Artículo de Revisión de Tema

Parametrizaciones robustas de Reconocimiento Automático de Habla (RAH) en redes de comunicaciones

Robust parameterizations of Automatic Speech Recognition (ASR) in communications networks

Diego Ferney Gómez Cajas, Franklin Alexander Sepúlveda Sepúlveda**, Mario Augusto Pinto Serrano****

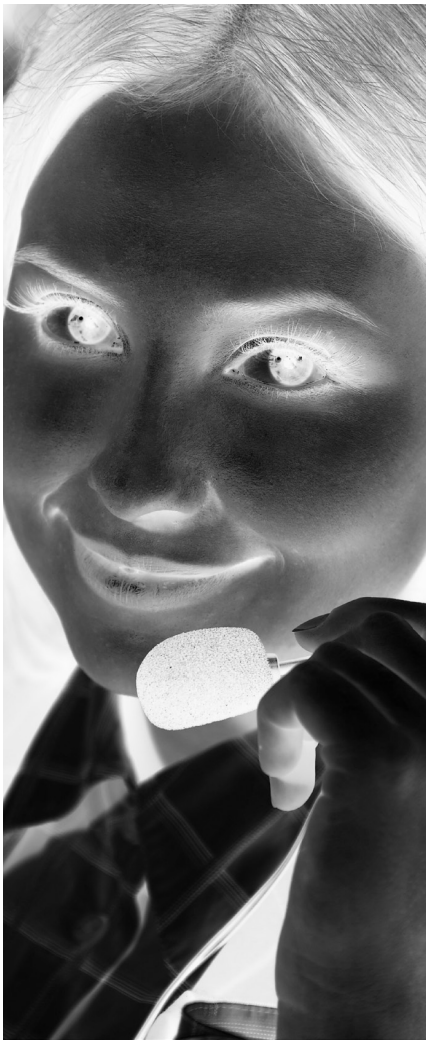
Abstract

In this paper we address the problem of Automatic Speech Recognition (ASR) when the speech signal has been transmitted over communications networks. In these conditions, the main causes of distortion in an ASR system are ambient noise, transmission errors and the encoding-decoding process [32]. In the literature we are able to find multiple solutions for this problem, from different points of views; however, in this paper we will focus the analysis on solutions with robust parameterizations for the above distortions.

Keywords: ASR, Speech Coding, CELP coders, packet networks, VoIP, transmission errors, packet loss, noise, mobile networks, UMTS, LTE.

Resumen

En este artículo abordamos la problemática de los sistemas de Reconocimiento Automático de Habla (RAH) cuando la señal de voz ha sido transmitida por una red de comunicaciones. Bajo este escenario, las distorsiones que más deterioran el funcionamiento del sistema de RAH son el ruido de ambiente, los errores de transmisión y la distorsión por el proceso de codificación-decodificación [32]. En la literatura se encuentran diversas soluciones que atacan el problema desde diferentes puntos de vista; no obstante, en este artículo centraremos nuestro análisis en las soluciones orientadas a conseguir parametrizaciones robustas bajo las distorsiones que hemos mencionado.



* PhD. en Multimedia y Comunicaciones, Universidad Carlos III de Madrid. Profesor Asistente, Universidad Antonio Nariño. dfgomezc@uan.edu.co

** PhD. en Ingeniería, Universidad Nacional de Colombia, Profesor en Universidad Industrial de Santander. alexander.sepulveda.sepulveda@gmail.com

*** Esp. en Automatización Industrial, Universidad de Ibagué, Estudiante de MSc. en Ingeniería - telecomunicaciones, Universidad Nacional de Colombia. Profesor en Universidad Antonio Nariño. mapintos@unal.edu.co

1. Introducción

Uno de los aspectos más importantes a tener en cuenta en el desarrollo de aplicaciones de RAH robusto es la parametrización [13][32]. En ésta, se extraen las características que permitan obtener una representación compacta de la voz pero que a su vez permitan obtener un alto desempeño del sistema de RAH.

Sin embargo, cuando la voz ha sido codificada (como en la mayoría de sistemas de transmisión de voz modernos), la voz está expuesta a diferentes tipos de distorsiones [5][47][35][24], tales como el ruido de ambiente, los errores de transmisión, la pérdida de paquetes, los errores debidos al proceso de codificación-decodificación, entre otros; que hacen que el rendimiento del sistema de RAH disminuya de forma considerable [27][45][29].

Dado lo anterior, existen en la literatura diversos trabajos que buscan obtener parametrizaciones robustas frente a este tipo de distorsiones. En este artículo realizaremos una exposición de algunos de los desarrollos más importantes en este ámbito.

2. Reconocimiento Automático de Habla en entornos de voz codificada

En algunos tipos de redes, la transmisión de voz se realiza utilizando codificadores de forma de onda del tipo PCM (*Pulse Code Modulation*) [20]; sin embargo, este tipo de codificación implica un elevado ancho de banda [4], y aunque consiguen una alta calidad en la señal de voz reconstruida [7], el sistema de transmisión resultante es muy costoso, especialmente en las redes en donde el ancho de banda es un recurso limitado (como es el caso de las comunicaciones inalámbricas). Es por este motivo, que la tendencia es a utilizar codificadores más eficientes en cuanto a la compresión de los parámetros generados en la codificación de la voz, pues en un sistema de telecomunicaciones moderno, se utilizan diferentes medios de transmisión (incluidos los medios blandos o inalámbricos) que hace que el ancho de banda sea una preocupación constante.

Dado lo anterior, aunque los codificadores con mayor eficiencia de compresión utilicen tasas binarias más bajas, pueden alcanzar calidades subjetivas similares a los codificadores de forma de onda de tasas binarias altas [21][22].

Dentro de las familias de codificadores que consiguen una buena relación tasa binaria vs calidad subjetiva, se encuentran los codificadores híbridos [25], y entre estos, uno de los algoritmos más utilizados en los actuales estándares de codificación, es el CELP (*CodeExcited Linear Prediction*) [40]. Por este motivo, centraremos nuestro análisis en este tipo de codificadores, describiendo especialmente la parametrización que se lleva a cabo.

3. Codificación y reconocimiento de voz: similitudes y diferencia

La parametrización es un proceso común tanto a la codificación como a los sistemas de RAH. Para entender mejor esta relación, nos podemos remontar al análisis fuente – filtro que utilizan los dos tipos de parametrización [30].

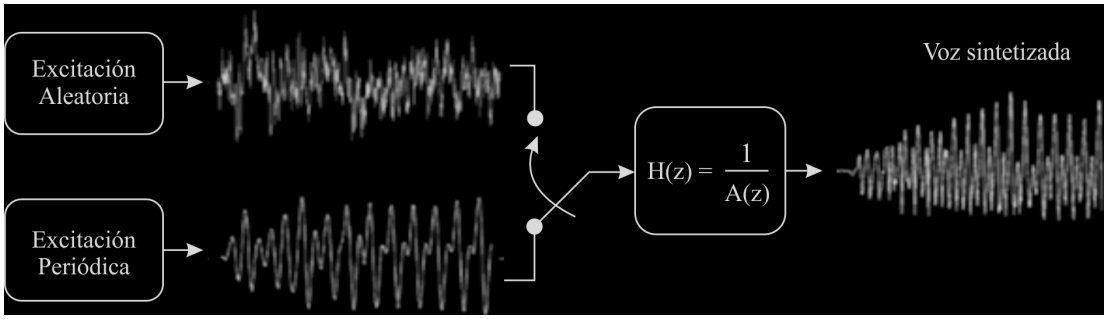
Modelo Fuente Filtro

Actualmente, la codificación de voz es utilizada fundamentalmente en las redes de telefonía móvil, y en la transmisión de voz sobre redes IP (Voiceover IP - VoIP) [32][13][17]. En estos codificadores, y concretamente en los codificadores tipo CELP, se hace uso del Modelo Fuente-Filtro que se ilustra en la Figura 1.

En este modelo, se simula el Sistema Fonador Humano [13][8][26], utilizando dos señales de excitación que modelan las ondas que se generan por el paso del aire a través de la laringe y faringe (incluidas las cuerdas vocales). De otro lado, el tracto vocal es caracterizado por un filtro que se encarga de sintetizar la señal de voz a partir de las dos componentes de excitación anteriores.

Bajo este modelo, el proceso de codificación busca obtener un conjunto de parámetros que caractericen las dos componentes principales del modelo:

Figura 1. Modelo Fuente Filtro



Fuente:

Modelada utilizando dos señales, una de naturaleza estocástica (Excitación Aleatoria), y otra de naturaleza determinística (Excitación Periódica).

Filtro:

Utilizan en su forma más esencial, un filtro todo polos cuyos coeficientes son obtenidos utilizando un Análisis de Predicción Lineal [40][28].Es por esto que dichos coeficientes son conocidos como LPC (*Linear Prediction Coefficients*) [25].

Parametrización en Codificación.

La parametrización que caracteriza tanto la información del filtro como de la excitación (o fuente), persigue dos objetivos desde el punto de vista de la codificación:

- Modelar la señal de voz con la menor pérdida de calidad perceptual posible.
- Reducir la tasa binaria necesaria para la codificación de los parámetros que modelan la señal de voz.

Codificación CELP

Entre los codificadores que utilizan rangos bajos de tasas binarias, uno de los algoritmos más populares es el CELP [46][50]. En éstos, el modelado de la Envolvente Espectral se realiza mediante el Análisis de Predicción Lineal (conocido también como Análisis LPC). No obstante, los LPC, no tienen unas buenas características para ser codificados y transmitidos, y por esta razón, en lugar de los LPC, el codificador codifica y transmite los denominados LSP (*Line Spectrum Pairs*)[42], pues estos últimos, entre otras ventajas, pueden ser cuantificados e interpolados de una manera más eficiente que los LPC. De otro lado, los coeficientes LSP, facilitan el análisis de estabilidad del filtro todo polos que caracterizará el tracto vocal del Modelo Fuente – Filtro [2][44].

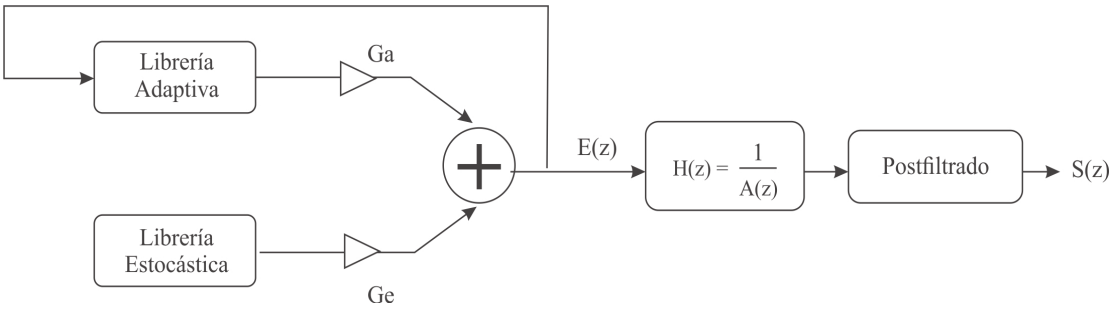
Es de destacar que la información que contienen los LSP es muy relevante para la tarea de reconstrucción de la voz en el proceso de decodificación [16], y por tanto, los codificadores que utilizan el esquema CELP utilizan un alto porcentaje de los bits transmitidos para codificar estos parámetros [3][21][22]. Además de lo anterior, en los sistemas de comunicación que utilizan codificación de canal (p. e. en telefonía móvil), los LSP suelen tener la más alta prioridad a la hora de brindarles protección [16].

De otro lado, en cuanto a la parametrización de la excitación, el CELP utiliza dos librerías de códigos, una denominada adaptativa (que modela la componente de excitación periódica) [21], y otra denominada estocástica (también llamada librería de códigos fijos) [13].

En el caso de la componente adaptativa, los parámetros a codificar son el periodo fundamental o *pitch* (T) y la ganancia de la librería adaptativa (G_a). Con estos parámetros se puede determinar el Vector de Códigos Adaptativos (V_p) que será el que modele la componente periódica de la excitación.

En cuanto a la librería estocástica, ésta se construye eliminando la componente adaptativa del

Figura 2. Estructura básica de un codificador CELP.



residuo de predicción del análisis LPC, de tal forma que los parámetros que modelan la componente estocástica son el Vector de Códigos Fijos (V_c) y su ganancia (G_c) [25].

Adicionalmente, es común utilizar en los codificadores CELP, una etapa de postfiltrado que busca mejorar la calidad de la voz sintetizada.

La Figura 2 ilustra las componentes básicas del codificador CELP.

4. Parametrización para reconocimiento

En el caso de la parametrización para reconocimiento, lo que se busca obtener un vector de características que represente de forma compacta de la voz reduciendo la variabilidad lingüística. Por lo tanto, una buena parametrización debería de un lado, captar las características más relevantes para el sistema de RAH (información del tracto vocal, evolución de la energía, etc.) y de otro lado, descartar información no relevante para el proceso de reconocimiento, concretamente, en una tarea de reconocimiento independiente de locutor, se busca eliminar la información de variabilidad acústica entre una persona y otra para dar mayor generalidad a la tarea de reconocimiento.

En este sentido, existen diferentes esquemas de parametrización; sin embargo los denominados cepstrum son - con diferencia -, lo más utilizados para modelar la envolvente espectral [32], y dentro de este tipo de parámetros, los MFCC (*Mel-Frequency Cepstral Coefficients*) o Coeficientes Cepstrales en escala Mel, los que más se utilizan [49][31]. Habitualmente, los MFCC se acompañan de otros parámetros como sus derivadas [38] [10], la energía [49][33][11], información de la excitación [16][12][19], etc.

Obtención de los MFCC

Cuando se utilizan Modelos Ocultos de Markov (*HMM – Hidden Markov Models*) en el modelado acústico [38][37], es muy conveniente utilizar los MFCC porque se puede asumir que están no correlacionados entre si y por tanto, facilitan la estimación de los parámetros del modelo [10].

Para la obtención de los MFCC existen clásicamente dos formas:

- Deconvolución Homomórfica.
- Predicción Lineal.

Los dos casos se basan en el modelo Fuente Filtro que se explicó en la Figura 1, donde los MFCC modelaran la envolvente espectral contenida en la respuesta en frecuencia del filtro solo polos. A continuación se explicarán los dos métodos.

Cepstrum por Deconvolución Homomórfica

En este caso, el cepstrum se obtiene a partir del logaritmo del espectro de la señal de voz. En la práctica, para obtener el espectro se utiliza la Transformada Discreta de Fourier (*DFT -Discrete Fourier Transform*) y a su módulo se le aplica el logaritmo (cepstrum real). Finalmente, con la Transformada Discreta de Coseno (*DCT – Discrete Cosine Transform*) se obtuvo el cepstrum de la señal de voz [41][13].

Es de destacar, que en el cepstrum obtenido para una señal de voz, la componente de la envolvente espectral decae rápidamente con, de tal manera que para obtener los Coeficientes Cepstrales podemos realizar un procedimiento denominado liftrado (filtrado en el dominio cepstral) para quedarse con los primeros valores de que representarán la envolvente espectral. Usualmente, en el liftrado se utilizan sólo los 12 primeros valores [32][41].

El cepstrum también puede ser obtenido en la escala MEL utilizando un banco de filtros sobre el espectro de la señal de voz [43], en este caso, los coeficientes obtenidos son los MFCC. Adicionalmente, se puede calcular la log-energía a partir de las muestras de voz directamente, y los denominados parámetros dinámicos que se calculan a partir de los parámetros anteriores.

Un resumen del procedimiento descrito se puede apreciar en la Figura 3.

Cepstrum a partir de análisis LPC.

En el procedimiento anterior, los coeficientes cepstrales que caracterizan el tracto vocal fueron obtenidos a partir de la señal de voz por deconvolución de sus dos componentes de acuerdo al modelo fuente-filtro. De esta manera, fueron separados los coeficientes que corresponden a la fuente (o excitación) de los que corresponden al filtro (tracto vocal) utilizando el liftrado. No obstante, si a la señal de voz se le aplica un proceso de predicción lineal, se puede separar también la información referente al tracto vocal, de la que corresponde a la excitación, siendo esta última el residuo del proceso de predicción.

Dado lo anterior, utilizando los coeficientes del predictor lineal (LPC) podemos obtener su función de transferencia y a partir de ella, su espectro y por tanto, sus coeficientes cepstrales [32]. No obstante, también es posible calcular directamente el cepstrum a partir de los coeficientes de predicción lineal utilizando la recursión mostrada en [1].

Finalmente, dado que los coeficientes cepstrales son obtenidos a partir del análisis LP, el cepstrum así obtenido es comúnmente llamado, cepstrum LP [32][13].

De forma similar a la obtención de por deconvolución homomórfica, es común añadir la log-energía calculada a partir de las muestras de la señal de voz reconstruida así como los parámetros dinámicos, tal como se observa en la Figura 4.

Cepstrum a partir del bitstream

En el apartado anterior, se describió la forma de obtener el cepstrum a partir de la información del filtro que caracteriza el tracto vocal. No obstante, como se describió en la sección de codificación, en un ambiente de voz codificada, la señal de voz reconstruida (en la etapa receptora) será el resultado de un proceso de síntesis de la señal de excitación en su paso a través del filtro

Figura 3. MFCC por Deconvolución Homomórfica

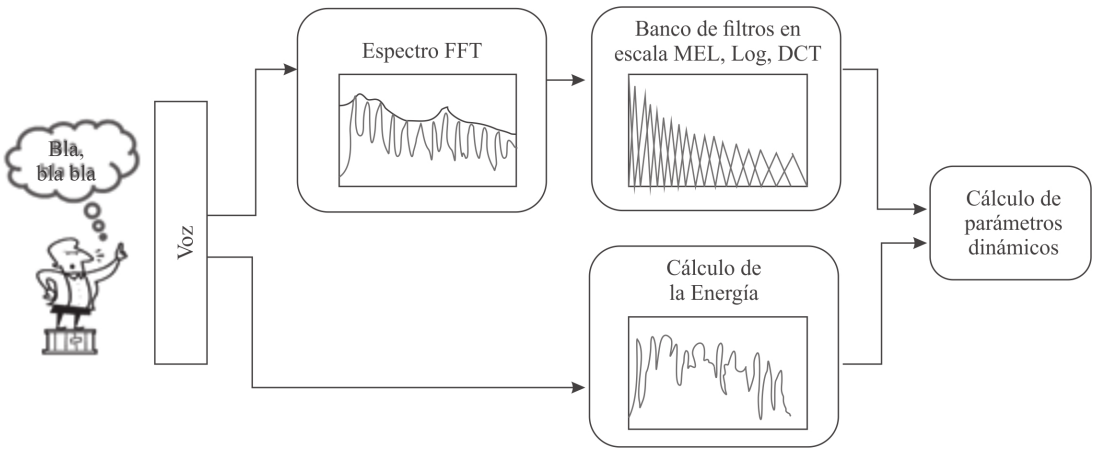
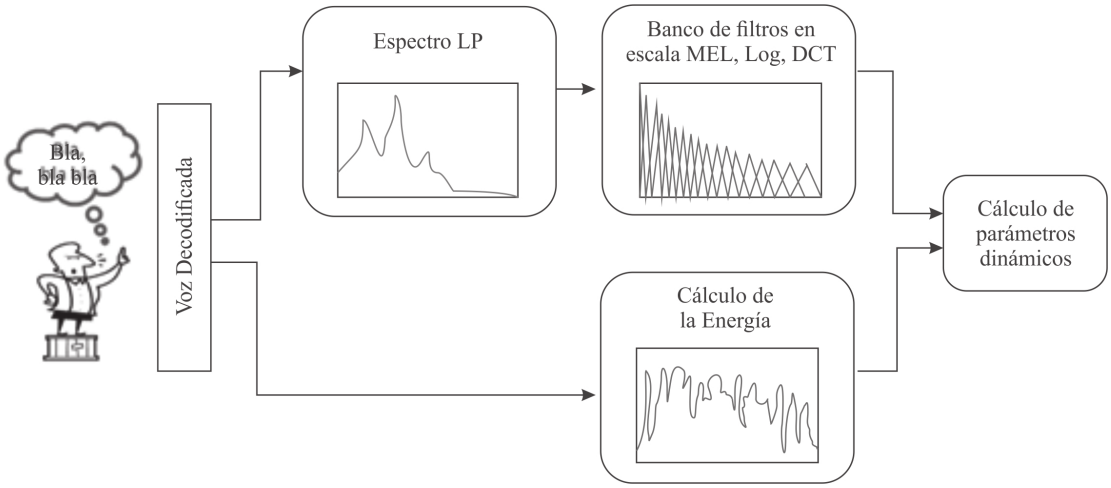


Figura 4. Cepstrum LP a partir de voz reconstruida (decodificada)



modelado en el codificador (en la etapa transmisora).

Lo anterior implica que el análisis de predicción lineal que se realiza para obtener el cepstrum, parta de la señal reconstruida en el receptor; y por tanto, involucre el uso de los parámetros de la excitación enviados por el codificador. Lo anterior hace que la señal de voz reconstruida pueda verse degradada por las distorsiones presentes en el canal de comunicaciones, y especialmente aquellas distorsiones producidas en los parámetros de la excitación que suelen ser menos protegidos, respecto de los parámetros que caracterizan el filtro [16][18].

Por estas razones, existe una solución alternativa que consiste en obtener el cepstrum directamente a partir de los parámetros enviados por el codificador, es decir, sin realizar la reconstrucción de la señal de voz en recepción.

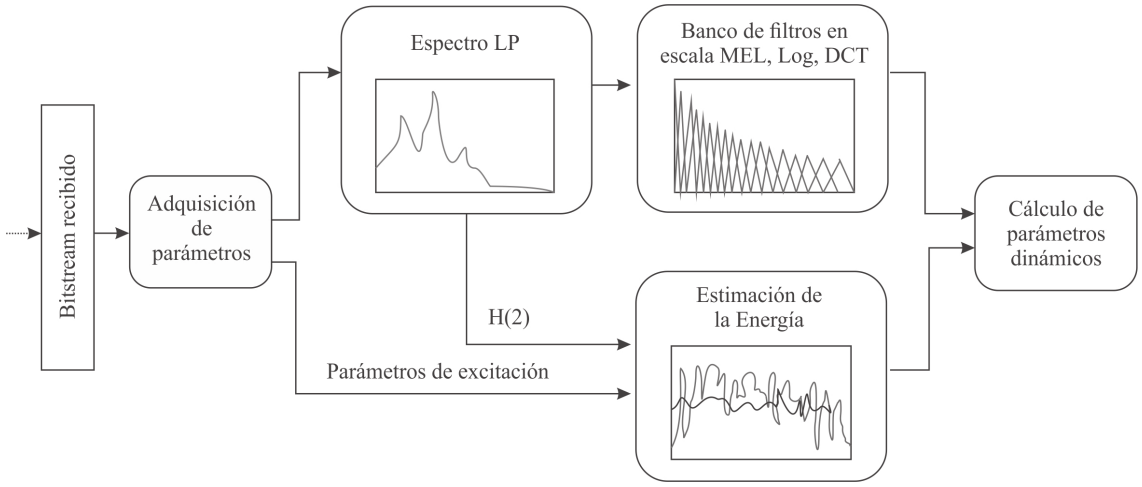
A esta técnica se le conoce como *Transparametrización* [32][13][23], dado que transforma los parámetros orientados a la codificación en parámetros orientados al reconocimiento. Sin embargo, en nomenclatura de habla inglesa, es común referirse a esta aproximación como "*bitstream based*" [23][11][12], pues el cepstrum LP se calcula a partir de los coeficientes de predicción lineal extraídos del bitstream enviado por el codificador.

El procedimiento es descrito con detalle en [33] [32], en donde además se describen los procesos para obtener algunos parámetros adicionales que conforman el vector de características de reconocimiento, tales como la energía [33], el periodo fundamental [13], etc. No obstante, el procedimiento tradicional para construir el vector de características se puede observar en la Figura 5, en donde es de destacar que la energía deber ser estimada a partir de los parámetros obtenidos del bitstream (excitación y envolvente espectral).

5. Distorsiones asociadas al RAH sobre redes de comunicaciones

Uno de los aspectos de especial importancia a tener en cuenta en un sistema de RAH sobre una red de comunicaciones, es la robustez que éste tenga frente a las distorsiones típicas a las que se enfrenta la señal de voz en su paso por la red. En este sentido, diversos autores coinciden en que las distorsiones más importantes que afectan el rendimiento del sistema de RAH son: los efectos de la codificación-decodificación [7] [27][32], los errores de transmisión [45][19][11] [12] y el ruido de ambiente [48][9]. No obstante, con los codificadores actuales, los errores de transmisión y el ruido de ambiente son los que más repercuten en la disminución de la tasa de error en el sistema de RAH [13].

Figura 5. Ceptrum LP mediante transparametrización



De otro lado, en cuanto a los errores de transmisión, estos se tornan de manera diferente en función de la red de comunicación a utilizar [19] [35], siendo la pérdida de paquetes en ráfaga, lo que más deteriora el funcionamiento de la tarea de RAH en un ambiente de transmisión de voz sobre IP [36][35][34], y de igual forma, la pérdida de ráfagas de bits en una transmisión de voz sobre una red de telefonía móvil [39][11]. En este sentido, diversos trabajos muestran como los cepstrum LP, en sus dos versiones, se muestran muy robustos frente a los errores de transmisión y frente a la distorsión por codificación; sin embargo, el cepstrum obtenido por deconvolución homomórfica no se muestra tan robusto [14][15] [32][13] .

Es de destacar también que en las redes de transmisión de voz que utilizan codificación de canal [51][13], concretamente las redes de telefonía celular como UTMS o LTE [51][52], éstas utilizan un esquema de protección desigual (UEP) [53], que introduce una protección más rigurosa a los parámetros que caracterizan la envolvente espectral, en deterioro de la protección de los parámetros de la excitación. Dado lo anterior, el procedimiento de transparametrización puede obtener un especial provecho, pues para la construcción del vector de características podría utilizar solamente los parámetros que son más protegidos, evitando utilizar los menos protegidos y así evitar también la degradación que estos introducen en la voz reconstruida y por tanto en el sistema de RAH. Lo anterior se puede observar en [16], en donde tienen en cuenta este esquema de protección desigual para conseguir mayor robustez en el sistema de reconocimiento utilizando la transparametrización, dado que esta última permite realizar una selección de los parámetros a utilizar para la construcción del vector de características, en comparación con las técnicas que utilizan voz reconstruida.

Por último, en cuanto al ruido de ambiente, si bien la transparametrización se muestra también muy robusta cuando el ruido se presenta en un efecto combinado con los errores de transmisión, no pasa lo mismo cuando el ruido se considera como la principal distorsión (en compañía de la distorsión por codificación que está presente en el entorno de análisis de este trabajo), pues en este caso, las dos aproximaciones que utilizan la señal de voz reconstruida, presentan un mejor comportamiento frente a este importante tipo de distorsión [13].

6. Conclusiones

Los sistemas de reconocimiento automático de habla que utilizan voz transmitida sobre una red de comunicaciones, deben tener en cuenta - entre otros aspectos relevantes - los procesos de codificación que se llevan a cabo, destacando especialmente la codificación de fuente que com-

prime la señal de voz para generar un ahorro en el ancho de banda necesario para su transmisión. No obstante, la codificación de canal también ha de ser tenida en cuenta, pues los esquemas típicos de protección frente a errores presentes en los sistemas de comunicación modernos, dan prioridad a la protección de algunos parámetros a transmitir (aquellos que caracterizan la envolvente espectral), en detrimento de otros (aquellos que caracterizan la excitación), lo que hace que las técnicas de reconocimiento que reconstruyen la señal de voz con todos los parámetros transmitidos se vean expuestas a los errores generados por los efectos adversos del canal que deterioran los parámetros menos protegidos.

De esta manera, tal como lo han advertido diversos trabajos previos, los efectos de la codificación deben ser tenidos en cuenta más allá de los efectos que introducen los errores de cuantificación de los codificadores de forma de onda tradicionales, pues con las técnicas de codificación moderna, se deben tener en cuenta también los nuevos efectos adversos que éstas introducen, así como las distorsiones tradicionales que introducen el ruido de ambiente, los errores de transmisión, entre otros.

En este sentido, las nuevas propuestas de transparametrización de parámetros se muestran robustas frente a las distorsiones mencionadas, pues de un lado evitan el coste computacional asociado al proceso de decodificación, y de otro lado, excluyen los parámetros que no son relevantes para la tarea de reconocimiento, o en su defecto le dan un mejor uso, que la tradicional reconstrucción de la voz a partir de todos los parámetros enviados por el codificador. La transparametrización se destaca especialmente en redes que utilizan codificación de canal, como las redes de telefonía móvil, y en entornos en donde los errores de transmisión en ráfagas son la principal causa de distorsión; sin embargo, en presencia de ruido de ambiente, como principal causa de distorsión, el uso del cepstrum LP obtenido a partir de voz reconstruida, se muestra muy robusto, incluso en diferentes tipos de ruido. Dado lo anterior, el uso del cepstrum LP (utilizando voz reconstruida o no), se muestra más robusto frente a las principales causas de distorsión, respecto del cepstrum obtenido por el tradicional método de deconvolución homomórfica.

7. Referencias

1] Atal, B. S. (1974). Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. the Journal of the Acoustical Society of America, 55, 1304.

2] Atal, B. S., Cox, R. V., & Kroon, P. (1989, May). Spectral quantization and interpolation for CELP coders. In Acoustics, Speech, and Signal Processing, 1989. ICASSP-89., 1989 International Conference on (pp. 69-72). IEEE.

3] Bessette, B., Salami, R., Lefebvre, R., Jelinek, M., Rotola-Pukkila, J., Vainio, J., ...&Jarvinen, K. (2002). The adaptive multirate wideband speech codec (AMR-WB). Speech and Audio Processing, IEEE Transactionson, 10(8), 620-636.

4] Carlson, A. B., & Contreras, J. R. S. (1980). Sistemas de comunicación. McGraw-Hill.

5] Chia-Ping Chen; Bilmes, J.; Ellis, D.P.W., "Speech Feature Smoothing for Robust ASR," *Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05). IEEE International Conference on* , vol.1, no., pp.525,528, March 18-23, 2005.

6] De Vicente Peña, J. (2007). Contribuciones al reconocimiento robusto de habla.

7] Euler, S., &Zinke, J. (1994, April). The influence of speech coding algorithms on automatic speech recognition. In Acoustics, Speech, and Signal Processing, 1994. ICASSP-94., 1994 IEEE International Conference on (Vol. 1, pp. I-621). IEEE.

8] Fant, G. (1982). The voice source-acoustic modeling. STLQPSR, 4, 28-48.

9] Flynn, R., & Jones, E. (2010). Robust distributed speech recognition in noise and packet loss conditions. Digital SignalProcessing, 20(6), 1559-1571.

10] Furui, S. (1986). Speaker-independent isolated word recognition using dynamic features of speech spectrum. Acoustics, Speech

and SignalProcessing, IEEE Transactionson, 34(1), 52-59.

11] Gallardo-Antolín, A., Peláez-Moreno, C., & Díaz-de-María, F. (2005). Recognizing GSM digital speech. Speech and Audio Processing, IEEE Transactions on, 13(6), 1186-1205.

12] Gómez, A. M., Peinado, A. M., Sánchez, V., & Rubio, A. J. (2006). Recognition of coded speech transmitted over wireless channels. Wireless Communications, IEEE Transactions on, 5(9), 2555-2562.

13] Gómez-Cajas, D. F. (2011). Contribuciones al reconocimiento robusto de habla en redes de comunicaciones mediante transparametrización: tesis doctoral (Doctoral dissertation, Universidad Carlos III de Madrid).

14] Gómez-Cajas, D. F., Peláez-Moreno, C., & Díaz-de-María, F. (2003) Reconocimiento robusto de habla en redes IP. In Actas de las XIII JORNADAS de I+D en Telecomunicaciones (TELECOMI+D+03) Madrid, España.

15] Gómez-Cajas, D. F., Peláez-Moreno, C., & Díaz-de-María, F. (2003) Reconocimiento robusto de habla en entornos IP. In Proceedings of the International Conference on Internet Technologies, Popayán, Colombia.

16] Gómez-Cajas, D. F., Peláez-Moreno, C., & Díaz-de-María, F. (2012, November). UEP-driven extended feature extraction for ASR over 3G speech channels. In Circuits and Systems (CWCAS), 2012 IEEE 4th Colombian Workshop on (pp. 1-5). IEEE.

17] Goode, B. (2002). Voice over internet protocol (VoIP). Proceedings of the IEEE,90(9), 1495-1517.

18] Toskala, A., & Holma, H. (Eds.). (2001). WCDMA for UMTS: Radio Access for Third Generation Mobile Communications. Wiley. Huerta, J. M., & Stern, R. M. (1998, November). Speech recognition from GSM codec parameters. In ICSLP.

19] ITU - Rec. G.711 (1993). Pulse code modulation (PCM) of voice frequencies.International Telecommunications Union, February.

20] ITU - Rec. G.729 (1996). Coding of speech at 8 kbit/s using conjugate structure algebraic-code-excited linear-prediction (CS-ACELP).

21] ITU - Rec.G723 (1996). Dual rate speech coder for multimedia communications transmitting at 5.3 and 6.3 kbit/s.

22] Kim, H. K., & Cox, R. V. (2001). A bitstream-based front-end for wireless speech recognition on IS-136 communications system. Speech and Audio Processing, IEEE Transactionson, 9(5), 558-568.

23] Kim, H. K., Kim, K. C., & Lee, H. S. (1993). Enhanced distance measure for LSP-based speech recognition. Electronics letters, 29(16), 1463-1465.

24] Kondoz, A. M. (2005). Digital speech: coding for low bit rate communication systems. Wiley.

25] Lieberman, P. (1988). Speech physiology, speech perception, and acoustic phonetics. Cambridge University Press.

26] Lilly, B. T., &Paliwal, K. K. (1996, October). Effect of speech coders on speech recognition performance. In Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on (Vol. 4, pp. 2344-2347). IEEE.

27] Makhoul, J. (1975). Linear prediction: A tutorial review. Proceedings of the IEEE, 63(4), 561-580.

28] Milner, B., &Semnani, S. (2000). Robust speech recognition over IP networks. In Acoustics, Speech, and Signal Processing, 2000. ICASSP›00. Proceedings. 2000 IEEE International Conference on (Vol. 3, pp. 1791-1794). IEEE.

29] Müller, J., andBaly, W. The Physiology of the Senses, Voice, and MuscularMotion, with the Mental Faculties... Taylor, Walton &Maberly, 1848.

30] Nishimura, Y., Shinozaki, T., Iwano, K., &Furui, S. (2004). Noise-robust speech recognition using multi-band spectral features. TheJournal of theAcousticalSociety of America, 116, 2480.

31] Peláez-Moreno, C. (2002). Reconocimiento de habla mediante transparametrización: una alternativa robusta para entornos móviles e IP (Doctoral dissertation, Universidad Carlos III de Madrid).

32] Peláez-Moreno, C., Gallardo-Antolín, A., & Díaz-de-María, F. (2001). Recognizing voice over IP: A robust front-end for speech recog-

nition on the World Wide Web. Multimedia, IEEE Transactionson, 3(2), 209-218.

33] Peláez-Moreno, C., Gallardo-Antolín, A., Gómez-Cajas, D. F., & Díaz-de-María, F. (2006). A comparison of front-ends for bitstream-based ASR over IP. Signal Processing, 86(7), 1502-1508.

34] Pelaez-Moreno, C.; Gallardo-Antolin, A.; Diaz-de-Maria, F., "Recognizing voice over IP: a robust front-end for speech recognition on the world wide web," Multimedia, IEEE Transactions on , vol.3, no.2, pp.209,218, Jun 2001.

35] Perkins, C., Hodson, O., & Hardman, V. (1998). A survey of packet loss recovery techniques for streaming audio. Network, IEEE, 12(5), 40-48.

36] Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE, 77(2), 257-286.

37] Rabiner, L., &Juang, B. H. (1993). Fundamentals of speechrecognition.

38] Rose, R. C., Parthasarathy, S., Gajic, B., Rosenberg, A. E., & Narayanan, S. (2001). On the implementation of ASR algorithms for hand-held wireless mobile devices. In Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on (Vol. 1, pp. 17-20). IEEE.

39] Schroeder, M., &Atal, B. S. (1985, April). Code-excited linear prediction (CELP): High-quality speech at very low bit rates. In Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'85. (Vol. 10, pp. 937-940). IEEE.

40] SMITH, J. O. (2008). Spectral Audio Signal Processing, Draft. Online: http://ccrma.stanford.edu/ jos/sasp/.

41] Soong, F., &Juang, B. (1984, March). Line spectrum pair (LSP) and speech data compression. In Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'84. (Vol. 9, pp. 37-40). IEEE.

42] Stevens, S. S., Volkmann, J., & Newman, E. B. (1937). A scale for the measurement of the psychological magnitude pitch TheJournal of theAcousticalSociety of America, 8, 185.

43] Sugamura, N., &Itakura, F. (1986). Speech analysis and synthesis methods developed at ECL in NTT—From LPC to LSP—. Speechcommunication, 5(2), 199-215.

44] Sun, L., Wade, G., Lines, B., &Ifeachor, E. (2001, April). Impact of packet loss location on perceived speech quality. In 2nd IP-Telephony Workshop (pp. 114-122).

45] Toga, J., &Ott, J. (1999). ITU-T standardization activities for interactive multimedia communications on packet-based networks: H. 323 and related recommendations. Computer Networks, 31(3), 205-223.

46] Tyagi, V.; McCowan, I.; Misra, H.; Bourlard, H., «Mel-cepstrum modulation spectrum (MCMS) features for robust ASR,» Automatic Speech Recognition and Understanding, 2003. ASRU ‹03. 2003 IEEE Workshop on, vol., no., pp.399,404, 30 Nov.-3 Dec. 2003

47] Vicente-Peña, J., Gallardo-Antolín, A., Peláez-Moreno, C., & Díaz-de-María, F. (2006). Band-pass filtering of the time sequences of spectral parameters for robust wireless speech recognition. Speechcommunication, 48(10), 1379-1398.

48] Zheng, F., Zhang, G., & Song, Z. (2001). Comparison of different implementations of MFCC. Journal of ComputerScience and Technology,16(6), 582-589.

49] Zhou, J., Wu, T., &Leng, J. (2010, June). Research on voice codec algorithms of SIP phone based on embedded system. In Wireless Communications, Networking and Information Security (WCNIS), 2010 IEEE International Conference on (pp. 183-187). IEEE.

50] Ziemer, R. E., Tranter, W. H., Buehrer, R. M., & Rappaport, T. S. (2000).Mobile Radio Communications. John Wiley & Sons, Inc.

51] 3GPP TSG-RAN, «3GPP TR 25.814, Physical Layer Aspects for Evolved UTRA (Release 7)», v1.3.1 (2006-05).

52] 3GPP TS 25.212, «Multiplexing and channel coding (FDD)». V6.2.0. 2004-06.

53] 3GPP TS 25.211, «Physical channels and mapping of transport channels onto physical channels (FDD)».V6.1.0. 2004-6.