de la curva de la evolución del pitch en el tiempo. Para obtener esta curva se utiliza un algoritmo de determinación de pitch basado en los valores de los coeficientes cepstrales, con selección de los mismos mediante el algoritmo de Viterbi. La curva así obtenida es postprocesada junto con la información sobre el carácter sordo o sonoro de cada trama, en un bloque de suavizado.

La estructura del módulo de detección de pitch se esquematiza en la figura 1. La única entrada que el módulo requiere es la señal de voz. Esto es importante, puesto que permite obtener las curvas de pitch directamente, sin necesidad de procesado previo para obtener alguna otra información (como podría ser la segmentación fonética por ejemplo), lo que posibilita calcular valores de pitch de cualquier señal desconocida.

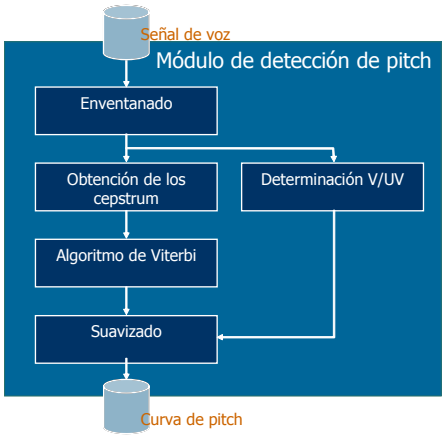

Figura 1. *Diagrama de bloques del MDP.*

La salida del algoritmo es un fichero con los valores de pitch a la frecuencia de muestreo establecida por el usuario, en formato PCM. Se detallan a continuación los pasos que se siguen para la determinación de estos valores.

2.1. Enventanado

El análisis de la periodicidad o aperiodicidad de la señal en cada momento debe realizarse sobre un fragmento de señal significativo. Por ello es necesario enventanar la señal tomando un fragmento lo suficientemente largo como para que la periodicidad sea detectable, pero lo suficientemente corto como para que no se pierda resolución temporal.

Para que la detección de periodicidad sea correcta se requiere que el enventanado contemple al menos dos periodos de pitch. Con esta idea, el tamaño de la ventana que se empleará queda fijado por un parámetro que es el pitch mínimo esperable en la señal. La ventana que se tomará será constante para todo el análisis y su longitud será de dos veces el periodo de pitch mínimo.

La ventana es de Hamming con lo que en realidad, debido a la atenuación de la misma en los bordes, la longitud efectiva de la trama será algo menor que dos periodos del pitch mínimo.

2.2. Determinación Sonoro/Sordo

La determinación de si una trama es sorda o sonora se hace evaluando la razón entre la potencia de la misma y su tasa de cruces por cero (Zero Crossing Rate, ZCR) también normalizada a la longitud de la trama. Para ello se calcula por un lado la potencia de la trama, como:

$$P_f = \frac{1}{N}\sum_{i=0}^{N-1}\left|x[i]\right|^2$$

Donde x[i] son cada una de las N muestras de la trama (incluye el efecto del enventanado por lo tanto).

La ZCR normalizada se basa en contar el número de veces que la señal cruza por cero en toda la longitud de la trama y dividir también entre el número de muestras de la trama.

Con este ratio, se trata de reforzar el efecto de dos medidas independientes relacionadas con el carácter sordo o sonoro de la señal de voz. En efecto, la tasa de cruces por cero es en general mayor para los sonidos sordos, que suelen tener componentes a más altas frecuencias. La energía o la potencia, por el contrario tiende a ser menor para los sonidos sordos. Por ello, calcular el ratio energía entre ZCR hace que ambos efectos se refuercen: para tramas sordas el valor será pequeño, y será mucho mayor (4 órdenes de magnitud aprox.) cuando la trama sea sonora.

2.3. Cálculo de los coeficientes cepstrales

El primer paso para determinar el valor de pitch de la trama es calcular sus coeficientes cepstrales. Recordemos que los cepstrum se calculan tomando logaritmos sobre la transformada de fourier de la señal y luego aplicando su transformada inversa:

$$\hat{x}[n] = TF^{-1}\{\log(X(\Omega))\}$$
$$\text{con } X(\Omega) = TF\{x[n]\}$$

Los valores de pitch están limitados fisiológicamente, por lo cual no es necesario examinar todos los coeficientes cepstrum. Por ello en primer lugar se calcula un intervalo limitado por las frecuencias mínima f_{min} y máxima f_{max} entre las que se considera que se moverá el pitch y sólo se buscarán los máximos entre los coeficientes de ese intervalo.

El siguiente paso, una vez definido el conjunto de coeficientes en los que se centrará la búsqueda, es encontrar cuáles son los máximos. Para ello, se normalizan todos los coeficientes al valor medio del conjunto de los coeficientes de la trama:

$$c'_i = \frac{c_i}{\bar{c}} \quad / \quad \bar{c} = \frac{1}{i_{\max} - i_{\min} + 1}\sum_{i=i_{\min}}^{i_{\max}} c_i$$

El uso de cepstrum normalizados permite independizar sus valores de factores como la energía de la señal en cada punto, produciendo una escala coherente a lo largo de toda la señal. Esto permitirá que se pueda definir un umbral de tal forma que si un coeficiente cepstrum es n veces mayor que la media de la trama, entonces se considerará que la trama es sonora, mientras que por el contrario, si el valor no supera ese umbral, entonces se considerará que la trama es sorda.

En cualquier caso, en este bloque lo único que se hace es calcular los cepstrum, normalizarlos respecto de la media de la trama y seleccionar los M coeficientes mayores, futuros candidatos para el algoritmo de Viterbi. La decisión final sobre si el tramo es sordo o sonoro se tomará en la etapa de Viterbi. Además de los M mayores coeficientes se añade otro más, la "no-frecuencia", correspondiente a la decisión de que la trama fuera sorda.

2.4. Algoritmo de Viterbi

Una vez que se han calculado el conjunto de M+1 candidatos para todas las tramas que forman la señal, se deben elegir aquéllos que formen una mejor curva de pitch. Para ello, como ya se ha comentado, se utiliza el algoritmo de Viterbi, que se basa en escoger aquellos valores que hagan que se minimice la suma de dos funciones de coste: el coste local (el coste de seleccionar un candidato por sí mismo, independientemente de los candidatos adyacentes); y el coste de transición (el coste de seleccionar un candidato teniendo en cuenta el candidato seleccionado para la trama anterior).

Estos costes se definen de forma que reflejen las características de la curva de pitch. El coste local se define mediante dos términos. Para el primero, se tiene en cuenta que el candidato más probable para la frecuencia fundamental de una trama es aquél cuyo coeficiente cepstrum asociado tiene el máximo valor.

Por ello, en términos de coste se define una función inversamente proporcional al valor del coeficiente cepstral.

El segundo término del coste local se fundamenta en que, para considerar la trama como sonora, el valor de los cepstrum debe superar un umbral respecto al valor medio de los cepstrum de su trama. Si no es así, probablemente la trama será sorda y el máximo será aleatorio. Así, se define un coste fijo en el caso de que se viole este criterio, es decir, que se considere una trama sonora a pesar de que no supere el umbral, o a la inversa, sorda aún superándolo.

El coste de transición se forma también mediante dos términos. El primero de ellos se basa en el criterio de que la parte sonora de la curva de pitch debe ser continua y sin saltos bruscos, pues estos se deberían probablemente a que se ha seleccionado un armónico o sub-armónico del valor correcto de pitch. Así pues se define un coste proporcional a la relación entre las frecuencias de pitch de una trama respecto a la anterior. Cuanto mayor sea esta diferencia mayor será el coste.

Finalmente, se define un segundo término de coste de transición, para recoger el criterio de que es muy poco probable que ocurran transiciones rápidas entre regiones sordas y sonoras. Es decir, es poco probable que pasemos de una región sorda a una sonora y rápidamente a otra sorda. En este caso, es muy probable que una región sonora (o sorda) suficientemente pequeña sea en realidad un error de decisión sordo-sonoro. Así se define un coste fijo que penaliza las transiciones sordo-sonoro y sonoro-sordo.

Con estas funciones el algoritmo de Viterbi encuentra la curva que menor coste acumulado arroja. Sin embargo, esta curva podría contener valores espurios de pitch y además no se ha tenido en cuenta la información proveniente del detector sordo/sonoro. De esto se encarga el último bloque.

2.5. Suavizado

La curva a la salida del algoritmo de Viterbi puede contener valores erróneos sobre todo en puntos en los que la periodicidad es débil. En estos puntos los cepstrum habrán estado cerca del umbral de detección y puede que se haya tomado un máximo erróneo. Para estos casos el algoritmo de detección sordo o sonoro proporciona información no relacionada con los cepstrum, que puede ser útil para dilucidar el carácter de la trama.

Por otro lado, es conocido [7] que el pitch de un locutor puede ser modelado por una distribución log-normal. Aquellos valores que se aparten en demasía de la media de esta distribución pueden considerarse como errores en la detección de pitch. Cuanto más alejados estén los valores de pitch de la media, más seguro será que son erróneos.

Por ello, existe un parámetro N que indica el umbral a partir del cual se considerará que un valor está fuera de rango. Un valor de la curva de pitch se considera fuera de rango si está alejado del valor medio N desviaciones típicas de la distribución log-normal.

Para aplicar este criterio, lo primero que se debe hacer es calcular la media y la desviación típica de la distribución log-normal que modela el pitch de la curva obtenida del Viterbi. Para ello se calcula el logaritmo de todos los valores de pitch de la curva y con estos valores escalados se obtiene su media y su desviación típica.

El bloque de suavizado comprueba en primer lugar si el valor está o no fuera de rango, y si está fuera de rango lo elimina, marcando la trama como sorda. En un paso posterior, se coteja el carácter sordo o sonoro de la trama según sale del Viterbi tras eliminar los valores fuera de rango con el resultado del detector sordo-sonoro. Así si el detector sordo sonoro indica que una trama es sonora y Viterbi dice que es sorda, se da preferencia al detector de sordo sonoro y se marca como sonora interpolando un valor de pitch para la misma, tomando para ello las muestras de pitch más cercanas e interpolando linealmente.

Si el detector sordo-sonoro dice que una trama es sorda y el algoritmo de Viterbi ha propuesto un valor de pitch y éste está dentro del rango, entonces tiene preferencia el algoritmo de Viterbi.

3. EVALUACIÓN

La evaluación de la eficiencia de estas herramientas requiere comparar sus resultados para un conjunto de señales en las que se conozca sus valores de pitch. El consorcio ECESS promovió una campaña de evaluación de herramientas para detección y marcado de pitch, y proporcionó una base de datos marcada manualmente. A continuación describiremos las características de esa base de datos para pasar luego a detallar los resultados de las evaluaciones.

3.1. Base de datos de evaluación

La base de datos que se ha utilizado para la evaluación consiste en un subconjunto de la base de datos SPEECON Spanish. Las bases de datos SPEECON fueron grabadas siguiendo las especificaciones del proyecto SPEECON de la Comisión Europea [8], y buscaba obtener grabaciones de señales de voz en diferentes entornos acústicos para su uso en reconocimiento, básicamente.

Las señales se adquirieron simultáneamente por cuatro canales utilizando diferentes micrófonos, en diferentes ambientes acústicos (coches, oficinas, lugares públicos, etc.). Así, el primero de los canales (C0) se grabó con un micrófono acoplado a unos auriculares, el canal C1 se adquirió con un micrófono Lavalier o de corbata, el C2 con un micrófono direccional situado a 1 metro del locutor y el C3 se grabó con uno omnidireccional situado a 2-3 metros del locutor.

En estas condiciones las relaciones de señal a ruido de los diferentes canales son muy diferentes. Así, el canal C0 tiene SNR's de unos 30 dB, y en el otro

extremo el canal C3, con un micro omnidireccional, tiene relaciones mucho más bajas, del orden de 0 dB's.

Para preparar la base de datos de referencia [6] se escogieron frases de 60 locutores, 30 hombres y 30 mujeres, desde 19 a 79 años, con grabaciones de 1 minuto por locutor, lo que hace un total de 60 minutos por cada canal. Las frases provienen de diferentes tipos de corpus desde el punto de vista de su dominio semántico.

Las señales del canal C0, las más libres de ruido, se marcaron a periodo de pitch mediante un sistema automático. Las marcas del sistema automático fueron revisadas manualmente en su totalidad, corrigiéndose cuando fue necesario. De estas marcas corregidas se derivaron los valores del pitch también para el canal 0, obteniéndose valores cada milisegundo.

Dado que las grabaciones se realizaron simultáneamente para todos los canales, las marcas y los valores de pitch serán idénticos para todos ellos, con la única diferencia que estarán más o menos retardados debido a las diferentes distancias que tiene que recorrer el sonido para llegar a cada micrófono. Con el fin de utilizar las mismas referencias para todos los canales, se compensaron estos retardos realineando las grabaciones, mediante correlación cruzada de cada canal con el C0.

3.2. Criterios de evaluación

Una vez descritas las características de la base de datos de referencia, veremos en este apartado cuáles son las magnitudes que nos van a permitir estudiar el rendimiento de nuestro módulo de detección de pitch. Para ello se definieron algunas medidas habituales en la literatura sobre estos temas [9].

- Valores erróneos por exceso y por defecto: Estas dos tasas de error miden el porcentaje de los valores de pitch correspondientes a segmentos sonoros cuyo valor supera, o queda por debajo, en más de un 20% el valor de pitch correcto. A estas medidas se les llama en inglés Gross Error High (GEH) y Gross Error Low (GEL), respectivamente, y se suelen visualizar acumulados, para dar una idea de la tasa de error total. Nótese que las tramas sordas o silenciosas (que no tienen ningún valor de pitch) no computan para este error.
- Porcentaje de error de tramas sonoras y de tramas sordas: El porcentaje de error de tramas sonoras (voiced error o VE) mide el porcentaje de las tramas sonoras que han sido erróneamente clasificadas como sordas, respecto del total de tramas sonoras de la BD. El porcentaje de error de tramas sordas (unvoiced error o UE) es, análogamente, el porcentaje de las tramas sordas, erróneamente clasificadas como sonoras, respecto al total de tramas sordas.
- Diferencias en la media y la desviación estándar: La última pareja de magnitudes que mediremos será la diferencia entre la media y la desviación estándar de todos los valores de pitch estimados y las de todos los datos de referencia.

3.3. Resultados

Los resultados que se muestran a continuación, buscan un equilibrio que dé resultados aceptables en todas las categorías y canales, lo que significa que a lo largo de las pruebas se vieron configuraciones de parámetros que hubieran permitido mejorar un canal o criterio específico. Por ello, los valores que aquí se proponen para los parámetros pueden mejorarse para adaptarse a las características de las señales que se vayan a analizar en un problema concreto.

Los resultados así obtenidos son los mostrados en la tabla 1:

	C0	C1	C2	C3
VE(%)	9,94	22,62	28,41	35,62
UE(%)	7,44	7,35	6,93	7,35
GEH(%)	0,65	0,31	0,57	0,97
GEL(%)	1,99	2,38	2,45	2,40
AbsMeanDiff(Hz)	0,54	1,87	7,57	11,92
AbsStdDiff(Hz)	1,45	4,46	4,86	6,03

Tabla 1. *Resultados finales*

Para analizar estos resultados los dividiremos en tres grupos. Por un lado tenemos los dos indicadores de error de trama sorda/sonora, UE y VE (figura 2). Como era previsible, el bloque específico de detección por Pot/ZCR era muy sensible al ruido, por lo que se deshabilitó para los canales ruidosos. De esta forma la discriminación sordo/sonoro para la trama se obtiene de los cepstrum en C1, C2 y C3. El resto de los ajustes de parámetros son comunes para todos los experimentos y canales.

En cualquier caso, la precisión de la clasificación se degrada rápidamente con el ruido del canal, y los resultados no son muy espectaculares, aunque, como se verá en los datos comparativos, la clasificación obtenida es de las mejores de entre los algoritmos comparados.

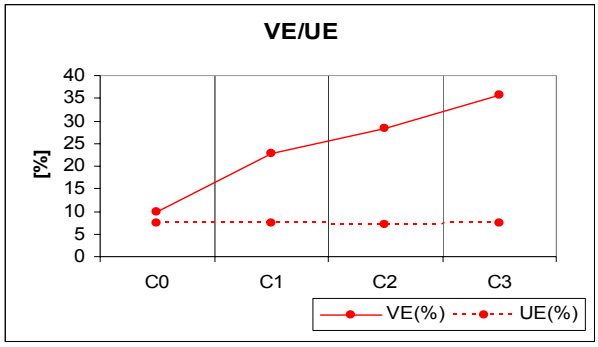

Figura 2 *Evolución VE/UE respecto del canal*

Es necesario hacer notar que la alta frecuencia del análisis, con tramas de 1 ms. hace que surjan una serie de cuestiones sobre los criterios para clasificar una trama como sonora. Debe tenerse en cuenta que 1ms. es un periodo entre 4 y 10 veces menor a un periodo de

pitch, por lo que cabe preguntarse al principio o al final de un fragmento periódico en qué punto del primer periodo de pitch que aparece (o en el último en que se desvanece), se debe comenzar a (o dejar de) considerar periódica la trama.

La siguiente pareja de datos es la formada por los errores en los valores de pitch por exceso (GEH) y por defecto (GEL). Los resultados en estas dos magnitudes son buenos, y lo son para todos los canales, empeorando sólo ligeramente al aumentar el ruido: su suma se mantiene siempre alrededor del 3% (ver figura 3).

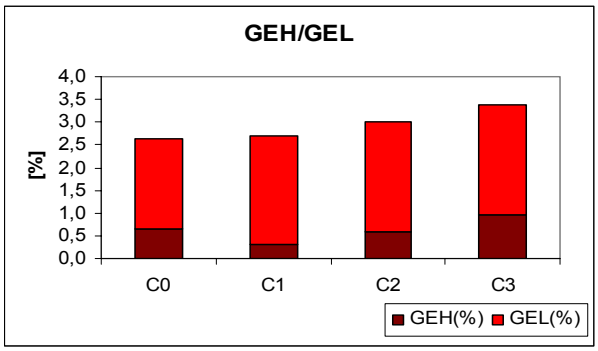

Figura 3. *Evolución GEH/GEL respecto del canal*

Las dos últimas medidas, la diferencia entre las medias y las desviaciones típicas de los valores de pitch de referencia y estimados, han resultado más afectadas por el ruido del canal, como se puede ver en la figura 4.

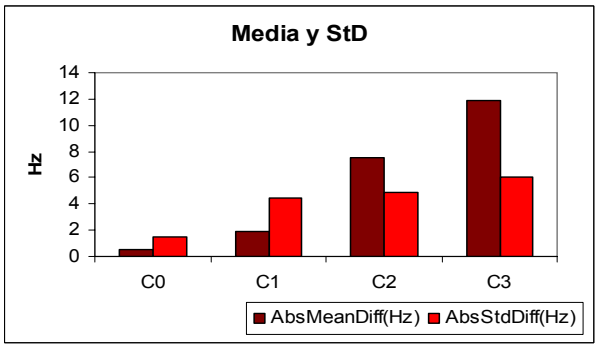

Figura 4. *Diferencias de media y desviación típica por canal*

3.4. Resultados comparativos

Para tener elementos de comparación contra los que validar la bondad de los resultados que estábamos obteniendo durante los experimentos, llevamos a cabo los mismos experimentos utilizando el algoritmo de autocorrelaciones del Praat [10], en su configuración por defecto. Este método consiste en calcular la autocorrelación de una trama de la señal con más de un periodo de pitch, lograda mediante enventanado. En la versión implementada en el Praat la autocorrelación está dividida por la autocorrelación de la ventana, para evitar el efecto distorsionador que ésta pudiera tener, según se describe en [11]. Además en [6] se habían publicado los resultados de otros dos algoritmos: el de los autores (KOT), basado en el cálculo de la transformada Hilbert de los residuos LPC; y el de Goncharoff [12], (GON), basado en la detección de la periodicidad en la energía de la señal a corto plazo, con programación dinámica para seleccionar los valores correctos.

Se muestran a continuación los resultados obtenidos por nuestro módulo comparándolo con el resto de los valores publicados.

3.4.1. Errores acumulados de trama sorda y sonora

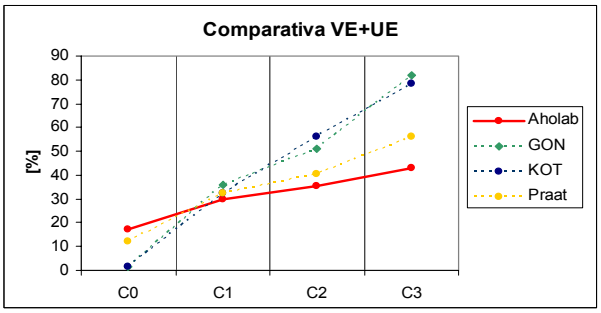

Figura 5. *Comparativa VE+UE, todos los canales*

Se ha comentado ya la relativa debilidad de los resultados de estos errores para el canal 0, sin ruido. Los algoritmos específicos en el dominio temporal, para la detección sordo-sonoro dan mejores resultados para este canal. Sin embargo, para los canales más ruidosos, los algoritmos temporales basados en ZCR empeoran rápidamente, mientras que la detección basada en cepstrum es considerablemente mejor (figura 5).

3.4.2. Errores acumulados de valor de pitch por defecto y por exceso

Los resultados de los errores en el valor del pitch estimado son en todos los casos los mejores de la comparativa (figura 6). Como hemos comentado son además buenos para todos los canales, por lo que su diferencia respecto al resto de los métodos considerados aumenta a medida que se incrementa el ruido del canal.

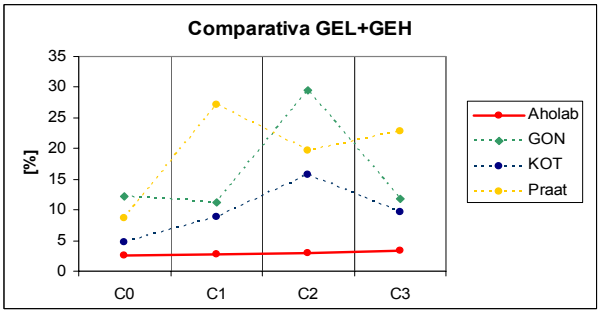

Figura 6. *GEL+GEH, todos los canales*

3.4.3. Diferencias en los estadísticos

Las diferencias en las medias comienzan siendo prácticamente despreciables para el canal 0 y van incrementándose con el ruido (figura 7). En nuestro caso el incremento es, sin embargo, menor que en los restantes algoritmos.

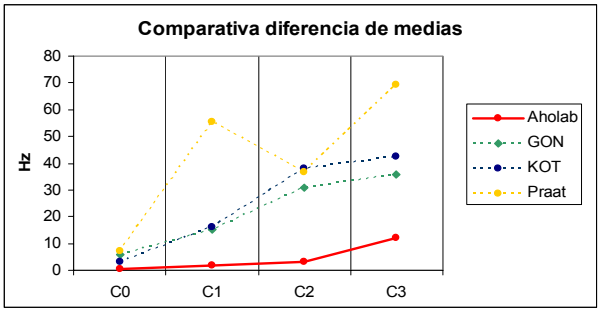

Figura 7. *Diferencia medias, todos los canales*

Respecto a la desviación estándar (figura 8) los valores se van incrementando en todos los casos a medida que sube el ruido, aunque es de resaltar la suave progresión que tiene en nuestro caso, lo que hace que obtengamos mejores resultados de nuevo para los canales ruidosos.

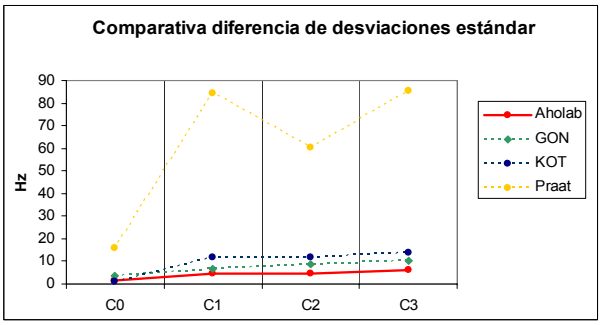

Figura 8. *Dif. desviación estándar, todos los canales*

Queda finalmente comentar en estos dos criterios de evaluación los erráticos resultados del algoritmo de Praat para los canales con ruido. Sin duda su aplicación para este tipo de señales requerirá ajustes en los parámetros por defecto, que no olvidemos que son los que se han utilizado en esta comparativa.

4. CONCLUSIONES

Las conclusiones de la evaluación del algoritmo de detección propuesto son muy prometedoras. El algoritmo ofrece unos resultados buenos en condiciones de bajo nivel de ruido; y es notoriamente robusto en condiciones ruidosas, ofreciendo el mejor rendimiento entre los algoritmos analizados.

Además el algoritmo participó en la campaña de evaluación del ECESS, con muy buenos resultados. Nuestra herramienta de detección de pitch obtuvo los mejores resultados en 19 de los 32 puntos de evaluación considerando todos los canales. Fue la mejor de las presentadas en la exactitud de los valores de pitch, para todos los canales; y obtuvo los mejores resultados en casi todas las medidas en los canales ruidosos.

Como continuación a estos trabajos, se pretenden desarrollar algunas ideas sobre posibles mejoras:

- Estudio de técnicas de detección del carácter sordo o sonoro de las tramas, que es la parte más débil del algoritmo: autocorrelación, filtrado paso bajo de ruido,…
- Inclusión de la información del bloque de detección sordo/sonoro en las funciones de coste del algoritmo de Viterbi.
- Comparación con otros algoritmos de detección y evaluación con otras bases de datos estándar.

5. BIBLIOGRAFÍA

[1] E. Navas, I. Hernáez, I. Luengo, J. Sánchez, I. Saratxaga. "Analysis of the Suitability of Common Corpora for Emotional Speech Modeling in Standard Basque". Lecture Notes in Artificial Intelligence, LNAI 3658, pp. 265-272, 2005.

[2] H.C.H. Huang, F. Seide. "Pitch tracking and tone features for Mandarin speech recognition". Procs. ICASSP 2000. Estambul. pp. 1523 - 1526 vol.3. Junio 2000.

[3] H. Ney, D. Suendermann, A. Bonafonte, H. Hoege. "A first step towards text-independent voice conversion". Procs. INTERSPEECH 2004, Jeju, Corea. pp. 1173-1176. Octubre 2004.

[4] S. Kim, T. Eriksson, H.G. Kang, D.H. Youn. "Pitch Synchronous Feature Extraction Method for Speaker Recognition". Procs. ICASSP 2004. Montreal. pp. 405-408. May 2004.

[5] I. Luengo, E. Navas, I. Hernáez. "Effectiveness of Short-Term Prosodic Features for Speaker Verification". Procs. The Fundamentals of Verbal and Non-verbal Communication and the Biometrical Issue. Vietri sul Mare, Italia. Septiembre 2006

[6] B. Kotnik, H. Höge, Z. Kacic. "Evaluation of Pitch Detection Algorithms in Adverse Conditions". Procs. 3rd International Conference on Speech Prosody, Dresden, Alemania, pp. 149-152, Mayo 2006.

[7] M. K. Sonmez, L. Heck, M. Weintraub, E. Shriberg. "A lognormal tied mixture model of pitch for prosody-based speaker recognition". Procs. EUROSPEECH '97. Rodas, Grecia. vol. 3, pp. 1391-1394. Septiembre 1997.

[8] D.J. Iskra et al. "SPEECON - Speech Databases for Consumer Devices: Database Specification and Validation". Procs. LREC'2002. Las Palmas de Gran Canaria. pp. 329-333. Junio 2002.

[9] X. Sun. "Pitch Determination and Voice Quality Analysis Using Subharmonic-To-Harmonic Ratio". Procs. ICASSP 2002. Orlando, EEUU. pp. 333-336. Mayo 2002.

[10] P. Boersma, D. Weenink. Praat: doing phonetics by computer (Version 4.3) [Computer program]. Retrieved from http://www.praat.org/

[11] P. Boersma. "Accurate short term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound". Procs. Institute of Phonetic Sciences 17. Univ. Amsterdam. pp. 97-110. 1993.

[12] V. Goncharoff, P. Gries. "An Algorithm for Accurately Marking Pitch Pulses in Speech Signals". IASTED International conference SIP '98. Nevada, USA. 1998.