RENDIMIENTO PERCEPTUAL Y RECONOCIMIENTO CON CODIFICADORES VOIP SOBRE REDES DE PAQUETES

José L. Carmona, Antonio M. Peinado, José L. Perez Córdoba, Ángel M. Gómez, Victoria Sánchez

Dpto. Teoría de la Señal, Telemática y Comunicaciones, Universidad de Granada

RESUMEN

El objetivo de este trabajo es llevar a cabo una exploración sobre la degradación introducida por una red IP sobre la calidad perceptual de la voz decodificada y el reconocimiento a partir de ésta. Destaca la evaluación de los codificadores más empleados en este entorno: G.729, G723.1 e ILBC, así como la amplia casuística de simulaciones de canal, caracterizadas por el porcentaje de pérdida de paquetes y el tamaño medio de las ráfagas. La evaluación de los codificadores se realizó en dos planos: por un lado, se analizó la calidad perceptual de la voz reconstruida mediante el sistema PESQ, mientras que por otro se estudió la eficiencia del sistema de reconocimiento Aurora-2 a partir de la voz sintetizada. Finalmente, se presentan interesantes comparativas de la eficiencia de los codificadores, poniendo de relevancia y justificando sus debilidades y fortalezas.

1. INTRODUCCIÓN

Este trabajo establece una comparativa del rendimiento de distintos codificadores VoIP (*Voice over IP*), con el fin de encontrar un codec que obtenga resultados óptimos tanto en la calidad de la voz reconstruida como en el RAH (Reconocimiento Automático del Habla) a partir de ésta. Esta topología de reconocimiento remoto recibe el nombre de arquitectura NSR (*Network Speech Recognition*), caracterizándose por situar el sistema reconocedor en su totalidad en el lado servidor.

Como alternativa se encuentra la arquitectura DSR (*Distributed Speech Recognition*), mostrada en la figura 1.a, que fracciona el sistema de reconocimiento entre el cliente y el servidor, de modo que la extracción de características es llevada a cabo por el primero y el reconocimiento en sí (comparación de patrones y decisión), por el segundo.

En la arquitectura NSR, la voz es codificada mediante un codec convencional y enviada hasta el extremo del servidor (véase figura 1.b) donde es posible llevar a cabo tareas RAH o reproducir la voz con otros fines. Por este motivo, evaluar la arquitectura NSR en IP exige evaluar los codificadores en reconocimiento y calidad perceptual, sometidos a las degradaciones propias de esta red. Hay

Este trabajo ha sido financiado por el proyecto MEC/FEDER TEC2004-03829.

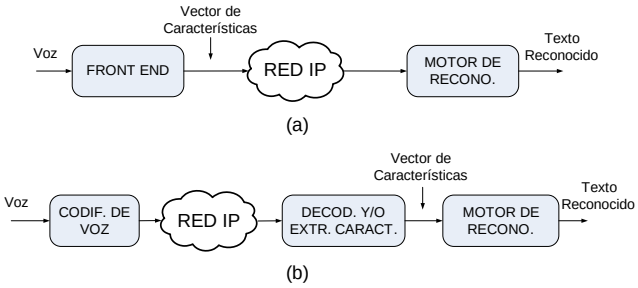

Figura 1. *Arquitecturas de reconocimiento remoto: (a) Topología DSR, (b) Topología NSR.*

que tener en cuenta que los codificadores tradicionales de voz tienen ciertas desventajas en el reconocimiento automático del habla, ya que no se encuentran diseñados a tal efecto. En este caso las redundancias de la voz relativas a la identificación del locutor, que no son necesarias en las tareas de reconocimiento, se encuentran aún en la voz codificada, lo que hace que los *bit-rates* sean mayores en las topologías NSR. Además, el procesado realizado por los codecs introduce ciertas distorsiones sobre la voz decodificada que son perjudiciales en las tareas de reconocimiento.

A pesar de las desventajas de la arquitectura NSR, ésta sigue presentando aspectos de elevado interés. Por un lado, supera las dificultades de implantación de los sistemas DSR, puesto que opera con los codecs convencionales de voz, mientras que al tener éstos como objetivo transmitir la señal con la máxima calidad perceptual posible, incluyen características del locutor que permiten su identificación. Esta última característica permitiría la implementación de servicios tan interesantes, desde el punto de vista comercial, como la firma de contratación de servicios por voz o la autoría de operaciones bancarias. Considérese que aunque existen sistemas DSR que incluyen parámetros relativos al locutor, como la frecuencia de *pitch*, la naturalidad de la voz sintetizada será menor y, por tanto, los procesos de identificación del locutor menos eficientes.

Obviamente, el rendimiento de la arquitectura NSR vendrá determinado por la robustez del codificador de voz y la calidad perceptual de ésta frente a las degradaciones introducidas por la red IP. He aquí la razón que motivó el desarrollo de este estudio comparativo, en el que

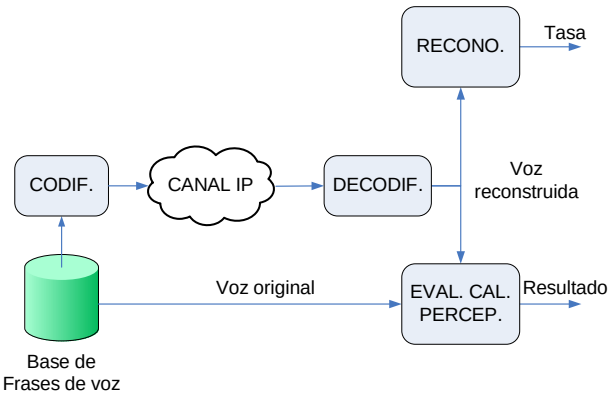

Figura 2. *Esquema general de las pruebas realizadas.*

se explora el comportamiento de los codificadores más utilizados en la actualidad en función de la degradación de la red, considerada ésta como ráfagas de paquetes perdidos.

2. ESQUEMA GENERAL

La estructura del experimento desarrollado se basa en la simulación de un sistema NSR en el que se irán sustituyendo las partes codificadora y decodificadora por los distintos codecs a evaluar. Una vez la información es codificada se segmenta en paquetes, para posteriormente pasar a la simulación de la condición de canal correspondiente. Esta simulación consistirá en marcar un conjunto de paquetes como perdidos, de modo que en el lado receptor éstos no sean utilizados para la síntesis de la señal de voz. En su lugar, cada uno de los codificadores utilizará su algoritmo PLC (*Packet Loss Concealment*) mediante el cual sustituirá la información perdida por aproximaciones derivadas de los paquetes correctamente recibidos con anterioridad. A medida que esta aproximación se desvíe de la información original, la voz sintetizada presentará mayor degradación.

Tal y como se observa en la figura 2, la degradación de la voz reconstruida en el extremo receptor se mide mediante dos sistemas de evaluación: un sistema RAH y un sistema evaluador de la calidad perceptual. El motivo por el que se han utilizado estos dos sistemas de evaluación se debe a que en los sistemas NSR sería posible reconstruir la voz con dos fines bien distintos, por un lado para ser reconocida por un sistema automático y por otro para ser escuchada por un oyente humano o realizar procesos de identificación del locutor. Bajo otra perspectiva, podría decirse que el sistema reconocedor mide la inteligibilidad del mensaje, mientras que el evaluador de la calidad perceptual tendría además en cuenta parámetros relacionados con la naturalidad del habla.

3. SELECCIÓN DE CODECS

Antes de comenzar a detallar los distintos elementos del experimento, es necesario seleccionar qué codificadores serán sometidos a prueba. A continuación se presentan las razones que justifican cada selección, así como las principales características de los codificadores elegidos.

3.1. G.729

Desarrollado por la ITU (*International Telecommunication Union*) y amparado por la recomendación H.323 que define los protocolos necesarios para proveer sesiones de comunicación audio-visual sobre redes de conmutación de paquetes, y que ha ido evolucionando para dirigir las crecientes necesidades de VoIP. Concretamente, G.729 presenta un bit-rate de 8 kbps (aunque posteriormente se han añadido anexos que ofrecen tasas de 6.4 y 11.8 kbps) y, según la recomendación H.323 se adecúa perfectamente a aquellos tipos de aplicaciones basadas sólo en audio y con un bajo requerimiento de ancho de banda. El principio de funcionamiento de este codificador es CS-ACELP (*Conjugate Structure - Algebraic CELP*) y actúa sobre tramas de voz de 10 ms, presentando un retardo algorítmico de 15 ms, ya que para el procesado de una trama precisa de los 5 ms iniciales de la siguiente (*look-ahead* de 5 ms). La definición de la carga útil del protocolo RTP viene definida, para los codificadores de la ITU, por el RFC 3551 que permite la inclusión de una o más tramas por paquete.

3.2. G.723.1

Al igual que G.729, G.723.1 es un estándar desarrollado por la ITU e incluido en la recomendación H.323, con la diferencia de presentar éste dos modos de trabajo. Su esquema de funcionamiento se encuentra basado en ACELP pero, a diferencia de AMR, el tamaño de trama utilizado es de 30 ms. Este codificador lleva a cabo un *look-ahead* de 7.5 ms, presentando, por tanto, un retardo algorítmico de 37.5 ms. Los dos modos de operación se denominan por sus *bit-rates*: 5.3 y 6.3 kbps.

3.3. iLBC

Nuevo codificador diseñado para presentar un óptimo rendimiento en las redes de conmutación de paquetes. En este caso, hace uso de una exclusiva estructura, distinta de la CELP, que codifica parte de la excitación (*start state*) mediante un codificador ADPCM, utilizando posteriormente ésta como diccionario adaptativo para la codificación del resto de la excitación de la trama. La codificación ADPCM condiciona valores altos de *bit-rate*, proporcionando dos modos, 13.33 y 15.2 kbps, con tamaños de trama de 30 y 20 ms respectivamente. En este caso, el codificador no lleva a cabo *look-ahead* en su procesado, limitándose el retardo algorítmico al tamaño de trama utilizado. La creación del RFC 3951, que define

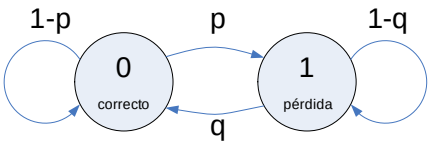

Figura 3. *Modelo de Gilbert.*

las especificaciones de este codec, y el RFC 3952, que marca la distribución de la carga útil del protocolo RTP, han hecho que este codec se difunda rápidamente en aplicaciones de VoIP. La actual predilección en su uso por sistemas tan extendidos como *Google Talk* y *Skype* [1], así como su caracter libre de derechos de autor y su diseño orientado directamente hacia VoIP [2], lo hacen un serio aspirante a convertirse en un codec predominante en este entorno.

De este modo, el conjunto de codecs y modos seleccionado queda de la siguiente manera: G.729 e iLBC {modos 13.33 y 15.2 kbps}. El número de tramas por paquete fue definido del siguiente modo: 1 trama/paquete para iLBC y G.723.1, mientras que 2 y 3 tramas/paquete para G.729. Es necesario remarcar que en el desarrollo de este trabajo cada codec emplea su propio algoritmo de mitigación de pérdidas, basados todos en ellos en técnicas de repetición y *muting.*

4. EMULADOR DEL CANAL IP

El uso de modelos de Markov se adapta bien al caracter rafagueante de las pérdidas [3], ya que son capaces de capturar la dependencia temporal de éstas. Concretamente, en este trabajo fue empleado el más simple de ellos, el modelo de dos estados: recepción correcta (estado 0) o pérdida de paquete (estado 1), presentados en la figura 3. El modelo queda totalmente definido por las probabilidades de transición p y q, tal y como se muestra en la figura.

A partir de estas probabilidades pueden determinarse tanto la probabilidad no condicional de pérdidas ulp (*unconditional loss packet*), como la probabilidad condicional de pérdidas clp (*conditional loss packet*):

$$\begin{array}{c} ulp = \frac{p}{p+q} \\ clp = 1 - q \end{array} \qquad (1)$$

donde ya puede identificarse que la probabilidad de pérdidas total P_{per} del modelo se corresponde con ulp, mientras que el tamaño medio de ráfaga L_{raf} vendrá determinado por clp a través del cálculo:

$$L_{raf} = \sum_{i=1}^{\infty} i \cdot (1 - clp) \cdot clp^{i-1} = \frac{1}{q} \qquad (2)$$

Siguiendo este modelo, se decidió evaluar un amplio abanico de condiciones de canal que vienen determinadas por una probabilidad de pérdidas y un tamaño medio de ráfaga. Concretamente, se realizaron 30 simulaciones de canal, más el canal limpio (sin pérdida de paquetes), obtenidas de evaluar el modelo con la selección de los parámetros L_{raf} y $P_{per}(\%)$ de los conjuntos $[1\ 2\ 4\ 8\ 12\ 16]$ y $[10\ 20\ 30\ 40\ 50]$, respectivamente.

5. PROCESOS DE EVALUACIÓN

En este apartado se describen los métodos de evaluación utilizados (reconocedor y evaluador perceptual) y se presentan las características de la base de datos de voz utilizadas para llevar a cabo las pruebas.

5.1. Sistema de Reconocimiento de Voz

El reconocedor es el provisto por Aurora-2 [4] y utiliza un modelo HMM (*Hidden Markov Model*) continuo de 16 estados para cada una de las once palabras que reconoce (más el silencio y la pausa, que tienen 3 y 1 estados, respectivamente), con 6 gausianas por estado. Tanto el entrenamiento como el conjunto de datos (base de datos de frases) sobre el que realizar la prueba son extraídos de la base de datos Aurora-2. La fase de entrenamiento se realiza sobre una base de 8440 frases en limpio (no contienen ruido), mientras que la prueba se lleva a cabo con el conjunto *test A* que se encuentra formado por 4004 frases en limpio distribuidas en 4 subconjuntos. El vocabulario está formado por los 11 dígitos comprendidos entre el 0 y el 9 (el cero presenta dos descripciones sonoras 'zero' y 'o'), y la duración media aproximada de las frases (dígitos conectados) es de 1.5 s.

La precisión de reconocimiento es medida mediante la tasa de palabras reconocidas (*word accuracy*, W_{acc}), que establece la relación entre el número de palabras correctamente reconocidas y el número total de palabras.

5.2. Evaluación de la Calidad Perceptual

El método más fiable y extendido para la medición de la calidad perceptual es la escala MOS (*Mean Opinion Score*), donde la calificación MOS de una señal se calcula como la media de las valoraciones dadas por los oyentes. Sin embargo, reunir a un número de oyentes tal y como se especifica en la prueba [5] es ciertamente difícil y, cuanto menos, costoso.

El algoritmo PESQ, desarrollado por la ITU en el año 2001, establece un método automático para la evaluación de la calidad perceptual. Puesto que este algoritmo presenta una exactitud aceptable en la estima de la calidad perceptual en un entorno de pérdida de paquetes para codificadores de voz [6], y dado que existe una correspondencia, mediante el documento [7], entre la nota PESQ y la puntuación MOS, se decidió emplear esta herramienta para la evaluación de la calidad perceptual. Aunque como se ha mostrado existe una correspondencia entre la puntuación PESQ y la MOS, es necesario exponer

que los resultados de este estudio son mantenidos como nota PESQ.

La duración media de las frases del *test A* de Aurora, supone un problema para la evaluación de este algoritmo, ya que, según el documento [6], no se encuentra preparado para evaluar frases de tan corta duración. Las longitudes aconsejadas se encuentran comprendidas en el rango de 8 a 20 s, por lo que se decidió agrupar las frases originales en grupos de 7 (resultando 572 frases), incrementando la duración media aproximada a 12 s, con una duración mínima de 7.5 s y máxima de 19.5 s.

Una vez obtenidos los resultados del algoritmo PESQ para los 572 archivos de la base de datos en cada condición de canal, es necesario promediarlos para obtener el resultado final. A tal efecto, se decidió realizar la media ponderada ($\bar{x}$) en función de la longitud de cada archivo:

$$\bar{x} = \sum_i w_i \cdot x_i \tag{3}$$

$$w_i = \frac{l_i}{L} \quad , \quad L = \sum_i l_i \tag{4}$$

donde x_i y l_i hacen referencia a la nota PESQ y la longitud del archivo i, respectivamente, mientras que w_i es el peso de ese archivo.

6. RESULTADOS EXPERIMENTALES

Debe de tenerse en cuenta en este apartado el resultado en limpio obtenido a partir de las frases de voz limpia $W_{acc} = 99{,}02\,\%$, que establece una cota superior en el desarrollo de las pruebas. Además, hay que hacer notar que los resultados PESQ se encuentran comprendidos en el rango de -0.5 a 4.5, siendo este último el valor obtenido por el algoritmo cuando evalúa la voz limpia.

6.1. Resultados en condición de canal limpio

Codec	*G.723.1*		*G.729*	*iLBC*	
Bit-rate (kbps)	5.3	6.3	8	13.2	15.2
$W_{acc}(\%)$	98.17	98.53	98.64	98.76	98.92
PESQ	3.69	3.80	3.95	3.92	3.94

Tabla 1. *Tabla comparativa de W_{acc} y PESQ obtenidos para los distintos codificadores con canal limpio.*

6.2. Resultados sobre canal con pérdidas

6.2.1. Resultados de Reconocimiento

Tasa de pérdidas	*Long. ráfaga*					
	1	*2*	*4*	*8*	*12*	*16*
10 %	95.90	89.67	87.37	86.60	86.46	86.49
20 %	91.26	80.04	76.82	73.93	74.39	74.57
30 %	85.00	69.74	64.73	62.67	62.61	63.24
40 %	76.70	58.37	53.62	52.46	50.28	51.69
50 %	66.05	46.74	41.62	42.35	41.88	41.42

Tabla 2. *Resultados de $W_{acc}(\%)$ con el codec G.729 (8 kbps), empaquetando 2 tramas (2 x 10 ms) por paquete.*

Tasa de pérdidas	*Long. ráfaga*					
	1	*2*	*4*	*8*	*12*	*16*
10 %	82.45	76.03	74.67	73.86	74.64	74.71
20 %	72.57	64.72	62.58	62.19	62.67	62.80
30 %	64.12	52.34	51.69	51.87	50.63	51.26
40 %	57.00	40.96	39.41	41.00	41.72	42.11
50 %	57.00	40.96	39.41	41.00	41.72	42.11

Tabla 3. *Resultados de $W_{acc}(\%)$ con el codec G.729 (8 kbps), empaquetando 3 tramas (3 x 10 ms) por paquete.*

Tasa de pérdidas	*Long. ráfaga*					
	1	*2*	*4*	*8*	*12*	*16*
10 %	93.84	89.46	86.96	85.93	86.15	86.12
20 %	88.31	80.51	75.61	73.68	73.88	74.09
30 %	81.28	70.71	64.23	61.82	62.60	62.23
40 %	72.51	59.47	53.49	50.77	50.09	51.06
50 %	62.32	48.42	41.15	40.11	40.24	40.80

Tabla 4. *Resultados de $W_{acc}(\%)$ con el codec G.723.1 (5.3 kbps).*

Tasa de pérdidas	*Long. ráfaga*					
	1	*2*	*4*	*8*	*12*	*16*
10 %	94.46	89.85	87.56	86.32	86.40	86.32
20 %	89.32	81.16	76.25	73.95	74.18	74.21
30 %	82.83	71.59	64.70	62.02	62.72	62.20
40 %	74.92	61.07	53.56	51.06	50.56	51.24
50 %	65.52	49.53	41.98	40.57	40.50	41.11

Tabla 5. *Resultados de $W_{acc}(\%)$ con el codec G.723.1 (6.3 kbps).*

Tasa de pérdidas	*Long. ráfaga*					
	1	*2*	*4*	*8*	*12*	*16*
10 %	97.83	96.03	92.88	90.06	89.10	88.56
20 %	96.83	93.25	86.97	80.61	79.15	77.88
30 %	95.63	90.32	80.17	71.81	68.92	67.75
40 %	94.21	86.55	73.71	63.94	58.68	57.10
50 %	93.29	82.36	65.90	55.86	51.03	48.41

Tabla 6. *Resultados de $W_{acc}(\%)$ con el codec iLBC en el modo 20 ms (15.2 kbps).*

Tasa de pérdidas	*Long. ráfaga*					
	1	*2*	*4*	*8*	*12*	*16*
10 %	96.41	93.89	90.76	88.73	87.86	88.03
20 %	93.86	89.39	82.95	78.60	77.43	76.63
30 %	91.25	84.19	74.88	69.06	66.65	65.93
40 %	88.42	78.76	67.19	60.23	55.41	54.79
50 %	85.83	73.51	58.65	50.46	47.64	46.33

Tabla 7. *Resultados de $W_{acc}(\%)$ con el codec iLBC en el modo 30 ms (13.3 kbps).*

6.2.2. Resultados con PESQ

Tasa de pérdidas	*Long. ráfaga*					
	1	*2*	*4*	*8*	*12*	*16*
10 %	2.96	2.79	2.84	2.96	3.03	3.08
20 %	2.55	2.25	2.29	2.39	2.46	2.50
30 %	2.29	1.91	1.88	1.98	2.05	2.09
40 %	2.09	1.66	1.60	1.67	1.71	1.77
50 %	1.91	1.48	1.36	1.43	1.49	1.53

Tabla 8. *Promedios del valor medio ponderado PESQ para el codec G.729 con empaquetado de 2 tramas (2x10 ms) por paquete.*

Tasa de pérdidas	*Long. ráfaga*					
	1	*2*	*4*	*8*	*12*	*16*
10 %	2.85	2.81	2.94	3.04	3.10	3.14
20 %	2.38	2.23	2.35	2.49	2.55	2.59
30 %	2.10	1.85	1.93	2.05	2.14	2.18
40 %	1.91	1.58	1.61	1.75	1.78	1.84
50 %	1.74	1.38	1.36	1.48	1.54	1.58

Tabla 9. *Promedios del valor medio ponderado PESQ para el codec G.729 con empaquetado de 3 tramas (3x10 ms) por paquete.*

Tasa de pérdidas	*Long. ráfaga*					
	1	*2*	*4*	*8*	*12*	*16*
10 %	2.97	2.84	2.85	2.90	2.93	2.95
20 %	2.59	2.37	2.34	2.39	2.42	2.45
30 %	2.34	2.03	1.94	1.98	2.04	2.07
40 %	2.15	1.76	1.64	1.68	1.72	1.76
50 %	2.00	1.55	1.37	1.42	1.48	1.51

Tabla 10. *Promedios del valor medio ponderado PESQ para el codec G.723.1 (5.3 kbps).*

Tasa de pérdidas	*Long. ráfaga*					
	1	*2*	*4*	*8*	*12*	*16*
10 %	3.01	2.88	2.90	2.95	2.99	3.02
20 %	2.62	2.39	2.36	2.42	2.45	2.49
30 %	2.35	2.04	1.96	2.00	2.07	2.09
40 %	2.16	1.77	1.65	1.70	1.73	1.78
50 %	2.00	1.55	1.37	1.42	1.49	1.52

Tabla 11. *Promedios del valor medio ponderado PESQ para el codec G.723.1 (6.3 kbps).*

Tasa de pérdidas	*Long. ráfaga*					
	1	*2*	*4*	*8*	*12*	*16*
10 %	3.28	3.14	3.10	3.11	3.13	3.15
20 %	2.95	2.74	2.65	2.60	2.61	2.62
30 %	2.70	2.46	2.32	2.25	2.24	2.25
40 %	2.51	2.25	2.07	1.99	1.94	1.94
50 %	2.32	2.07	1.85	1.75	1.74	1.72

Tabla 12. *Promedios del valor medio ponderado PESQ para el codec iLBC modo 20 ms (15.2 kbps).*

Tasa de pérdidas	*Long. ráfaga*					
	1	*2*	*4*	*8*	*12*	*16*
10 %	3.22	3.12	3.13	3.14	3.15	3.18
20 %	2.88	2.69	2.64	2.63	2.64	2.67
30 %	2.63	2.39	2.28	2.25	2.27	2.27
40 %	2.46	2.16	2.01	1.97	1.94	1.95
50 %	2.30	1.98	1.78	1.72	1.71	1.71

Tabla 13. *Promedios del valor medio ponderado PESQ para el codec iLBC modo 30 ms (13.3 kbps).*

7. DISCUSIÓN DE RESULTADOS

Partiendo de la condición de canal limpia (sin pérdidas de paquetes), mostrada en la tabla 1, se deben de resaltar los magnificos resultados obtenidos por los codificadores CELP G.729 y G.723.1 que haciendo uso de *bit-rates* inferiores a 8 kbps obtienen buenos resultados tanto en reconocimiento como en calidad. Particularmente, hay que resaltar en esta condición ideal los resultados del codificador G.729, que haciendo uso de un *bit-rate* moderado (8 kbps) obtiene $W_{acc} = 98{,}64$, compaginado con una calificación PESQ de 3.95. De este modo, obtiene similares resultados a iLBC (13.33 y 15.2 kbps), codificador con *bit-rates* superiores.

Cond. Canal	*0*	*1*	*2*	*3*	*4*	*5*
$P_{per}(\%)$	0	10	20	30	40	50
L_{raf}	–	1	2	4	8	16

Tabla 14. *Condiciones de canal representadas.*

Sin embargo, la situación cambia cuando se introducen las pérdidas de paquetes. Para llevar a cabo una comparativa ante esta nueva situación, es necesario agrupar los codificadores de modo que la longitud del segmento de voz correspondiente a un paquete sea igual. Teniendo en cuenta este aspecto, aparecen dos grupos comparativos. Por un lado iLBC (modo 15.2) y G.729 (tomando 2 tramas/paquete), con un empaquetado de 20 ms, mientras que por otro lado se encontrarían G.723.1, iLBC (13.33 kbps) y G.729 (3 tramas/paquete), correspondiéndose este caso con 30 ms.

Es de resaltar el resultado de reconocimiento obtenido por el algoritmo iLBC para pérdidas de un único paquete (columna 1 de las tablas 6 y 7) con $P_{per} = 50\,\%$, donde sólo cae 5.6 % del resultado en limpio en el modo 15.2 kbps (20 ms), frente al 32.6 % de G.729 (2 tramas/paq.) (véase tabla 2). Para el modo 13.3 kbps (30 ms) iLBC cae 12.9 %, frente a la caída del 41.6 % de G.729 (3 tramas/paq.) (tabla 3) y del 35.9 % y 33 % de los modos 5.3 y 6.3 kbps de G.723.1, respectivamente (tablas 4 y 5). iLBC tiene también mejor comportamiento a medida que se aumenta el tamaño de ráfaga, sin embargo, estas diferencias son menos notables.

Paralélamente, los resultados PESQ obtenidos para iLBC son mejores que para el resto de los codificadores. Sin embargo, la calidad perceptual es más sensible a la pérdida de paquetes que la tasa de reconocimiento. Mientras que, como se veía en el parrafo anterior, la degradación causada por $P_{per} = 50\,\%$ y $L_{raf} = 1$ originaba una pérdida del 5.6 % en reconocimiento para iLBC (20 ms), en la nota PESQ provoca una caída de 1.62 puntos (en una escala de 5).

Los valores representados en las figuras 4 y 5 se corresponden con las condiciones de canal expuestas en la tabla 14, siendo este muestreo de las condiciones de canal, en cierto modo, representativo de la degradación sufrida ante porcentajes de pérdidas y tamaños de ráfaga crecientes. En ambas situaciones, tal y como se apuntaba con anterioridad, el codificador iLBC consigue los mejores resultados convirtiéndose en el codificador más robusto ante pérdidas.

En general, los codificadores CELP no obtienen buenos resultados en canales degradados puesto que para la generación de la excitación hacen uso de filtros de *pitch* que propagan el error cometido más allá de los segmentos de voz perdidos. Concretamente, G.729 consigue óptimos

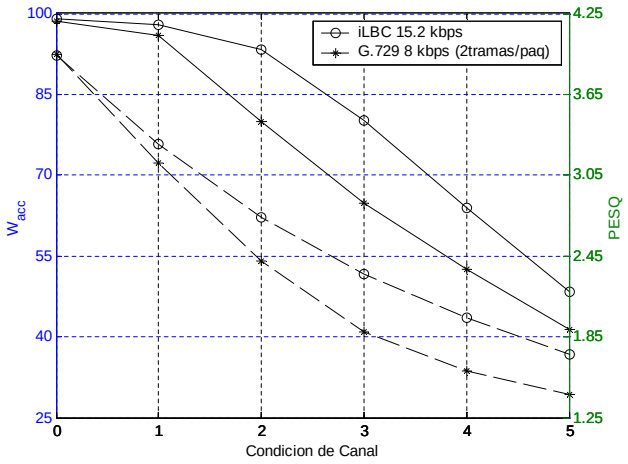

Figura 4. *Comparativa de codificadores con empaquetado de 20 ms. El trazo continuo representa el W_{acc} (escala izquierda), mientas que el discontinuo se corresponde con el promedio PESQ ponderado (escala derecha).*

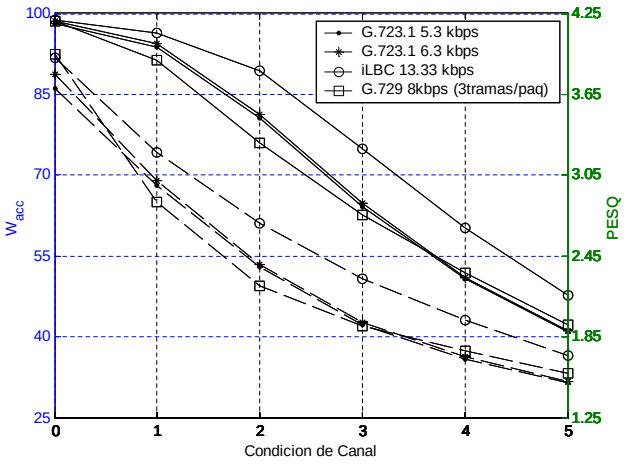

Figura 5. *Comparativa de codificadores con empaquetado de 30 ms. El trazo continuo representa el W_{acc} (escala izquierda), mientas que el discontinuo se corresponde con el promedio PESQ ponderado (escala derecha).*

resultados en limpio transmitiendo los coeficientes LSP, mediante un cuantificador diferencial predictivo. Sin embargo, esta estrategia hace al codec más vulnerable ante la pérdida de tramas consecutivas, puesto que una vez finalizada una ráfaga de pérdidas la predicción LSP del decodificador se verá notablemente degradada, obteniendo peores resultados incluso que el codec G.723.1 que mantiene *bit-rates* inferiores. El codificador iLBC resuelve esta problemática eliminando todo tipo de dependencias intertrama tanto en la generación de la excitación como en la codificación de los coeficientes LSP. El precio pagado es un considerable aumento de la tasa frente a las distintas configuraciones CELP.

8. CONCLUSIONES

En este trabajo se ha presentado una comparativa del rendimiento de los codificadores más extendidos en VoIP, atendiendo a la tasa de reconocimiento a partir de la voz sintetizada y la calidad perceptual de ésta, en un amplio abanico de condiciones de pérdidas de paquetes. Aunque iLBC y G.729 parten de resultados similares para una condición de canal limpia, los resultados obtenidos presentan a iLBC como el codificador más robusto a costa de un mayor *bit-rate*.

9. BIBLIOGRAFÍA

[1] B.W. Wah y B. Sat, "Analysis and evaluation of skype and google talk," Technical report, Department of Electrical and Computer Engineering and the Coordinated Science Laboratory, January 2006.

[2] S.V. Andersen, W.B. Kleijn, R. Hagen, J. Linden, M.N. Murthi, y J. Skoglund, "ilbc - a linear predictive coder with robustness to packet losses," *Proceedings IEEE Speech Coding Workshop*, pp. 23–25, 2002.

[3] M. Yajnik, S. Moon, J. Kurose, y D. Towsley, "Measurement and modelling of the temporal dependence in packet loss," *Proceedings of IEEE INFOCOM*, 1999.

[4] D. Pearce y H. G. Hirsch, "The aurora experimental framework for the performance evaluations of speech recognition systems under noisy conditions," *Proc. ICSLP'00*, vol. 4, pp. 29–32, 2000.

[5] Recommendation ITU-T P.830, "Subjective performance assessment of telephone-band and wideband digital codecs," 1996.

[6] Recomendación UIT-T P.862, "Evaluación de la calidad vocal por percepción: Un método objetivo para la evaluación de la calidad vocal de extremo a extremo de redes telefónicas de banda estrecha y codecs vocales," *Serie P.800*, Febrero 2001.

[7] Recomendación UIT-T P.862.1, "Función de correspondencia para convertir los resultados brutos de la prueba p.862 en nota media de opinión de la calidad de escucha objetiva," *Serie P.800*, Noviembre 2003.