Evaluation of the Effects of Speech Enhancement Algorithms on the Detection of Fundamental Frequency of Speech

Nicanor García Ospina, Juan Camilo Vásquez Correa, Jesús Francisco Vargas Bonilla, Juan Rafael Orozco Arroyave, Julian David Arias Londoño

Facultad de Ingeniería
Universidad de Antioquia

nicanor.garcia@udea.edu.co

18 de septiembre de 2014

UNIVERSIDAD DE ANTIOQUIA
1803

SISTEMIC
Sistemas Embebidos e Inteligencia Computacional

Contenido

UNIVERSIDAD DE ANTIOQUIA 1803

SISTEMIC
Sistemas Embebidos e Inteligencia Computacional

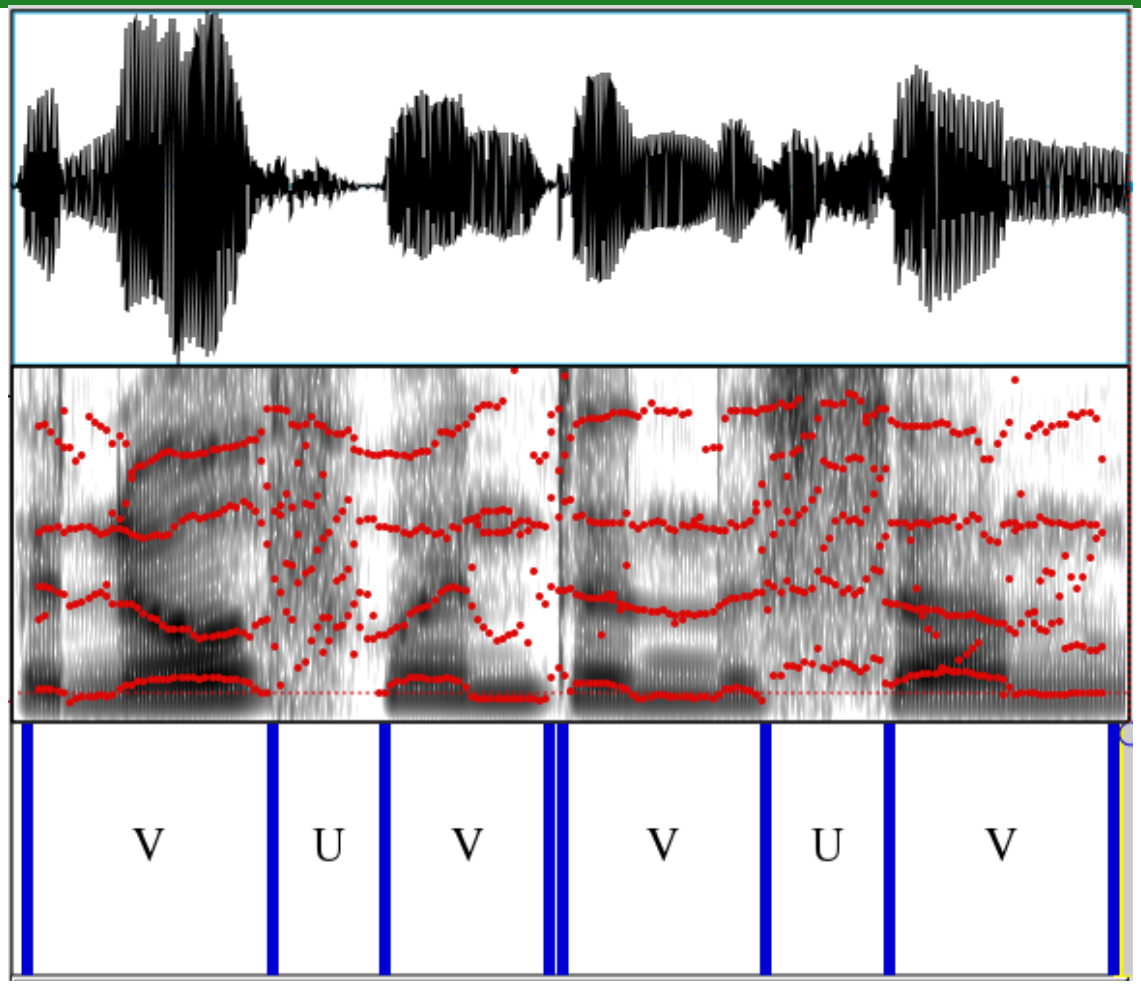
Introducción

Dos tipos de sonidos [1]:

- **Vocálicos** (*voiced*).
- **No-vocálicos** (*unvoiced*).

Segmentación v/uv:

- Segmentación automática [2].
- Mejor desempeño [3, 4].



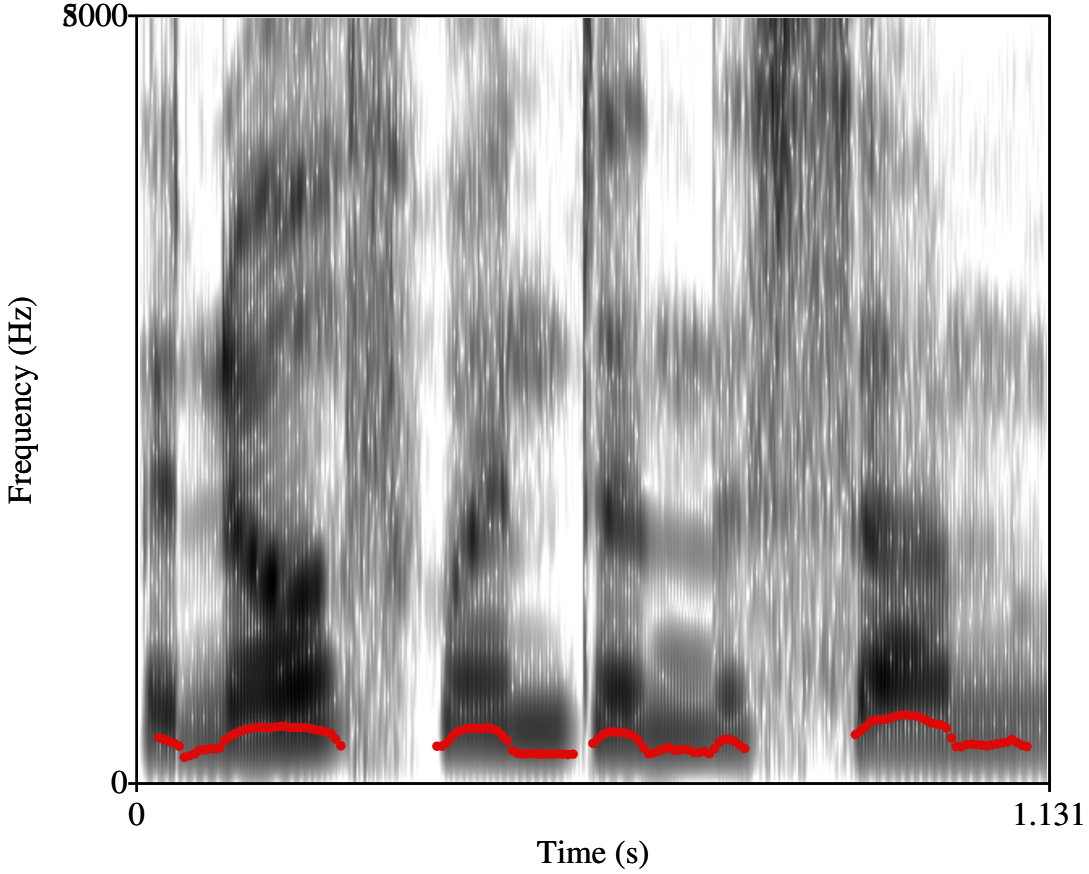
UNIVERSIDAD DE ANTIOQUIA 1803

SISTEMIC
Sistemas Embebidos e Inteligencia Computacional

Introducción

Pitch: Frecuencia de vibración de los pliegues vocales [2].

- Su presencia es necesaria para que un sonido sea vocálico.
- Se puede usar para realizar **segmentación v/uv**.
- Independiente del hablante y del idioma [1].



UNIVERSIDAD DE ANTIOQUIA 1803

SISTEMIC
Sistemas Embebidos e Inteligencia Computacional

Problema del ruido

En ambientes no-controlados:

- La **segmentación v/uv** y la **detección del pitch** se ven afectadas por ruido aditivo [2].

Para remediar esto se debe:

- Pre-procesar la señal.
- Técnicas de *Speech Enhancement*.
- **Su efecto no ha sido evaluado a profundidad.**

UNIVERSIDAD DE ANTIOQUIA 1803

SISTEMIC Sistemas Embebidos e Inteligencia Computacional

Objetivo

- Evaluar de manera objetiva el efecto de las técnicas de *Speech Enhancement* en el desempeño de sistemas para detección del pitch y segmentación v/uv.

UNIVERSIDAD DE ANTIOQUIA 1803

SISTEMIC Sistemas Embebidos e Inteligencia Computacional

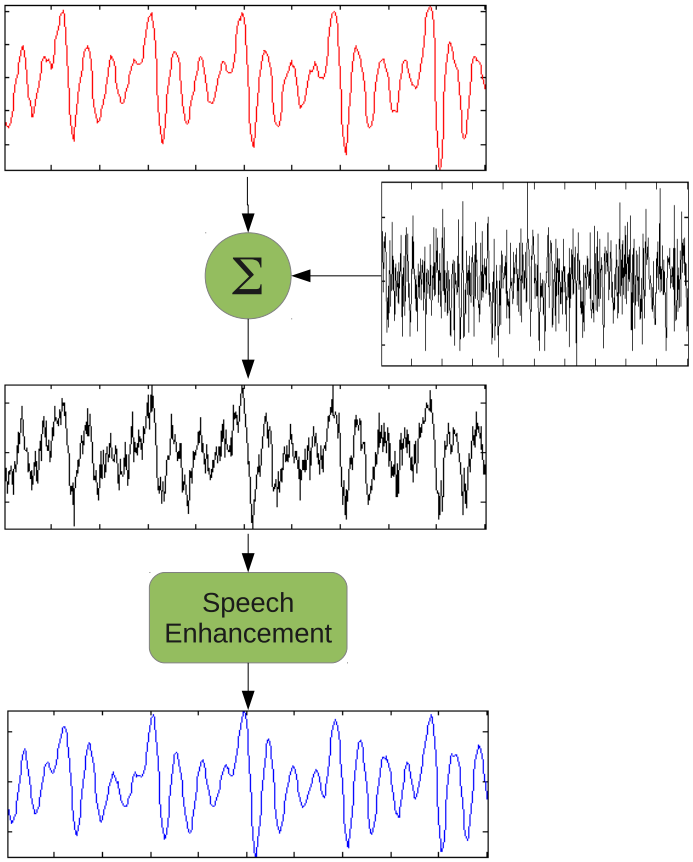
Speech Enhancement

Técnicas de eliminación de ruido diseñadas para la señal de voz [5].

- Buscan un estimador de la señal de voz limpia.
- Su objetivo es la calidad e inteligibilidad.
- Mejor desempeño en algunos sistemas [6, 7]

Se pueden dividir en cuatro tipos de algoritmos:

- Sustracción Espectral (Spectral Substraction, SS)
- Filtros de Wiener (Wiener Filter, WF)
- Basadas en modelos estadísticos (ME)
- Subespacio (Subspace)



UNIVERSIDAD DE ANTIOQUIA 1803

SISTEMIC
Sistemas Embebidos e Inteligencia Computacional

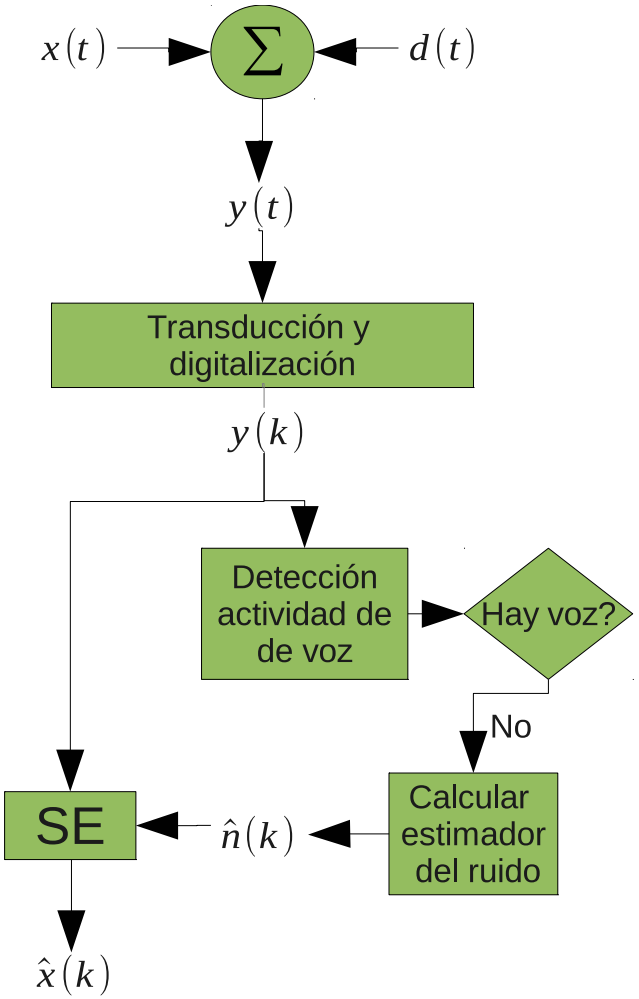
Modelo de señal

Se utiliza el siguiente modelo para la señal contaminada [5]

$$y(n) = x(n) + d(n) \qquad (1)$$

donde

- $x[n]$ es la señal de voz limpia.
- $d(n)$ es el ruido.
- Se asume independencia y no-correlación.



UNIVERSIDAD DE ANTIOQUIA 1803

SISTEMIC Sistemas Embebidos e Inteligencia Computacional

Sustracción espectral

Ventajas:

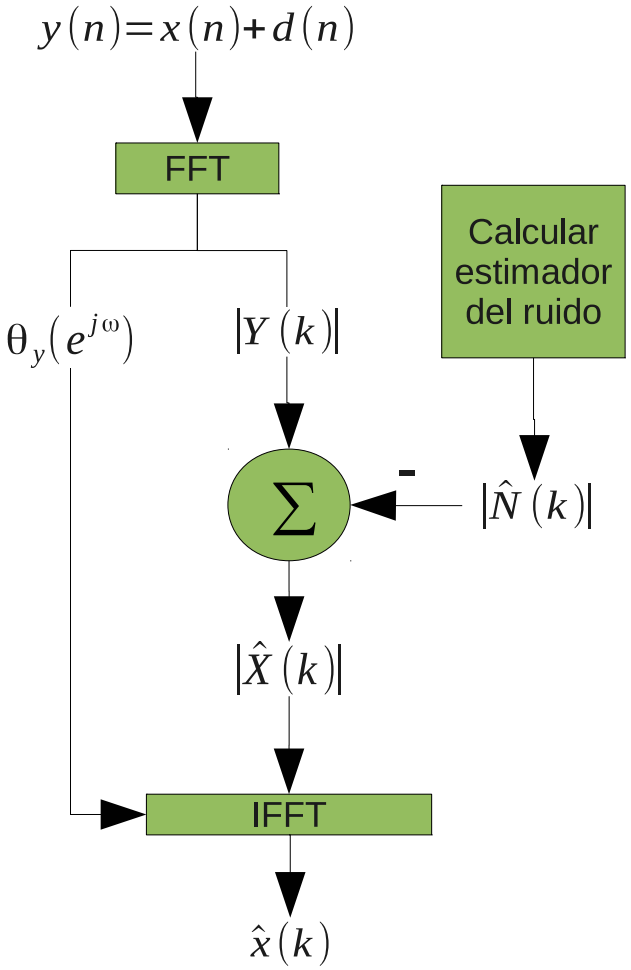
- Muy simple computacionalmente.
- Relativamente buena efectividad.
- Muchas versiones y mejoras [5].

Desventajas:

- Produce *ruido musical*[8].

Versiones evaluadas:

- Sustracción Espectral con sobre-sustracción (**SSBerouti**) [9].
- Sustracción Espectral Multi-Banda (**SSMB**) [10].



Filtro de Wiener

Es un estimador lineal del espectro complejo de la señal limpia [5].
Ventajas:

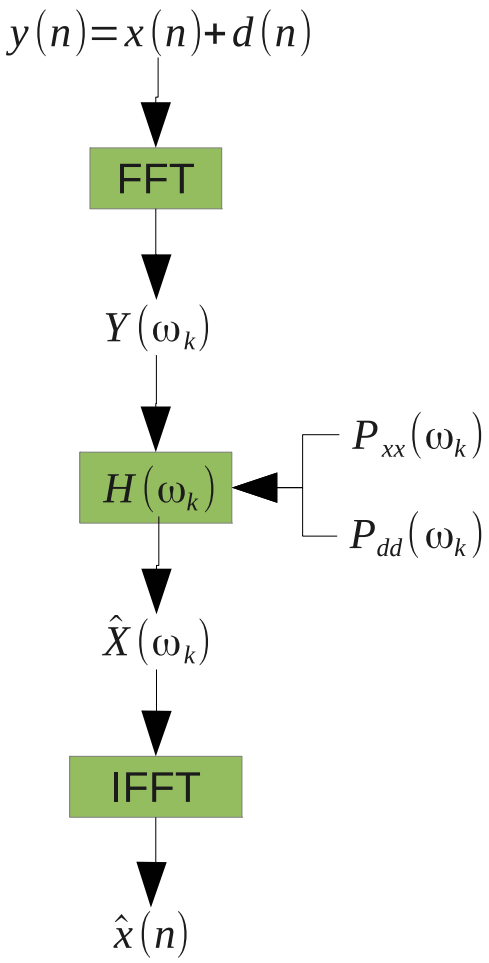
- Es óptimo en términos del mínimo error cuadrático medio.
- Poca distorsión de la señal.

Desventajas:

- Elimina poco ruido a SNRs bajas.
- No es posible calcular directamente $H(\omega_k)$.

Versión evaluada:

- Estimación de la SNR *a priori* (**WF**) [11].



Métodos basados en modelos estadísticos

Desarrollados a partir de modelos estadísticos de los espectros de magnitud de la señal y el ruido.
Ventajas:

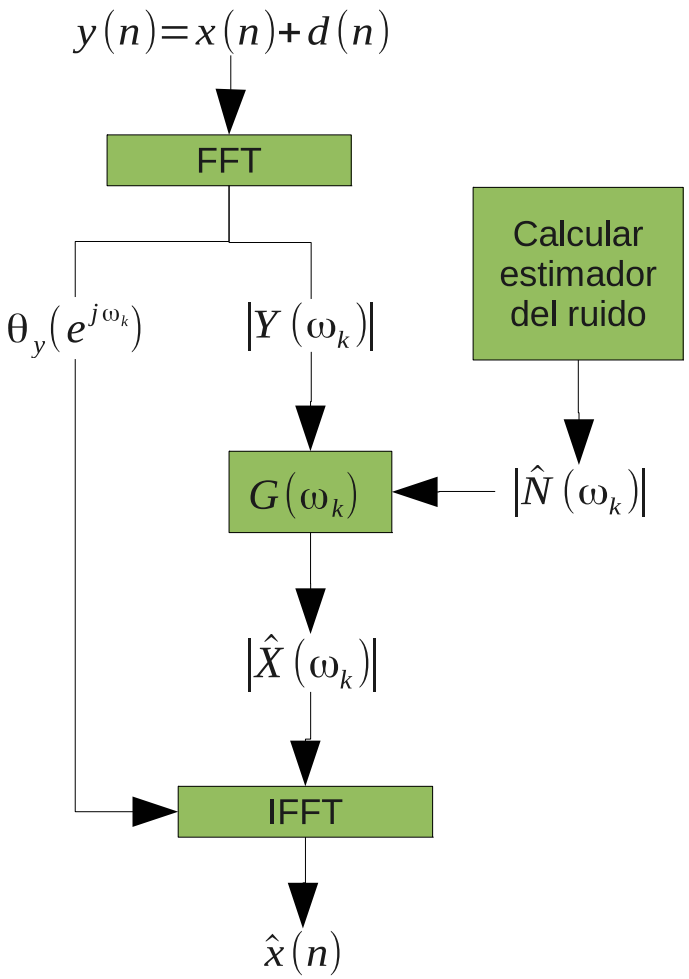
- Estimadores no lineales [5].
- Disminuyen el *ruido musical* [12].

Desventajas:

- Computacionalmente complejos[13].

Estimadores evaluados:

- *Minimum Mean Square Error* (**MMSE**) [14].
- *Log-spectra MMSE* (**LogMMSE**) [15].



UNIVERSIDAD DE ANTIOQUIA 1803

SISTEMIC Sistemas Embebidos e Inteligencia Computacional

Algoritmos de subespacio

La señal contaminada define un espacio vectorial que se puede descomponer en:

1. **Sub-espacio de señal.**
2. **Sub-espacio de ruido.**

Ventajas:

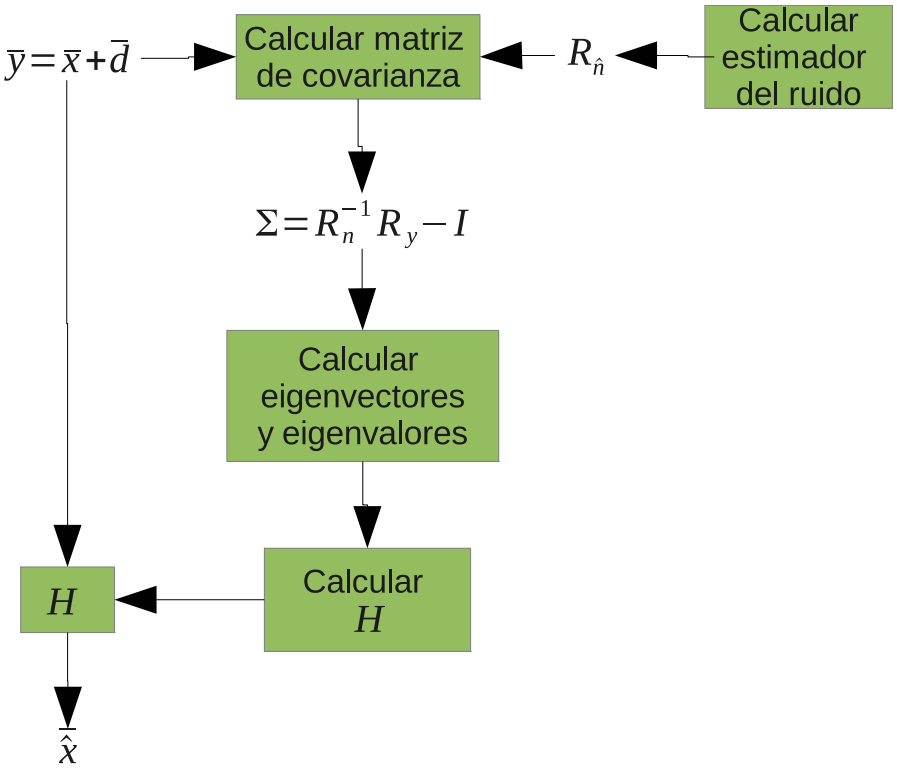
- Transformada dependiente de la señal.
- Disminución de ruido musical.

Desventajas:

- Computacionalmente complejo.

Versión evaluada:

- Enfoque de sub-espacio generalizado (**KLT**) [16].



Métodos de detección del *pitch* y segmentación v/uv

En este trabajo se escogieron dos de los algoritmos más usados [17, 18]:

- Método modificado de la autocorrelación implementado en Praat [19].
- Método Suma de armónicos de la onda residual (SRH) [20].

Ambos algoritmos cuentan con métodos de segmentación v/uv basados en umbrales.

UNIVERSIDAD DE ANTIOQUIA 1803

SISTEMIC Sistemas Embebidos e Inteligencia Computacional

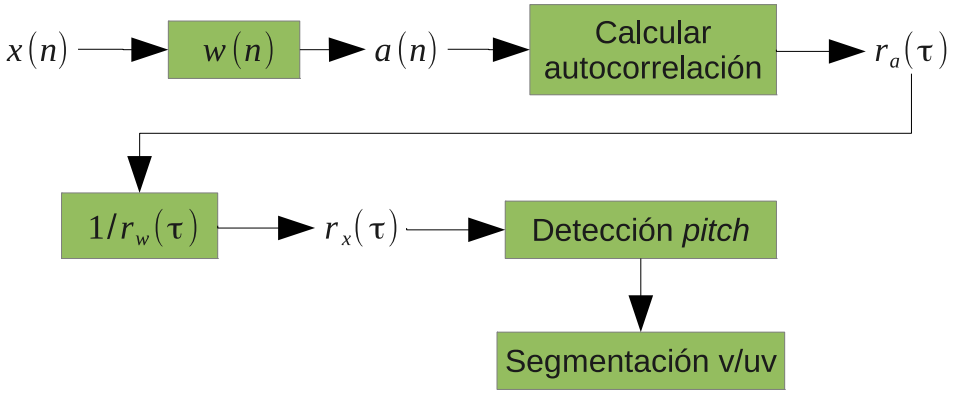
Detección de *pitch* y segmentación v/uv en Praat

Praat es uno de los programas más utilizados para el análisis de la voz [21]. El método modificado de la autocorrelación [19]:

- Toma en cuenta el efecto de la ventana.
- Robusto a errores de octava [19].
- Segmentación v/uv a partir de la energía del *pitch*.

Se evaluaron las configuraciones:

- **Praat1:** Ventana de Hamming.
- **Praat2:** Ventana Gaussiana.

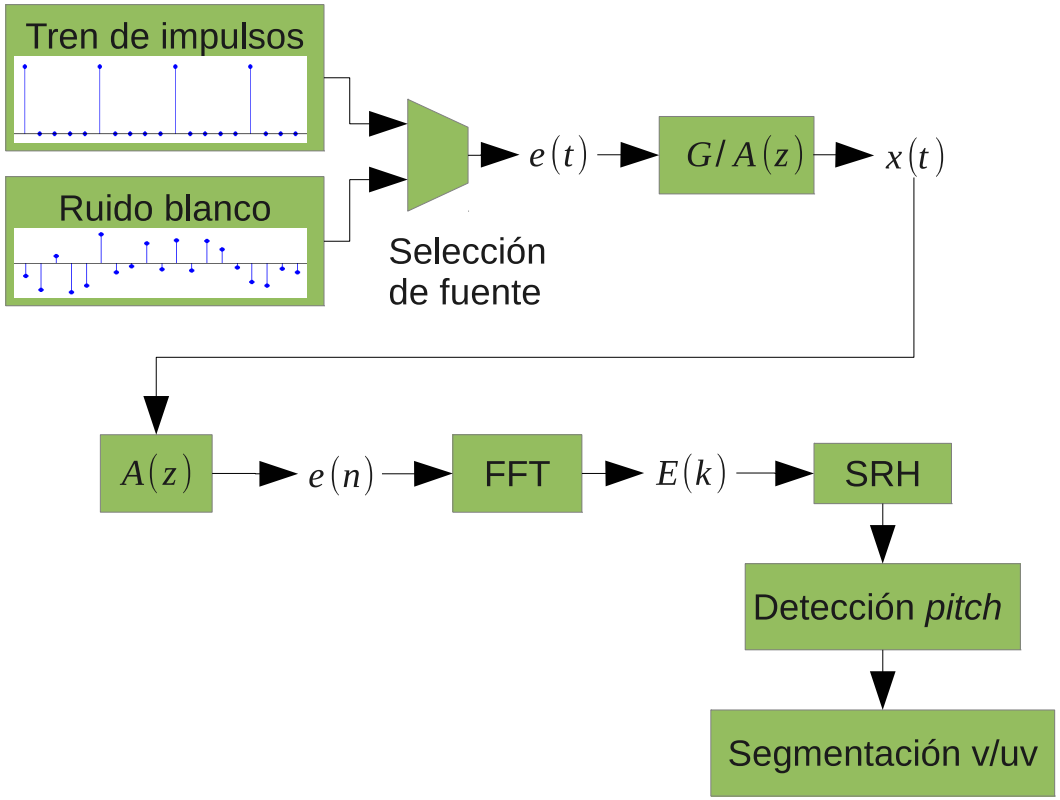


UNIVERSIDAD DE ANTIOQUIA 1803

SISTEMIC Sistemas Embebidos e Inteligencia Computacional

Método SRH

- Análisis en tiempo corto de la señal $e(t)$.
- Disminución de los efectos del ruido y las resonancias del tracto vocal[20].
- Detección en dos iteraciones.
- Segmentación v/uv a partir del valor del SRH.

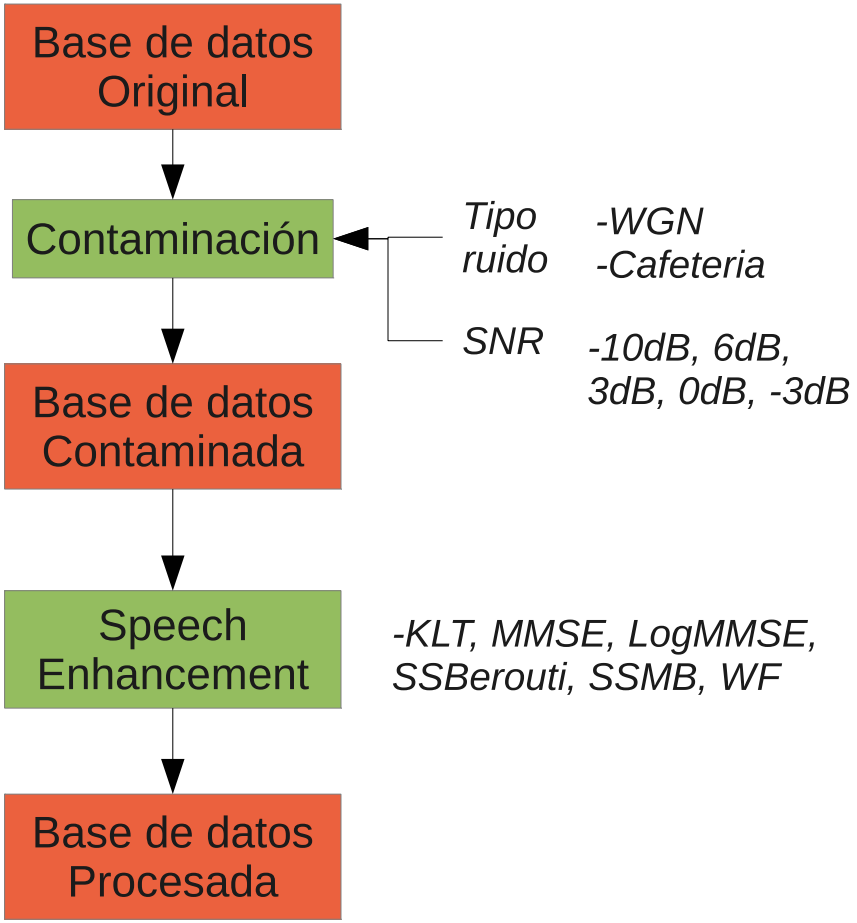


Metodología: Bases de datos

Se utilizó la base de datos de la Universidad Keele [22].

- Creada para la evaluación de algoritmos de detección del *pitch* y segmentación v/uv.
- 10 grabaciones de la fábula *The North Wind and the Sun*.
- Balanceada: 5 mujeres y 5 hombres.
- Etiquetas de referencia.

Se usó la configuración por defecto de los algoritmos de *Speech Enhancement*.



UNIVERSIDAD DE ANTIOQUIA 1803

SISTEMIC Sistemas Embebidos e Inteligencia Computacional

Metodología: Medidas de evaluación

Evaluación de los segmentadores v/uv [2]:

- *Voicing Detection Error* (VDE).

Considera:

- El número de segmentos vocálicos etiquetados como no vocálicos: $N_{V \to U}$
- El número de aquellos no-vocálicos etiquetados como vocálicos $N_{U \to V}$
- El número total de segmentos en la señal N

Se calcula como

$$\mathrm{VDE} = \frac{N_{V \to U} + N_{U \to V}}{N} \times 100\,\% \tag{2}$$

UNIVERSIDAD DE ANTIOQUIA 1803

SISTEMIC Sistemas Embebidos e Inteligencia Computacional

Metodología: Medidas de evaluación

Evaluación de la detección del *pitch* [2]:

- *Gross Pitch Error* (GPE)

Considera:

- El número de segmentos vocálicos con un error relativo mayor al 20 % N_{F0E}
- El número de segmentos vocálicos correctamente identificados N_{VV}

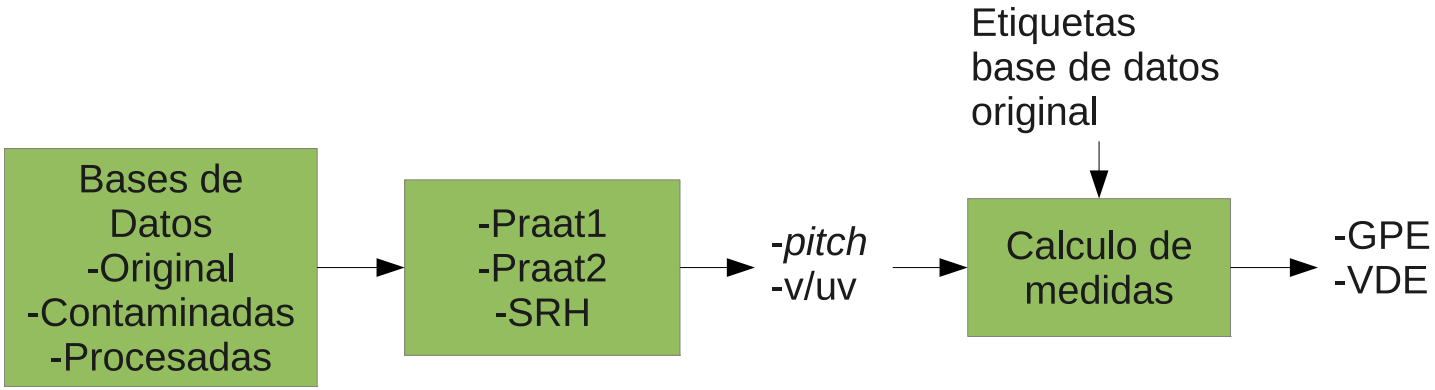
Se calcula como:

$$\mathrm{GPE} = \frac{N_{F0E}}{N_{VV}} \times 100\,\% \tag{3}$$

UNIVERSIDAD DE ANTIOQUIA 1803

SISTEMIC Sistemas Embebidos e Inteligencia Computacional

Experimento



Configuración de los algoritmos de detección y segmentación:

	Praat1	Praat2	SRH
Paso [ms]	10		
Ventana de análisis [ms]	40	80	100
Solape [%]	75	87.5	90
Mín. *pitch* [Hz]	75		
Máx. *pitch* [Hz]	500		

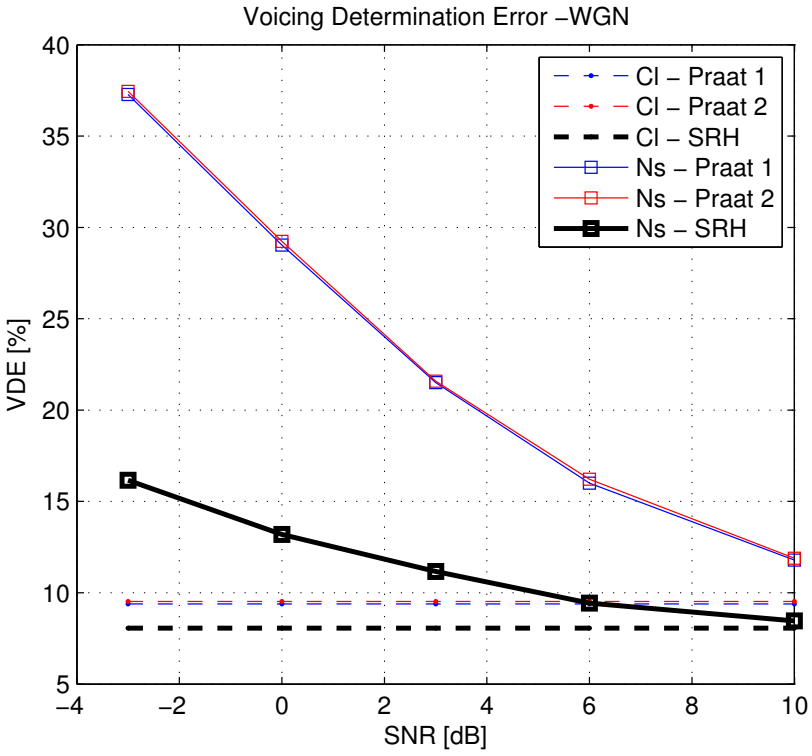
- Los valores presentados aquí son las media de los GPE y VDE en todas las señales de la base de datos.

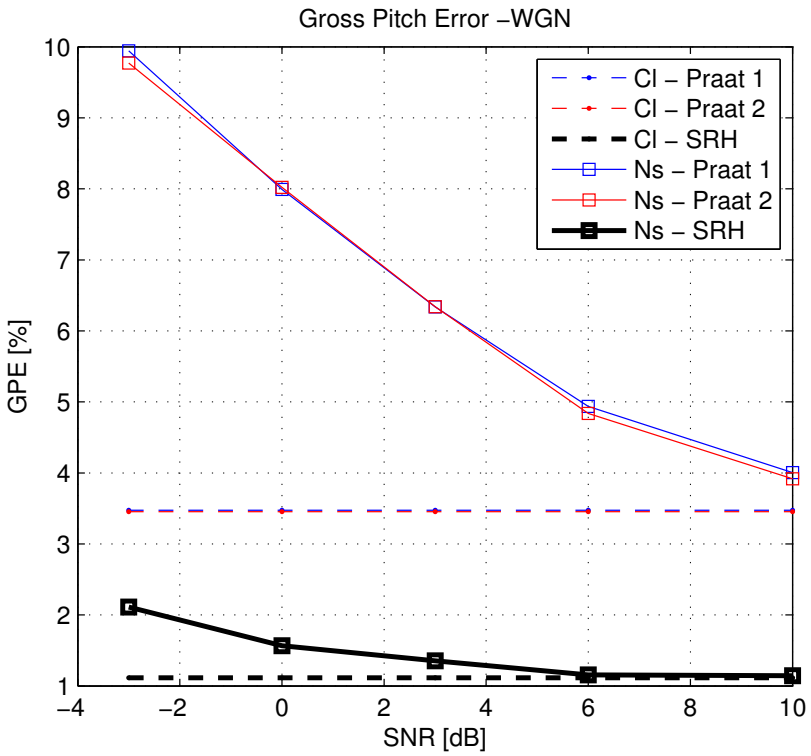
UNIVERSIDAD DE ANTIOQUIA 1803

SISTEMIC Sistemas Embebidos e Inteligencia Computacional

Resultados

Evaluación de los algoritmos de segmentación y detección.
Ruido WGN sin procesamiento.



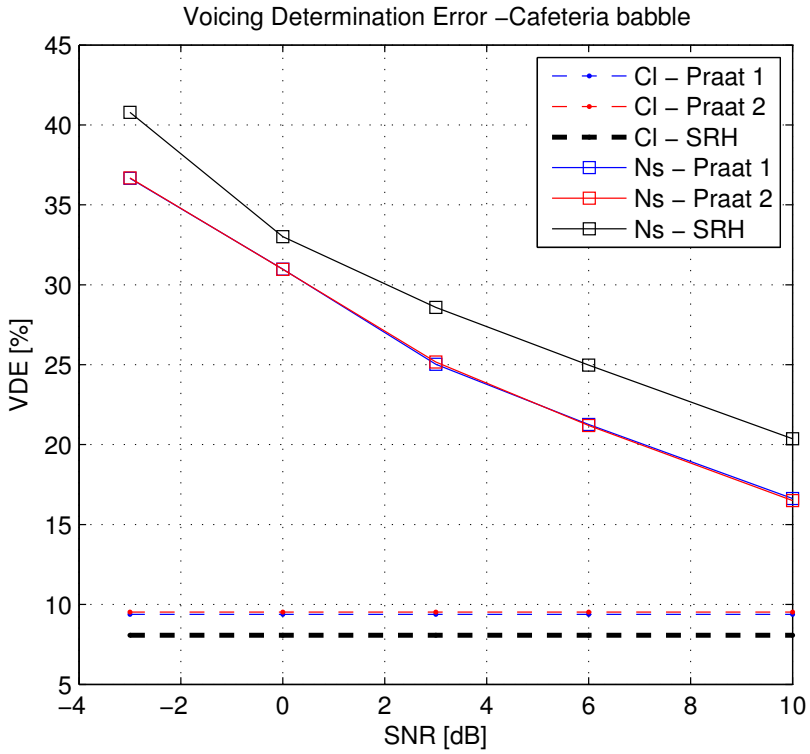


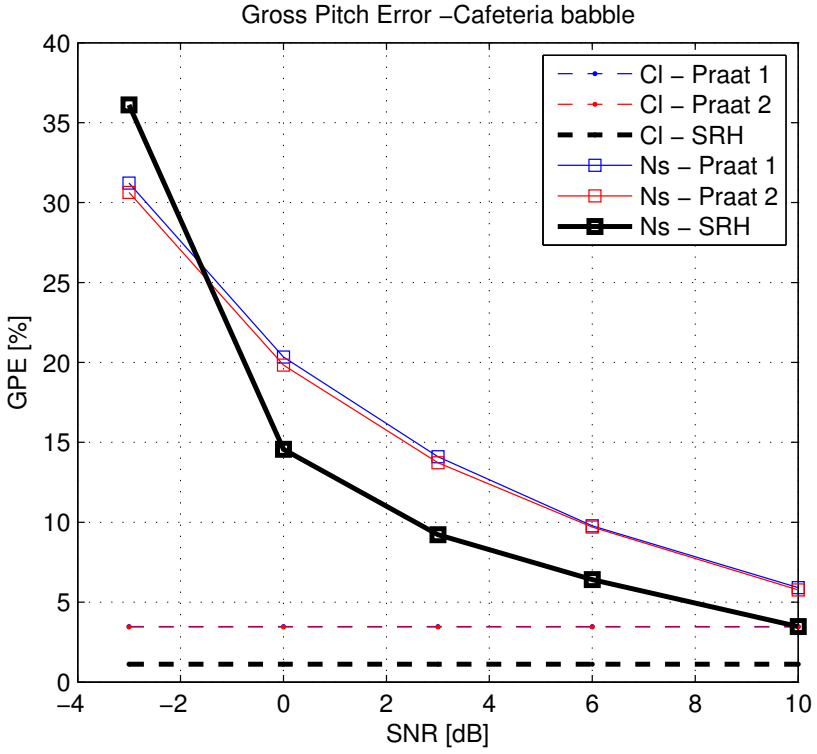
UNIVERSIDAD DE ANTIOQUIA 1803

SISTEMIC Sistemas Embebidos e Inteligencia Computacional

Resultados

Evaluación de los algoritmos de segmentación y detección.
Ruido Cafeteria sin procesamiento.



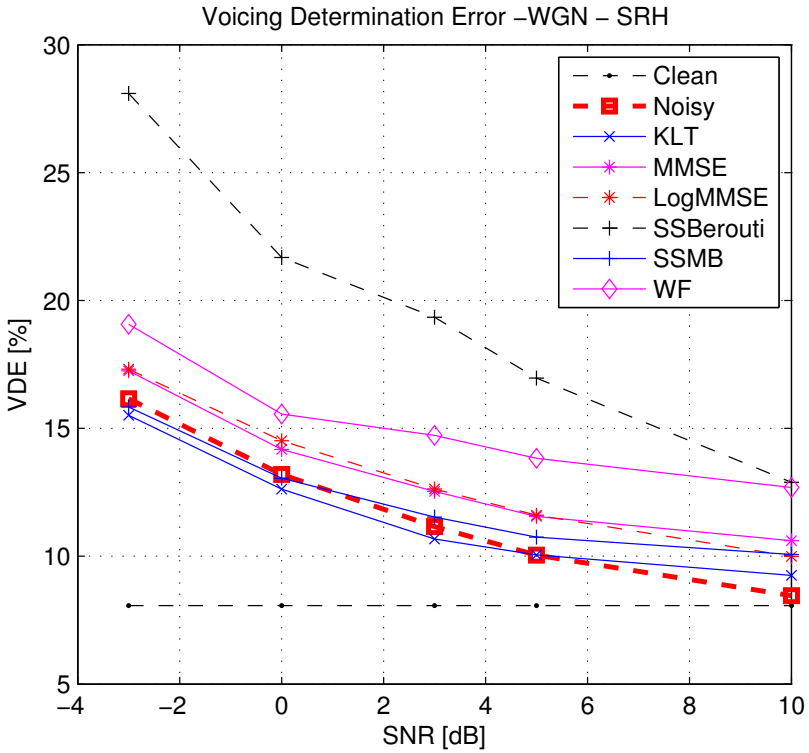


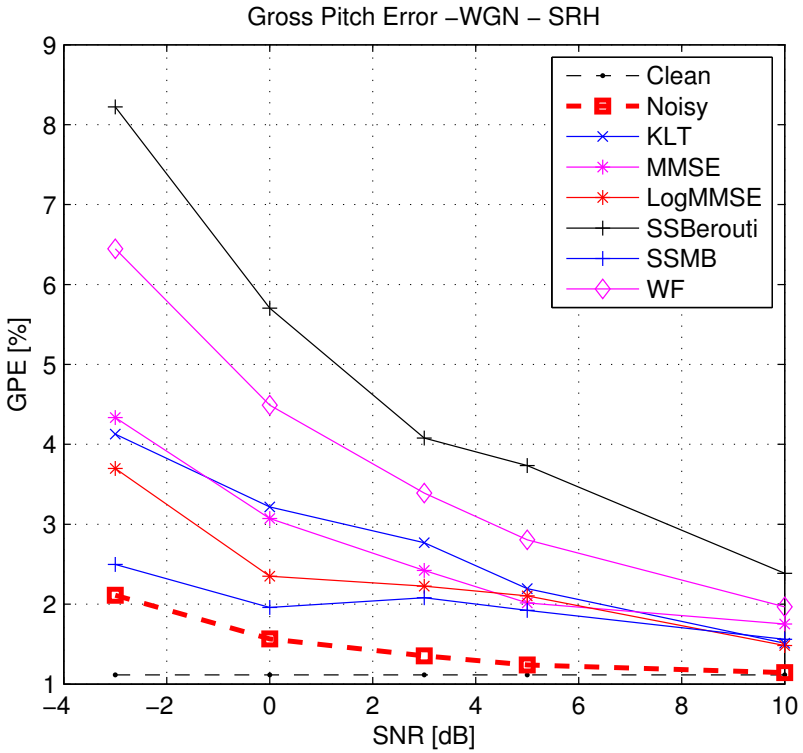
UNIVERSIDAD DE ANTIOQUIA 1803

SISTEMIC Sistemas Embebidos e Inteligencia Computacional

Resultados

Evaluación de los algoritmos de *Speech Enhancement*.
Ruido WGN. Segmentación y detección con SRH.



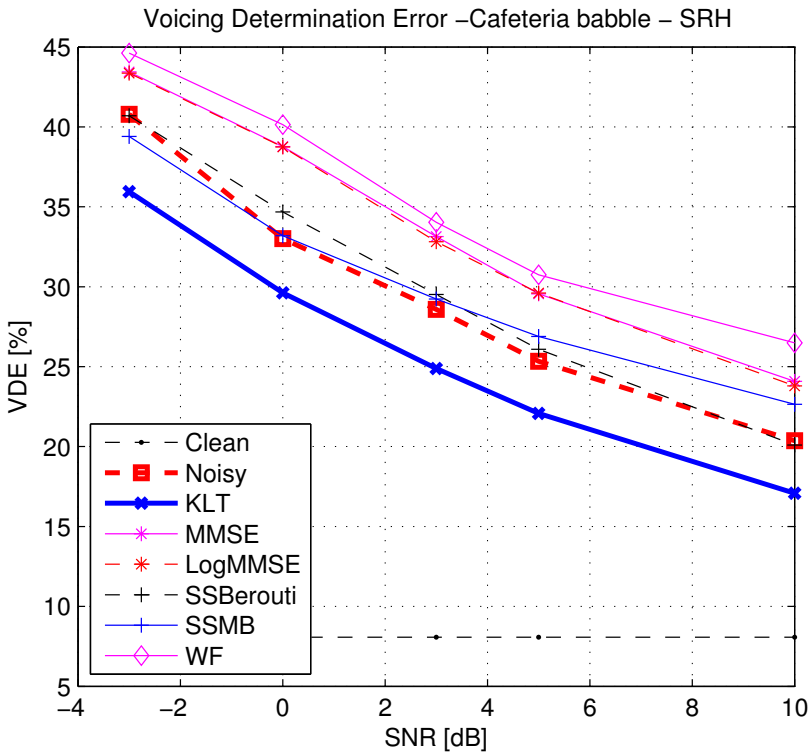


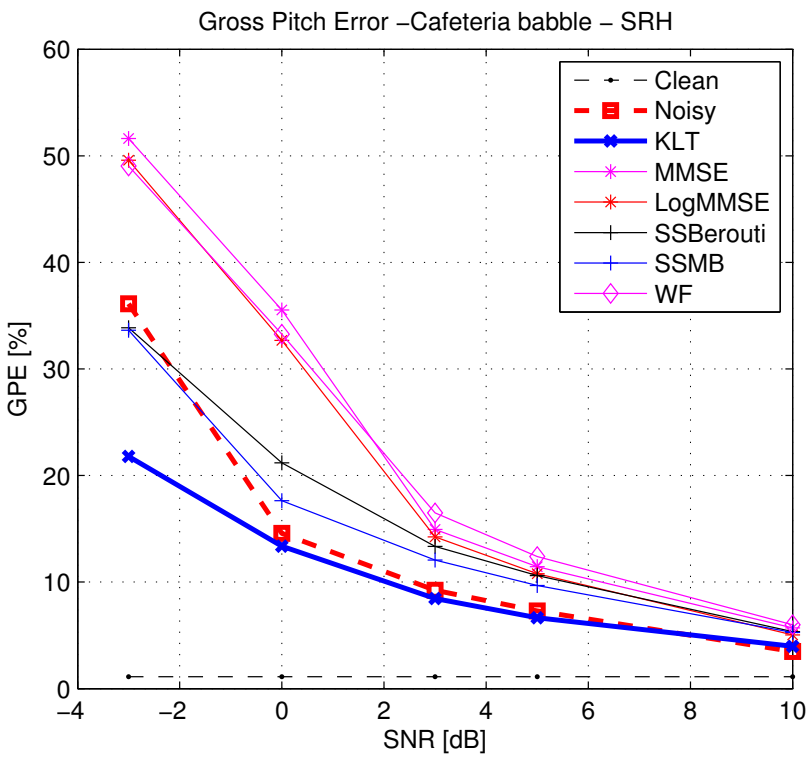
UNIVERSIDAD DE ANTIOQUIA 1803

SISTEMIC
Sistemas Embebidos e Inteligencia Computacional

Resultados

Evaluación de los algoritmos de *Speech Enhancement*.
Ruido Cafeteria. Segmentación y detección con SRH.





UNIVERSIDAD DE ANTIOQUIA 1803

SISTEMIC Sistemas Embebidos e Inteligencia Computacional

Conclusiones

- SRH tiene un desempeño mejor que Praat en la segmentación v/uv y la detección del *pitch* tanto para las señales limpias como para aquellas contaminadas con ruido WGN.
- Ambos, SRH y Praat, tienen desempeños similares para señales contaminadas con ruido tipo Cafeteria.
- Sólo el algoritmo KLT es capaz de mejorar la segmentación v/uv, realizada con SRH, en señales contaminadas con ruido tipo Cafeteria.
- Ninguna técnica de *Speech Enhancement* logra mejorar significativamente la detección del *pitch*.
- Como trabajo futuro se propone evaluar el efecto de las técnicas de *Speech Enhancement* en otras técnicas de detección del *pitch* y segmentación v/uv y en características como medidas de ruido y MFCCs.

UNIVERSIDAD DE ANTIOQUIA 1803

SISTEMIC Sistemas Embebidos e Inteligencia Computacional

Agradecimientos

Se agradece a Colciencias, mediante la convocatoria de jovenes investigadores e innovadores 2013, de la cual es beneficiario Juan Camilo Vásquez Correa, y mediante la Convocatoria 528 para estudios de doctorado en Colombia 2011 de la cual es beneficiario Juan Rafael Orozco Arroyave. Este trabajo también es parcialmente financiado por el proyecto de colciencias numero 111556933858.

UNIVERSIDAD DE ANTIOQUIA 1803

SISTEMIC Sistemas Embebidos e Inteligencia Computacional

Gracias

¿Preguntas?

Bibliografia I

[1] R. W. Schafer L. R. Rabiner.
Introduction to Digital Speech Processing, Foundations and Trends in Signal Processing, volume 1.
now Publishers Inc., Hanover, MA, 4 edition, 2007.

[2] W. J. Hess.
Spriger Handbook of Speech Processing, chapter 10. Pitch and Voicing Determination of Speech with an Extension Toward Music Signals, pages 181–212.
Springer-Verlag, Berlin, 2008.

[3] Eliza Concepcion, Mary Shalom, B. Lopez, Michael Gringo, Angelo Bayona, Michael Morales, Franz De Leon, Belen Calingacion, Prospero Naval, Rowena Cristina, and L. Guevara.
Determination of prosodic feature set for emotion recognition in acted call center speech.
2008.

[4] J. Pohjalainen and P. Alku.
Automatic detection of anger in telephone speech with robust autoregressive modulation filtering.
In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 7537–7541, May 2013.

[5] Philipos C. Loizou.
Speech Enhancement: Theory and Practice.
CRC Press, Inc., Boca Raton, FL, USA, 2nd edition, 2013.

[6] Jia-Ching Wang, Chung-Hsien Yang, Jhing-Fa Wang, and Hsiao-Ping Lee.
Robust speaker identification and verification.
Computational Intelligence Magazine, IEEE, 2(2):52–59, 2007.

[7] R. Saeidi, J. Pohjalainen, T. Kinnunen, and P. Alku.
Temporally weighted linear prediction features for tackling additive noise in speaker verification.
Signal Processing Letters, IEEE, 17(6):599–602, June 2010.

Bibliografia II

[8] S. Boll.
Suppression of acoustic noise in speech using spectral subtraction.
Acoustics, Speech and Signal Processing, IEEE Transactions on, 27(2):113–120, 1979.

[9] M. Berouti, R. Schwartz, and J. Makhoul.
Enhancement of speech corrupted by acoustic noise.
In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '79.*, volume 4, pages 208–211, 1979.

[10] Sunil Kamath and Philipos Loizou.
A multi-band spectral subtraction method for enhancing speech corrupted by colored noise.
In *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, volume 4, pages IV–4164–IV–4164, 2002.

[11] P. Scalart and J.V. Filho.
Speech enhancement based on a priori signal to noise estimation.
In *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, volume 2, pages 629–632 vol. 2, 1996.

[12] O. Cappe and J. Laroche.
Evaluation of short-time spectral attenuation techniques for the restoration of musical recordings.
Speech and Audio Processing, IEEE Transactions on, 3(1):84–93, Jan 1995.

[13] P.J. Wolfe and S.J. Godsill.
Simple alternatives to the ephraim and malah suppression rule for speech enhancement.
In *Statistical Signal Processing, 2001. Proceedings of the 11th IEEE Signal Processing Workshop on*, pages 496–499, 2001.

[14] Y. Ephraim and D. Malah.
Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator.
Acoustics, Speech and Signal Processing, IEEE Transactions on, 32(6):1109–1121, 1984.

Bibliografia III

[15] Y. Ephraim and D. Malah.
Speech enhancement using a minimum mean-square error log-spectral amplitude estimator.
Acoustics, Speech and Signal Processing, IEEE Transactions on, 33(2):443–445, 1985.

[16] Yi Hu and P.C. Loizou.
A generalized subspace approach for enhancing speech corrupted by colored noise.
Speech and Audio Processing, IEEE Transactions on, 11(4):334–341, 2003.

[17] M.A Little, P.E. McSharry, E.J. Hunter, J. Spielman, and L.O. Ramig.
Suitability of dysphonia measurements for telemonitoring of parkinson's disease.
Biomedical Engineering, IEEE Transactions on, 56(4):1015–1022, April 2009.

[18] Thomas Drugman, Myriam Rijckaert, Claire Janssens, and Marc Remacle.
Tracheoesophageal speech: A dedicated objective acoustic assessment.
Computer Speech & Language, (0):–, 2014.

[19] Paul Boersma.
Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound.
In *IFA Proceedings 17*, pages 97–110, 1993.

[20] Thomas Drugman and Abeer Alwan.
Joint robust voicing detection and pitch estimation based on residual harmonics.
In *INTERSPEECH*, pages 1973–1976. ISCA, 2011.

[21] Paul Boersma and David Weenink.
Praat: doing phonetics by computer, 2013.

[22] F. Plante, Georg F. Meyer, and William A. Ainsworth.
A pitch extraction reference database.
In *EUROSPEECH*. ISCA, 1995.

Bibliografia IV

[23] L.T. McWhorter and L.L. Scharf.
Multiwindow estimators of correlation.
Signal Processing, IEEE Transactions on, 46(2):440–448, Feb 1998.

Sustracción espectral con sobre-sustracción

Para disminuir la incidencia del ruido musical Beroutti et al. proponen en 1979 la siguiente modificación [9]

$$|\widehat{X}(\omega)| = \begin{cases} |Y(\omega)| - \alpha|\widehat{D}(\omega)| & \text{si } |Y(\omega)| > (\alpha+\beta)|\widehat{D}(\omega)| \\ \beta|\widehat{D}(\omega)| & \text{en otro caso} \end{cases}$$

Se propone utilizar dos parámetros en la técnica, el de sobre-sustracción α y el parámetro de piso de ruido β. Volver

Parámetro de sobre-sustracción

El parámetro de sobre-sustracción se calcula a partir de la SN por una relación lineal de la siguiente manera

$$\alpha = \begin{cases} \alpha_0 + \frac{5}{s} & \text{si } SNR < -5dB \\ \alpha_0 - \frac{1}{s}(SNR) & \text{si } -5dB \leq SNR \leq 20dB \\ 1 & \text{si } SNR > 20dB \end{cases}$$

donde $1/s$ es la pendiente de la recta.

El parámetro β es constante. Volver

Sustracción espectral multi-banda

Debido a que la señal de voz no se ve afectada de igual forma en todo el espectro, Kamath y Loizou proponen en 2002 la siguiente modificación [10]

$$|\widehat{X}_i(\omega)|^2 = \begin{cases} |Y_i(\omega)|^2 - \alpha_i \delta_i |\widehat{D}_i(\omega)|^2 & \text{si } |Y(\omega)|^2 > |\widehat{D}(\omega)|^2 \\ \beta |\widehat{D}_i(\omega)|^2 & \text{en otro caso} \end{cases}$$

El parámetro δ se calcula de la siguiente manera

$$\delta_i = \begin{cases} 1 & \text{si } f_i < 1\text{kHz} \\ 2{,}5 & \text{si } 1\text{kHz} < f_i < \frac{f_s}{2} - 2\text{kHz} \\ 1{,}5 & \text{si } f_i > \frac{f_s}{2} - 2\text{kHz} \end{cases}$$

Donde f_i es la frecuencia superior de la i-ésima banda y f_s la frecuencia de muestreo.

El parámetro de sobre-sustracción α_i también se calcula para cada banda con la SNR respectiva. Volver

Detalles, filtros de Wiener

Para desarrollar el filtro de Wiener en la frecuencia se parte de un filtro de respuesta infinita al impulso para obtener un estimador de la señal deseada d:

$$\hat{d}(n) = \sum_{k=-\infty}^{\infty} h_k y(n-k) = h(n) * y(n) \tag{4}$$

En el dominio de la frecuencia

$$\hat{D}(\omega) = H(\omega)Y(\omega) \tag{5}$$

El error de esta estimación viene dado por

$$E(\omega) = D(\omega) - \hat{D}(\omega) = D(\omega) - H(\omega)Y(\omega) \tag{6}$$

El error cuadrático medio viene dado por

$$\begin{aligned} E[|E(\omega)|^2] = \quad & E[|D(\omega)|^2] - H(\omega)E[D*(\omega)Y(\omega)] \\ & -H*(\omega)E[Y*(\omega)D(\omega)] + |H(\omega)|^2 E[|Y(\omega)|^2] \end{aligned} \tag{7}$$

Volver

Detalles, desarrollo filtros de Wiener

Notando que $P_{yy} = E[|Y(\omega)|^2]$ es el espectro de potencia de $y(n)$ y que $P_{yd} = E[Y(\omega)D*(\omega)]$ es el espectro cruzado de potencia, se puede expresar el MMSE como

$$E[|E(\omega)|^2] = E[|D(\omega)|^2] - H(\omega)P_{yd}(\omega) - H*(\omega)P_{dy}(\omega) + |H(\omega)|^2 P_{yy}(\omega) \tag{8}$$

derivando con respecto a $H(\omega)$ e igualando a cero se obtiene que el filtro óptimo está dado por

$$H(\omega) = \frac{P_{dy}(\omega)}{P_{yy}(\omega)} \tag{9}$$

Volver

Detalles, filtros de Wiener para Speech Enhancement

En el caso de la eliminación de ruido en señales de voz y asumiendo no-correlación entre señal de voz y ruido se tiene que

$$\begin{aligned} P_{dy} &= E[X(\omega)\{X(\omega) + N(\omega)\}*] \\ &= E[X(\omega)X*(\omega) + E[X(\omega)N*(\omega)] \\ &= P_{xx}(\omega) \end{aligned} \tag{10}$$

y de manera similar

$$\begin{aligned} P_{yy} &= E[\{X(\omega) + N(\omega)\}\{X(\omega) + N(\omega)\}*] \\ &= P_{xx}(\omega) + P_{nn}(\omega) \end{aligned} \tag{11}$$

Volver

Relación señal-a-ruido *a priori* y *a posteriori*

La relación señal a ruido (SNR) *a priori* es la razón entre la señal de voz original y el ruido, así se expresa de diferentes maneras de acuerdo a las características que se observan [5]

$$\xi_k \triangleq \frac{P_{xx}(\omega_k)}{P_{dd}(\omega_k)} \triangleq \frac{\lambda_x(k)}{\lambda_d(k)} \triangleq \frac{E[|X(\omega)|^2]}{E[|D(\omega)|^2]} \tag{12}$$

No se puede determinar directamente en condiciones reales. Se han desarrollado diferentes algoritmos para estimarla.
La Relación señal a ruido *a posteriori* se define como la razón entre la señal de voz contaminada y el ruido. Se puede expresar como

$$\gamma_k \triangleq \frac{P_{yy}(\omega_k)}{P_{dd}(\omega_k)} \triangleq \frac{Y_k^2}{\lambda_d(k)} \triangleq \frac{E[|Y(\omega)|^2]}{E[|D(\omega)|^2]} \tag{13}$$

Tampoco se puede determinar exactamente, pues no se puede determinar el ruido directamente.

Estimación de la SNR *a priori* I

Una de los métodos propuestos para obtener un estimado de la SNR *a priori* es conocido como *Decision-Directed Approach*. En este se parte de las siguiente igualdades [14]

$$\xi_k(m) = \frac{E[X_k^2(m)]}{\lambda_d(k,m)} \tag{14}$$

$$\xi_k(m) = E[\gamma_k(m)] - 1 \tag{15}$$

donde m denota la ventana de análisis actual. Al combinarlas resulta

$$\xi_k(m) = E\left[\frac{1}{2}\frac{X_k^2(m)}{\lambda_d(k,m)} + \frac{1}{2}(\gamma_k(m) - 1)\right] \tag{16}$$

El estimador final se logra al hacer la anterior ecuación recursiva.

$$\hat{\xi}_k(m) = \alpha\frac{|\hat{X}_k(m-1)|^2}{|\hat{D}_k(m-1)|^2} + (1-\alpha)\,\text{máx}(\gamma_k(m) - 1, 0) \tag{17}$$

donde α es un factor de ponderación que reemplaza el $1/2$ y $\hat{X}_k(m-1)$ es el espectro del estimador de la señal limpia obtenido para la ventana anterior.

Detalles algoritmo MMSE

En el enfoque Bayesiano del error cuadrático medio el valor esperado se realiza con respecto a la pdf conjunta $p(\mathbf{Y}, X_k)$ y está dado por

$$BMSE(\hat{X}_k) = \int \int (X_k - \hat{X}_k)^2 p(\mathbf{Y}, X_k) \mathbf{Y} dX_k \tag{18}$$

La minimización de la anterior resulta en

$$\hat{X}_k = E[X_k | Y(\omega_0) Y(\omega_1) ... Y(\omega_{N-1})] \tag{19}$$

y asumiendo independencia entre los coeficientes de la transformada de Fourier

$$\hat{X}_k = E[X_k | Y(\omega_k)] = \int_0^\infty X_k p(x_k | Y(\omega_k)) dX_k \tag{20}$$

Volver

Detalles algoritmo MMSE

Utilizando la regla de Bayes para determinar la probabilidad $p(x_k|Y(\omega_k))$

$$\hat{X}_k = \frac{\int_0^\infty X_k p(Y(\omega_k|X_k)p(X_k)dX_k}{\int_0^\infty P(Y(\omega_k)|X_k)p(X_k)dX_k} \tag{21}$$

Además

$$p(Y(\omega_k|X_k)p(X_k) = \int_0^{2\pi} p(Y(\omega_k)|x_k,\theta_x)p(x_k,\theta_x)d\theta_x \tag{22}$$

Donde θ_x es la realización de la variable aleatoria de la fase de $X(\omega_k)$. Ahora se deben estimar $p(Y(\omega_k)|x_k,\theta_x)$ y $p(x_k,\theta_x)$. Asumiendo $Y(\omega_k)$ como la suma de dos variables aleatorias Gaussianas de media cero

$$p(Y(\omega_k)|x_k,\theta_x) = \frac{1}{\pi\lambda_d(k)}\exp(\left\{-\frac{1}{\lambda_d(k)}|Y(\omega_k)-X(\omega_k)|^2\right\} \tag{23}$$

donde $\lambda_D(k)$ es la varianza de la k-ésima componente espectral del ruido. Volver

Detalles algoritmo MMSE

Para variables aleatorias Gaussianas complejas como se asumió $X(\omega_k)$ se sabe que la magnitud, X_k y fase, $\theta_x(k)$, son independientes y que se puede evaluar la pdf conjunta como el producto de las pdfs individuales. La pdf de X_k sigue una distribución de Rayleigh pues $X_k = \sqrt{r(k)^2 + i(k)^2}$ donde $r(k) = \mathrm{Re}\{X(\omega_k)\}$ y $i(k) = \mathrm{Im}\{X(\omega_k)\}$ son variables aleatorias Gaussianas. La pdf de $\theta_x(k)$ is uniforme en $(-\pi, \pi)$, así

$$p(x_k, \theta_x) = \frac{X_k}{\pi \lambda_x(k)} \exp\left\{ -\frac{X_k^2}{\lambda_x(k)} \right\} \tag{24}$$

donde $\lambda_x(k)$ la varianza de la k-ésima componente espectral de la señal limpia.

Volver

Detalles algoritmo MMSE

El paso final para obtener el estimador de la magnitud de la señal limpia se da reemplazando las ecuaciones 24 y 23 en 22 y esta a su vez en 21 y utilizando la representación integral de la función de Bessel modificada, así

$$I_n(z) = \frac{1}{2\pi} \int_0^{2\pi} \cos \beta n \exp(z \cos \beta)\, d\beta \qquad \text{(A.1)}$$

we obtain

$$\hat{A}_k = \frac{\int_0^{\infty} a_k^2 \exp\left(-\frac{a_k^2}{\lambda(k)}\right) I_0\left(2a_k \sqrt{\frac{v_k}{\lambda(k)}}\right) da_k}{\int_0^{\infty} a_k \exp\left(-\frac{a_k^2}{\lambda(k)}\right) I_0\left(2a_k \sqrt{\frac{v_k}{\lambda(k)}}\right) da_k} \qquad \text{(A.2)}$$

con

$$\frac{1}{\lambda(k)} = \frac{1}{\lambda_x(k)} + \frac{1}{\lambda_d(k)} \qquad (25)$$

Volver

Detalles algoritmo MMSE

Se pueden evaluar las integrales en la anterior diapositiva como

$$\int_0^\infty x^c e^{-ax^2} I_0(bx)dx = \frac{\Gamma(0{,}5c + 0{,}5)}{2a^{(c+1)/2}} \Phi\left(\frac{c+1}{2}, 1; \frac{b^2}{4a}\right) \tag{26}$$

donde $\Phi(a, b; z)$ es la función hipergeométrica confluente, definida como

$$\Phi(a, b; z) = 1 + \frac{a}{b}\frac{z}{1!} + \frac{a(a+1)}{b(b+1)}\frac{z}{2!} + \frac{a(a+1)(a+2)}{b(b+1)(b+2)}\frac{z}{3!} + \dots \tag{27}$$

Volver

Estimador de la magnitud LogMMSE

Basado en las suposiciones ya mencionadas para los coeficientes y minimizando el error cuadrático medio del logaritmo del espectro de magnitud [15]

$$E\{(\log X_k - \log \hat{X}_k)^2\} \tag{28}$$

se obtiene un estimador con la función de ganancia

$$G(\xi_k, \gamma_k) = \frac{\xi_k}{\xi_k + 1} \exp\left\{\frac{1}{2}\int_{\nu_k}^{\infty} \frac{e^{-t}}{t} dt\right\} \tag{29}$$

Volver

Detalles algoritmo LogMMSE

El estimador óptimo se puede obtener evaluando la media del logaritmo de X_K

$$\log \hat{X}_k = E[\log X_k | Y(\omega_k)] \tag{30}$$

Denotando $Z_k = \log X_k$, su función generadora de momentos condicionada en $Y(\omega_k)$ está dada por

$$\Phi_{Z_k|Y(\omega_k)}(\mu) = E[\exp(\mu Z_k)|Y(\omega_k)] = E[X_k^{\mu}|Y(\omega_k)] \tag{31}$$

Utilizando el mismo modelo estadístico que se utilizó para obtener el estimador MMSE, se obtiene

$$\Phi_{Z_k|Y(\omega_k)}(\mu) = \lambda_k^{\mu/2}\Gamma\left(\frac{\mu}{2}+1\right)\Phi\left(-\frac{\mu}{2},1;-\nu_k\right) \tag{32}$$

donde $\Gamma(\cdot)$ es la función gamma y $\Phi(a,b;x)$ es la función hipergeométrica confluente.

Volver

Detalles algoritmo LogMMSE

Tomando la derivada con respecto μ y evalúandola en $\mu = 0$ es obtiene la media condicional del logiaritmo de X_k

$$E[\log X_k | Y(\omega_k)] = \frac{1}{2}\left[\log \lambda_k + \log \nu_k + \int_{\nu_k}^{\infty} \frac{\exp -t}{t} dt\right] \tag{33}$$

Volver

Enfoque de subespacio generalizado

En esta propuesta se busca una matriz H óptima que pueda diagonalizar simultáneamente las matrices de covarianza de la señal limpia y el ruido, $R_{\mathbf{x}}, R_n \in \Re^{K \times K}$. Así, [16]

$$H = V^{-T} Q V^T \tag{34}$$

donde V es la matriz de eigenvectores de la matriz Σ estimada como

$$\Sigma = R_{\mathbf{n}}^{-1} R_{\mathbf{y}} - I \tag{35}$$

y Q es una matriz diagonal cuyos elementos q_{ii} se calculan como:

$$q_{ii} = \frac{\lambda_{\mathbf{x}}^{(i)}}{\lambda_{\mathbf{x}}^{(i)} + \mu} \tag{36}$$

donde los $\lambda_{\mathbf{x}}$ se estiman como los eigenvalores positivos de la matriz Σ y μ es el multiplicador de Lagrange, un parámetro escalar que sirve para controlar el equilibrio entre la cantidad de ruido removida y la distorsión de la señal.

Volver

Detalles enfoque de subespacio generalizado

Para ello se sigue el siguiente procedimiento: 1. Calcular la matriz de covarianza conjunta Σ:

$$\Sigma = R_{\mathbf{n}}^{-1} R_{\mathbf{x}} \tag{37}$$

Como no es posible acceder $\mathbf{x}$, this matrix se estima como

$$\Sigma = R_{\mathbf{n}}^{-1} R_{\mathbf{y}} - I \tag{38}$$

donde $R_{\mathbf{y}} \in \Re^{K \times K}$ es la matriz de covarianza de la señal contaminada, estimada utilizando el método *multi-taper* [23].

Volver

Detalles enfoque de subespacio generalizado

La matriz de covarianza del ruido $R_{\mathbf{n}}$ se estima inicialmente encontrando el vector de autocorrelación en una ventana de la señal que sólo contiene ruido y organizando este vector en una matriz de Toeplitz y se actualiza cuando se cumple:

$$\frac{R_{\mathbf{y}}(1,1)}{R_{\mathbf{n}}(1,1)} \leq \gamma \tag{39}$$

donde γ es un umbral. Luego, se actualiza como:

$$R_{\mathbf{n}}^{(\text{new})} = \alpha R_{\mathbf{n}}^{(\text{old})} + (1-\alpha) R_{\mathbf{y}} \tag{40}$$

donde α es el coeficiente de adaptación, que controla la tasa de actualización.

Volver

Detalles enfoque de subespacio generalizado

2. Obtener las matrices de eigenvalores y ($\Lambda_{\mathbf{x}}$) eigenvectores (V) of Σ, i.e.,

$$\Sigma V = \Lambda_{\mathbf{x}} V \tag{41}$$

Estos diagonalizan simultáneamente a $R_{\mathbf{x}}$ y $R_{\mathbf{n}}$:

$$\Lambda_{\mathbf{x}} = V^T R_{\mathbf{x}} V \tag{42}$$

$$I = V^T R_{\mathbf{n}} V \tag{43}$$

Volver

Detalles enfoque de subespacio generalizado

3. Se calcula el estimador óptimo como:

$$H = V^{-T} Q V^{T} \tag{44}$$

donde Q es una matriz diagonal cuyos elementos q_{ii} se calculan como:

$$q_{ii} = \frac{\lambda_{\mathbf{x}}^{(i)}}{\lambda_{\mathbf{x}}^{(i)} + \mu} \tag{45}$$

donde $\lambda_{\mathbf{x}}$ son los eigenvalores de la matriz de covarianza de la señal limpia y μ es el multiplicador de Lagrange, un parámetro escalar que sirve para controlar el equilibrio entre la cantidad de ruido removida y la distorsión de la señal.
Como no es posible acceder a la señal limpia, los eigenvalores de $R_{\mathbf{x}}$ se estiman como los eigenvalores positivos de Σ.

Volver

Detalles método de detección de *pitch* en Praat

La señal enventanada se calcula como

$$a(t) = [x(t_{\text{mid}} - 0{,}5\,T + t) - \mu_x]\, w(t) \tag{46}$$

donde μ_x es la media de la señal en el segmento.
La función de autocorrelación normalizada se calcula como

$$r_a(\tau) = \frac{\int_0^{T-\tau} a(t)a(t+\tau)dt}{\int_0^T a^2(t)dt} \tag{47}$$

El estimado de la función de autocorrelación es:

$$r_x(\tau) \approx \frac{r_a(\tau)}{r_w(\tau)} \tag{48}$$

- La longitud de análisis de la ventana está ligada a la frecuencia mínima a considerar como *pitch*.

Volver

Detalles método de Suma de Armónicos de la Señal Residual (SRH)

- Filtrado inverso por predicción lineal de orden 12 para obtener $e(t)$.
- Se calcula la FFT de un tamaño igual a la frecuencia de muestreo para obtener $E(k)$.
- El valor del SRH se calcula como

$$\mathrm{SRH}(f) = E(f) + \sum_{k=2}^{N_{harm}} [E(k \cdot f) - E((k - 0{,}5) \cdot f)] \tag{49}$$

- Dos iteraciones, la segunda con adaptación de las frecuencias máximas y mínimas al hablante.

Volver