DETECCIÓN DE ACTIVIDAD DE VOZ ROBUSTA AL RUIDO BASADA EN MODELOS ACÚSTICOS

Ángel de la Torre, Javier Ramírez, Carmen Benítez, José C. Segura, Luz García, Antonio J. Rubio

Departamento de Teoría de la Señal Telemática y Comunicaciones, Universidad de Granada
[atv,javierrp,carmen,segura,luzgm,rubio]@ugr.es

RESUMEN

En este trabajo proponemos un detector de actividad de voz (VAD) derivado de la aproximación VTS (Vector Taylor Series approach). Se hace uso de una mezcla de Gaussianas (entrenada con voz limpia) para proporcionar una decisión apropiada para la detección de voz/no-voz. La aproximación VTS adapta la mezcla de Gaussianas a las condiciones de ruido, lo que proporciona un rendimiento estable para un rango amplio de SNRs. Hemos evaluado el método propuesto en cuanto a la habilidad en la detección de voz/no-voz, y también en cuanto a la aplicación de VAD para reconocimiento robusto de voz. En comparación con otros métodos de detección de actividad de voz, el VAD propuesto presenta el mejor compromiso en detección de voz/no-voz. Cuando el VAD es aplicado para filtrado de Wiener y eliminación de frames de silencio, el método propuesto también muestra los mejores resultados de reconocimiento.

1. INTRODUCCIÓN

Los sistemas de reconocimiento automático de voz se ven fuertemente afectados por el ruido. Se han desarrollado numerosas técnicas para paliar el efecto del ruido sobre la tasa de reconocimiento. La mayor parte de ellos frecuentemente requieren estimar una estadística del ruido mediante un detector de actividad de voz (VAD) suficientemente preciso. La tarea de clasificación en voz/no-voz no es tan trivial como inicialmente podría parecer y la mayor parte de los algoritmos para VAD fallan cuando el nivel de ruido de fondo se incrementa. Durante la última década numerosos investigadores han desarrollado diferentes estrategias para detectar voz en señales ruidosas [1, 2, 3, 4], poniendo especial atención en la derivación y estudio de características y reglas de decisión robustas al ruido.

En este artículo proponemos un algoritmo de detección de actividad de voz basado en modelos acústicos. La decisión voz/no-voz está basada en la aproximación VTS (Vector Taylor Series approach) [5, 6, 7, 8]. La aproximación VTS, propuesta inicialmente como una técnica de compensación de ruido para reconocimiento robusto de voz, se ha adaptado para la detección de actividad de voz. La formulación VTS está basada en una mezcla de Gaussianas en el dominio de las energías del banco de filtros en escaladas logarítmicamente (log-FBE). La mezcla de Gaussianas se utiliza para calcular la probabilidad de cada Gaussiana dada la trama de entrada ruidosa. En la formulación VTS estas probabilidades son usadas para obtener una estimación limpia de la trama. En el algoritmo VAD basado en VTS que proponemos, estas probabilidades son usadas para calcular la probabilidad de que la trama sea voz. La decisión del VAD es realizada comparando la probabilidad de que la trama sea voz con un umbral. Con este enfoque se esperan dos ventajas. Por una parte, la decisión del VAD se apoya en un modelo de mezcla de Gaussianas entrenado con voz limpia, y por tanto, la decisión del VAD está basada en eventos de voz observados en la base de datos de entrenamiento. Por otra parte, la aproximación VTS proporciona un método para adaptar la mezcla de Gaussianas a las condiciones de ruido. De este modo, el método propuesto permite la adaptación del VAD a las condiciones de ruido y, por tanto, cabe esperar que el rendimiento del VAD se mantenga estable para un rango extenso de SNRs.

2. VAD BASADO EN VTS

2.1. Vector Taylor Series approach

La aproximación VTS [5, 6, 7, 8] es un método de compensación del ruido que proporciona una representación de la voz limpia eliminando el ruido aditivo. Esta compensación del ruido se realiza en el dominio log-FBE y está basada en una mezcla de Gaussianas. La formulación VTS asume que el efecto del ruido puede describirse como un término aditivo en el dominio log-FBE,

$$\mathbf{y}(\mathbf{x},\mathbf{n}) = \mathbf{x} + \mathbf{g}(\mathbf{x},\mathbf{n}) \quad (1)$$

donde $\mathbf{x}$ y $\mathbf{y}$ son vectores en este dominio que representan la voz limpia y ruidosa, respectivamente para un frame dado, y $\mathbf{n}$ representa el ruido aditivo que afecta a este frame. Para el canal i-ésimo, $\mathbf{g}$ queda descrita por la ecuación,

$$g(i) = \log\left(1 + \exp\left(n(i) - x(i)\right)\right) \quad (2)$$

Se pueden definir dos funciones auxiliares $f(i)$ y $h(i)$ del siguiente modo,

$$f(i) \equiv \frac{1}{1+\exp(x(i)-n(i))} = -\frac{\partial g(i)}{\partial x(i)} = \frac{\partial g(i)}{\partial n(i)} \quad (3)$$

$$h(i) \equiv (1-f(i))f(i) = \frac{\partial^2 g(i)}{\partial x^2(i)} = \frac{\partial^2 g(i)}{\partial n^2(i)} \quad (4)$$

y usando estas definiciones, podemos aproximar $y(i)$ usando un desarrollo en serie de Taylor alrededor de unos valores $x_0(i)$ y $n_0(i)$. Similarmente, usando la aproximación del desarrollo en serie de Taylor podemos describir cómo una función densidad de probabilidad Gaussiana en el dominio log-FBE se ve afectada por el ruido aditivo. Consideremos una función densidad de probabilidad Gaussiana que representa voz limpia, con media $\mu_x(i)$ y matriz de covarianza $\Sigma_x(i,j)$, y asumamos un proceso de ruido aditivo de tipo Gaussiano, con media $\mu_n(i)$ y matriz de covarianza $\Sigma_n(i,j)$. Podemos hacer un desarrollo en serie de Taylor alrededor de $x_0(i) = \mu_x(i)$ y $n_0(i) = \mu_n(i)$. La media y la matriz de covarianza de la función densidad de probabilidad que describe la voz ruidosa puede obtenerse como los valores esperados,

$$\mu_y(i) = E[y(i)] \quad (5)$$

$$\Sigma_y(i,j) = E[(y(i)-\mu_y(i))(y(j)-\mu_y(j))] \quad (6)$$

y pueden ser estimados como una función de $\mu_x(i)$, $\mu_n(i)$, $\Sigma_x(i,j)$ y $\Sigma_n(i,j)$ usando el desarrollo en serie de Taylor como,

$$\mu_y(i) \approx \mu_x(i) + g_0(i) + \frac{1}{2}h_0(i)[\Sigma_x(i,i) + \Sigma_n(i,i)] \quad (7)$$

$$\Sigma_y(i,j) \approx (1-f_0(i))(1-f_0(j))\Sigma_x(i,j) + f_0(i)f_0(j)\Sigma_n(i,j) + \frac{1}{2}h_0^2(i)(\Sigma_x(i,i)+\Sigma_n(i,i))^2\delta_{i,j} \quad (8)$$

donde $g_0(i)$, $f_0(i)$ y $h_0(i)$ están evaluadas en $x_0(i) = \mu_x(i)$ y $n_0(i) = \mu_n(i)$. De este modo, la aproximación VTS proporciona una función densidad de probabilidad Gaussiana que representa la voz ruidosa, calculada a partir de la Gaussiana que describe la voz limpia y la Gaussiana que describe el ruido.

Si la voz limpia es modelada como una mezcla de K Gaussianas, la aproximación VTS proporciona una estimación de la voz limpia $\hat{\mathbf{x}}$ dadas la voz ruidosa observada $\mathbf{y}$ y la estadística del ruido (μ_n y Σ_n) como,

$$\hat{\mathbf{x}} \approx \mathbf{y} - \sum_k P(k|\mathbf{y})\mathbf{g}(\mu_{x,k}, \mu_n) \quad (9)$$

donde $\mu_{x,k}$ es la media de la k-ésima Gaussiana limpia y $P(k|\mathbf{y})$ es la probabilidad de que la observación $\mathbf{y}$ haya sido generada por la Gaussiana k, dada por,

$$P(k|\mathbf{y}) = \frac{P(k)\mathcal{N}(\mathbf{y}, \mu_{y,k}, \Sigma_{y,k})}{\sum_{k'} P(k')\mathcal{N}(\mathbf{y}, \mu_{y,k'}, \Sigma_{y,k'})} \quad (10)$$

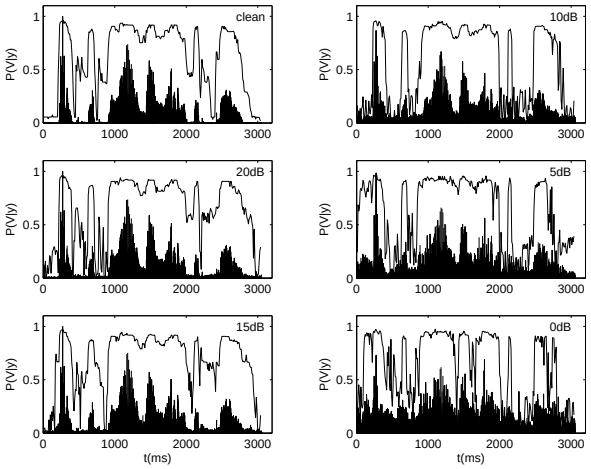

Figura 1. Probabilidad $P(V|\mathbf{y})$ de que el frame $\mathbf{y}$ sea voz para cada frame de una frase, evaluada a diferentes SNRs. La amplitud de la señal (para valores positivos) también se ha representado en cada gráfica.

donde $P(k)$ es la probabilidad a priori de la k-ésima Gaussiana y $\mathcal{N}(\mathbf{y}, \mu_{y,k}, \Sigma_{y,k})$ es la k-ésima Gaussiana ruidosa (con media $\mu_{y,k}$ y matriz de covarianza $\Sigma_{y,k}$) evaluada en $\mathbf{y}$. La media y la matriz de covarianza de la Gaussiana k-ésima ruidosa puede estimarse a partir de la estadística del ruido (μ_n y Σ_n) y de la k-ésima Gaussiana limpia ($\mu_{x,k}$ y $\Sigma_{x,k}$) usando las ecuaciones (7) y (8).

2.2. Aplicación de VTS a la detección de actividad de voz

Si a cada Gaussiana k se le asigna una probabilidad $P(V|k)$ (la probabilidad de que la Gaussiana k-ésima sea voz), la probabilidad de que un frame ruidoso de entrada $\mathbf{y}$ sea voz puede evaluarse como,

$$P(V|\mathbf{y}) = \sum_k P(V|k)P(k|\mathbf{y}) \quad (11)$$

donde $P(k|\mathbf{y})$ viene dada por la aproximación VTS (ecuación (10)).

La probabilidad $P(V|k)$ se puede estimar fácilmente para cada Gaussiana, dado que la mezcla de Gaussianas se ha construido a partir de una base de datos de entrenamiento limpia. En este trabajo hemos considerado la media para la componente de energía de la Gaussiana limpia E_k para la estimación de esta probabilidad, y se ha considerado una función lineal entre dos límites de energía (E_0 y E_1),

$$P(V|k) = \begin{cases} 0 & E_k < E_0 \\ (E_k - E_0)/(E_1 - E_0) & E_k \in [E_0, E_1] \\ 1 & E_k \geq E_0 \end{cases} \quad (12)$$

Los parámetros E_0 y E_1 se han ajustado empíricamente.

La figura 1 muestra la evolución en el tiempo de la probabilidad $P(V|\mathbf{y})$ evaluada para una frase extraída de la base de datos AURORA-2 [9], obtenida para diferentes SNRs. En este ejemplo se puede observar la efectividad del método a bajas SNRs. Se ontiene una probabilidad mayor que 0.8 para, al menos, algunos frames en cada sílaba incluso para SNRs bajas (esta frase corresponde con la cadena de dígitos ingleses 86Z1162). La decisión voz/no-voz puede realizarse usando un umbral T. El frame $\mathbf{y}$ es etiquetado como voz si $P(V|\mathbf{y}) > T$, y como no-voz en caso contrario. Adicionalmente, se puede incluir un time-in y un time-out de algunos frames con objeto de evitar que el VAD descarte aquellos frames de voz con poca energía al principio y al final de algunas sílabas.

3. MARCO EXPERIMENTAL

Se suelen llevar a cabo varios experimentos con objeto de evaluar el rendimiento de los algoritmos para VAD. El análisis está normalmente enfocado en la determinación de las probabilidades de error para diferentes escenarios de ruido y valores de la SNR [10, 4], así como en la influencia de la decisión del VAD sobre el rendimiento de los sistemas de procesamiento de voz [11, 12]. El marco experimental y los test objetivos llevados a cabo para evaluar el rendimiento del algoritmo propuesto se describen en esta sección.

3.1. Evaluación en diferentes entornos de ruido

En primer lugar, el VAD propuesto se evaluó en términos de la habilidad para discriminar entre voz y no-voz en diferentes escenarios y con diferentes niveles de ruido usando la base de datos AURORA-2 [9]. Esta base de datos se ha construido a partir de la base de datos limpia TIdigits (que consiste en secuencias de hasta siete dígitos conectados pronunciados por locutores de inglés americano), usada como fuente de voz, y una selección de 8 ruidos reales diferentes, que han sido añadidos artificialmente a la voz a SNRs de 20dB, 15dB, 10dB, 5dB, 0dB y -5dB. En el análisis de discriminación, la base de datos limpia TIdigits se usó para etiquetar manualmente cada frame de cada frase como voz o no-voz para referencia. El rendimiento de detección de actividad de voz es evaluado en términos de tasa de acierto de no-voz (HR0) y tasa de acierto de voz (HR1), definidos como la fracción de todos los frames de no-voz o voz que han sido correctamente detectados como no-voz o voz, respectivamente,

$$\text{HR1} = \frac{N_{1,1}}{N_1^{\text{ref.}}} \qquad \text{HR0} = \frac{N_{0,0}}{N_0^{\text{ref.}}} \tag{13}$$

donde $N_1^{\text{ref.}}$ y $N_0^{\text{ref.}}$ son el número de frames de no-voz o voz que realmente hay en la base de datos, y $N_{1,1}$ y $N_{0,0}$ son el número de frames de voz y no-voz que han sido correctamente clasificados, respectivamente.

La figura 2 compara el VAD basado en VTS propuesto (usando un umbral T=0.5) frente a algoritmos estandarizados incluyendo ITU-T G.729 [13], ETSI AMR [14] y ETSI AFE [15], y otros algoritmos recientemente reportados [1, 2, 3, 4]. Se muestran la tasa de acierto de no-voz (HR0) y la tasa de acierto de voz (HR1) para condiciones limpias y para niveles de SNR de entre 20 y -5 dB. Nótese que se proporcionan los resultados para los dos VADs definidos en el estándar [15] (uno para la estimación del espectro del ruido en la fase de filtrado de Wiener, WF, y el otro usado en la fase de eliminación de frames de no-voz, FD). Los resultados mostrados en estas figuras son valores promedio para el conjunto completo de ruidos. Puede concluirse, de la figura 2 que:

- El VAD ITU-T G.729 muestra una precisión pobre en detección de voz cuando se incrementa el nivel de ruido, mientras que la detección de no-voz es buena en condiciones limpias (85 %) y pobre (20 %) en condiciones ruidosas.
- El VAD ETSI AMR1 presenta un comportamiento extremadamente conservativo con una precisión alta en la detección de voz para todo el rango de niveles de SNR, pero a costa de unos resultados en detección de no-voz muy pobres a medida que se incrementa el nivel de ruido. Aunque el AMR1 parece un procedimiento apropiado para detección de voz en condiciones de ruido desfavorables, su comportamiento extremadamente conservativo degrada la precisión en detección de no-voz, llegando el HR0 a ser menor del 10 % por debajo de 10 dB, lo que lo hace poco útil para sistemas de procesamiento de voz.
- El VAD ETSI AMR2 alcanza una mejora considerable sobre los detectores G.729 y AMR1, alcanzando mejores precisiones en detección de no-voz, aunque aun sufre una degradación rápida de la detección de voz en condiciones de ruido desfavorables.
- El VAD usado en el estándar AFE para la estimación del espectro del ruido en la etapa de filtrado de Wiener está basado en la energía de la banda completa y proporciona un rendimiento pobre en detección de voz con una caída rápida del HR1 para valores bajos de la SNR. Por otra parte, el VAD usado en el AFE para eliminación de frames de no-voz (frame-dropping, FD) alcanza una alta precisión en detección de voz pero resultados moderados en detección de no voz.
- Finalmente, el VAD basado en VTS que proponemos alcanza el mejor compromiso de entre los dife-

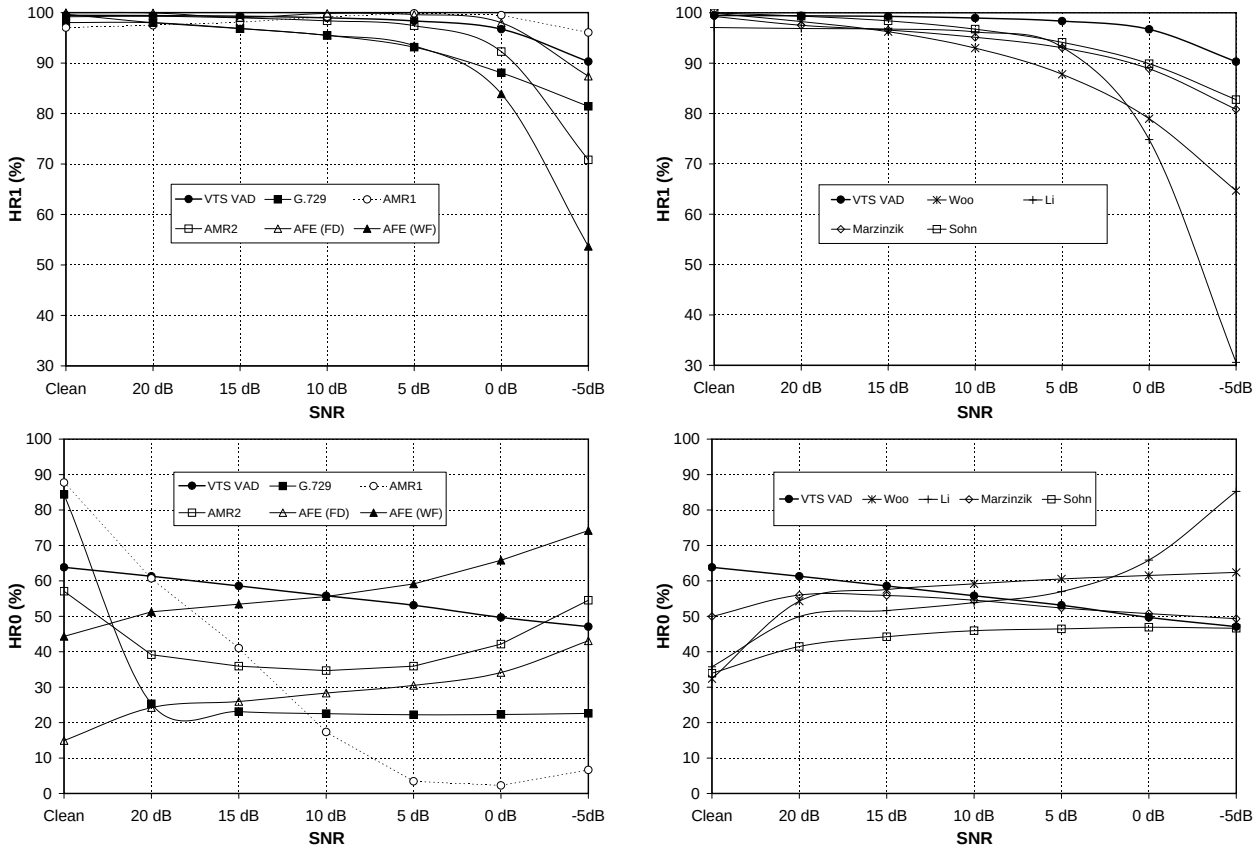

Figura 2. Análisis de la discriminación voz/no-voz en función de la SNR. Resultados promediados para todos los ruidos considerados en la base de datos AURORA-2. Se representan la tasa de acierto de voz y de no-voz (comparados con los obtenidos estándares y otros VADs recientemente reportados.)

rentes procedimientos evaluados. Este VAD obtiene un buen comportamiento en detección de periodos de no-voz y además muestra una caída lenta en el rendimiento para condiciones de ruido desfavorables en la detección de frames de voz (90 % a -5 dB).

La tabla 1 muestra un resumen de las ventajas proporcionadas por el VAD basado en VTS sobre otros métodos de detección de voz en términos de las tasas de detección de voz y no-voz promediadas sobre el rango completo de SNRs. De la tabla puede observarse que el método propuesto, con un HR1 promedio del 97.50 % y un HR0 promedio del 55.62 %, alcanza el mejor compromiso en detección de voz/no-voz cuando se compara con el resto de detectores de voz evaluados.

3.2. Evaluación del VAD en un sistema de reconocimiento robusto

Si bien el análisis de discriminación presentado en la sección anterior es efectivo para la evaluación de un algoritmo de detección de voz/no-voz, también hemos estudiado la influencia del VAD sobre un sistema de reconocimiento de voz. El marco de referencia es front-end para reconocimiento distribuido de voz (DSR) [16] propuesto por el grupo de trabajo STQ de la ETSI para la evaluación de algoritmos de extracción de características robustos al ruido para DSR. Se ha estudiado la influencia de la decisión del VAD sobre el rendimiento de diferentes aspectos de la extracción de características. La primera aproximación incorpora filtrado de Wiener (WF) al sistema base como método de supresión de ruido. El segundo algoritmo de extracción de características combina filtrado de Wiener y eliminación de frames de no-voz (frame dropping, FD). La tabla 2 muestra los resultados de reconocimiento para AURORA-2 en función de la SNR para experimentos de reconocimiento de voz basados en los

Tabla 2. Tasa de reconocimiento promedio (Word Accuracy) para experimentos "clean training" con la base de datos AURORA-2.

	Base	**Base + WF**					**Base + WF + FD**				
		G.729	AMR1	AMR2	AFE	**VTS-VAD**	G.729	AMR1	AMR2	AFE	**VTS-VAD**
Clean	99.03	98.81	98.80	98.81	98.77	98.85	98.41	97.87	98.63	98.78	98.58
20 dB	94.19	87.70	97.09	97.23	97.68	97.41	83.46	96.83	96.72	97.82	97.51
15 dB	85.41	75.23	92.05	94.61	95.19	95.07	71.76	92.03	93.76	95.28	95.37
10 dB	66.19	59.01	74.24	87.50	87.29	88.11	59.05	71.65	86.36	88.67	89.34
5 dB	39.28	40.30	44.29	71.01	66.05	70.93	43.52	40.66	70.97	71.55	73.83
0 dB	17.38	23.43	23.82	41.28	30.31	40.65	27.63	23.88	44.58	41.78	44.81
-5 dB	8.65	13.05	12.09	13.65	4.97	13.16	14.94	14.05	18.87	16.23	17.61
Average	60.49	57.13	66.30	78.33	75.30	**78.43**	57.08	65.01	78.48	79.02	**80.17**

Tabla 1. Tasas de acierto de voz y no-voz promediadas para SNRs entre condiciones limpias y -5dB.

Comparación con VADs estándar			
	G.729	AMR1	AMR2
HR0 (%)	31.77	31.31	42.77
HR1 (%)	93.00	98.18	93.76
	AFE (WF)	AFE (FD)	**VTS-VAD**
HR0 (%)	57.68	28.74	**55.62**
HR1 (%)	88.72	97.70	**97.50**

Comparación con otros VADs					
	Sohn	Woo	Li	Marz.	**VTS-VAD**
HR0 (%)	43.66	55.40	57.03	52.69	**55.62**
HR1 (%)	94.46	88.41	83.65	93.04	**97.50**

algoritmos de VAD G.729, AMR, AFE, y VTS VAD. Estos resultados se han promediado sobre los tres conjuntos de test definidos para los experimentos de reconocimiento de AURORA-2. Como conclusión puede apreciarse que el VAD propuesto supera a los estándares de detección de voz G.729, AMR1, AMR2 y AFE, cuando es aplicado para el filtrado de Wiener y también cuando es aplicado conjuntamente para filtrado de Wiener y para eliminación de frames de no-voz.

4. CONCLUSIONES

En este trabajo hemos propuesto un algoritmo VAD basado en modelos acústicos derivado de la aproximación VTS (Vector Taylor Series approach). El uso de una mezcla de Gaussianas (en el dominio log-FBE) entrenada con una base de datos de voz limpia proporciona una regla de decisión de voz apropiada para la detección de voz/no-voz. Por otra parte, la formulación VTS permite la adaptación de la mezcla de Gaussianas a las condiciones de ruido, alcanzándose un rendimiento estable del VAD propuesto para un rango extenso de SNRs y de tipos de ruido. El VAD basado en VTS propuesto se ha evaluado en términos de la capacidad para discriminación entre voz y no-voz en diferentes escenarios de ruido. Cuando se ha comparado con otros VADs estándar, hemos encontrado que el VAD propuesto proporciona el mejor compromiso en detección de voz/no-voz, con un HR1 promedio del 97.50 % y un HR0 promedio del 55.62 % (promediados entre condiciones limpias y -5dB). Con respecto al rendimiento en reconocimiento de voz, el VAD propuesto también proporciona los mejores resultados de reconocimiento cuando es usado para la estimación del espectro del ruido en el filtrado de Wiener y cuando es usado para eliminación de tramas de no-voz.

5. AGRADECIMIENTOS

Este trabajo ha recibido soporte del 6^o Programa Marco de la Unión Europea, a través del contrato número IST-2002-507943 (HIWIRE, Human Input that Works in Real Environments), y del Gobierno Español a través del proyecto SR3-VoIP (TEC2004-03829/TCM). Los puntos de vista expresados en este trabajo corresponden únicamente a los autores. La Unión Europea no es responsable de ningún uso que se pueda hacer de la información contenida en este trabajo.

6. BIBLIOGRAFÍA

[1] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Processing Letters*, vol. 16, no. 1, pp. 1–3, 1999.

[2] K. Woo, T. Yang, K. Park, and C. Lee, "Robust voice activity detection algorithm for estimating noise

spectrum," *Electronics Letters*, vol. 36, no. 2, pp. 180–181, 2000.

[3] Q. Li, J. Zheng, A. Tsai, and Q. Zhou, "Robust endpoint detection and energy normalization for real-time speech and speaker recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 3, pp. 146–157, 2002.

[4] M. Marzinzik and B. Kollmeier, "Speech pause detection for noise spectrum estimation by tracking power envelope dynamics," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 6, pp. 341–351, 2002.

[5] P.J. Moreno, *Speech Recognition in Noisy Environments*, Ph.D. thesis, Carnegie Mellon University, Pittsburgh, Pensilvania, 1996.

[6] P.J. Moreno, B. Raj, and R.M. Stern, "A vector Taylor series approach for environment-independent speech recognition," in *Proc. of ICASSP-96*, 1996, pp. 733–736.

[7] R.M. Stern, B. Raj, and P.J. Moreno, "Compensation for environmental degradation in automatic speech recognition," in *ESCA-NATO Tutorial and Research Workshop on Robust Speech Recognition for Unknown Communication Channels*, 1997, pp. 33–42.

[8] J.C. Segura, A. de la Torre, M.C. Benítez, and A.M. Peinado, "Model-based compensation of the additive noise for continuous speech recognition. Experiments using the Aurora-II database and tasks," in *Proc. of EuroSpeech-2001*, 2001, pp. 221–224.

[9] H.G. Hirsch and D. Pearce, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noise conditions," in *ISCA ITRW ASR2000 Automatic Speech Recognition: Challenges for the Next Millennium*, Paris, France, September 2000.

[10] F. Beritelli, S. Casale, G. Rugeri, and S. Serrano, "Performance evaluation and comparison of G.729/AMR/Fuzzy voice activity detectors," *IEEE Signal Processing Letters*, vol. 9, no. 3, pp. 85–88, 2002.

[11] J. Ramírez, J. C. Segura, M. C. Benítez, A. de la Torre, and A. Rubio, "Efficient voice activity detection algorithms using long-term speech information," *Speech Communication*, vol. 42, no. 3-4, pp. 271–287, 2004.

[12] J. Ramírez, José C. Segura, C. Benítez, A. de la Torre, and A. Rubio, "An effective subband OSF-based VAD with noise reduction for robust speech recognition," *IEEE Trans. on Speech and Audio Processing*, vol. 13, no. 6, pp. 1119–1129, 2005.

[13] ITU, "A silence compression scheme for G.729 optimized for terminals conforming to recommendation V.70," *ITU-T Recommendation G.729-Annex B*, 1996.

[14] ETSI, "Voice activity detector (VAD) for Adaptive Multi-Rate (AMR) speech traffic channels," *ETSI EN 301 708 Recommendation*, 1999.

[15] ETSI, "Speech processing, transmission and quality aspects (STQ); distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithms," *ETSI ES 202 050 Recommendation*, 2002.

[16] ETSI, "Speech processing, transmission and quality aspects (stq); distributed speech recognition; front-end feature extraction algorithm; compression algorithms," *ETSI ES 201 108 Recommendation*, 2000.