CUANTIFICACIÓN VECTORIAL DIFERENCIAL PARA LA TRANSMISIÓN EFICIENTE DE PARÁMETROS ACÚSTICOS EN SISTEMAS DE RECONOCIMIENTO AUTOMÁTICO DEL HABLA DISTRIBUIDO

José Enrique García, Alfonso Ortega, Antonio Miguel y Eduardo Lleida.

Grupo de Tecnologías de las Comunicaciones (GTC)
I3A, Universidad de Zaragoza
{jegarlai,ortega,amiguel,lleida}@unizar.es

RESUMEN

El reconocimiento automático del habla distribuido surge como solución a limitaciones de capacidad computacional en dispositivos portátiles de uso cotidiano como teléfonos móviles o PDAs. Debido a las restricciones de ancho de banda que en ocasiones estos dispositivos pueden presentar, se considera necesario el desarrollo de técnicas de transmisión eficientes para el envío de los vectores de parámetros acústicos desde el *Front-End* hacia el *Back-End* del sistema de reconocimiento automático del habla. En este trabajo se presenta un estudio para mejorar la eficiencia en la transmisión de parámetros acústicos basado en el empleo de técnicas de cuantificación vectorial diferencial (DVQ) con el objetivo de reducir al máximo el ancho de banda empleado sin deterioro de las prestaciones del sistema de reconocimiento automático del habla en términos de WER. Se han alcanzado tasas de error comparables a las que se obtendrían sin realizar ningún tipo de compresión con velocidades de transmisión tan bajas como 2.1 Kbps.

1. INTRODUCCIÓN

Los sistemas de reconocimiento automático del habla distribuido se basan en una estructura cliente-servidor en la que uno de los extremos de la comunicación, generalmente el cliente, extrae y envía vectores de características acústicas al otro extremo, generalmente el servidor, que realiza la decodificación acústica. El primero de los extremos recibe el nombre de *Front-End*, mientras que el segundo se denomina *Back-End*. Este reparto de las tareas permite llevar a cabo el desarrollo de aplicaciones con interfaces orales para dispositivos portátiles con baja capacidad. En ocasiones, este tipo de dispositivos cuentan con un reducido ancho de banda, por lo que la transmisión deberá ser realizada del modo más eficiente posible. Además, un servidor debe dar servicio a un elevado número de clientes, siendo conveniente la minimización del ancho de banda utilizado por cada uno de ellos.

Este trabajo ha sido parcialmente financiado a través del proyecto TIN2005-08660-C04.

En este trabajo se presenta una técnica de codificación de parámetros acústicos para su posterior transmisión al *Back-End.* Dicha técnica hace uso de una cuantificación vectorial diferencial (DVQ) para conseguir la mayor compresión posible sin degradar las prestaciones del sistema de reconocimiento automático del habla en términos de su tasa de error.

El presente artículo está organizado del siguiente modo: En la Sección 2 se realiza un breve repaso a las técnicas de cuantificación vectorial. La Sección 3 se dedica a la presentación del bloque extractor de parámetros acústicos y la descripción de las técnicas de compresión de los mismos utilizadas. En la Sección 4 se ofrece un estudio de las prestaciones de las mismas y por último, la Sección 5 presenta las conclusiones.

2. CUANTIFICACIÓN VECTORIAL

La cuantificación vectorial es una técnica de codificación de fuente empleada para representar de un modo compacto un conjunto de valores y tiene sus bases en la Teoría de la Información. Ésta demuestra que siempre que la información mutua entre las diferentes componentes sea no nula, el uso de cuantificación vectorial conjunta conseguirá una representación más compacta que la cuantificación escalar de cada componente por separado. No fue hasta la década de los años 80, cuando su realización práctica se hizo posible a través del trabajo de Linde, Buzo y Gray [1]. Dicho método surge como generalización del algoritmo de Lloyd [2] que a su vez puede verse como una solución heurística del problema de *k-means* [3].

El algoritmo *k-means* describe el procedimiento para realizar la agrupación (*clustering*) de un conjunto dado de elementos en k clases. Así, dicho algoritmo constituye un método para obtener un *codebook* (diccionario) formado por *codewords* (palabras), de manera que en el proceso de cuantificación se represente cada uno de los vectores de entrada con el *codeword* más próximo en el sentido de mínima distorsión. Es el índice de dicho *codeword*, lo que es enviado al decodificador para posteriormente reconstruir el vector de entrada. Una de las medidas de distorsión más comúnmente utilizadas es la distancia euclídea.

$$d(\mathbf{x},\tilde{\mathbf{x}}) = (\mathbf{x}-\tilde{\mathbf{x}})^T(\mathbf{x}-\tilde{\mathbf{x}}) \quad (1)$$

El algoritmo *k*-means guarda ciertas similitudes con el algoritmo *Expectation-Maximization (EM)* para mezclas de gaussianas [4] bajo ciertas condiciones: a) Las componentes de la mezcla poseen matrices de covarianza diagonales con elementos unitarios con la distancia euclídea como medida de distorsión. b) Los pesos de las componentes son todos iguales. c) Mientras que el algoritmo *k-means* realiza asignaciones "*hard"* (deterministas) de los elementos a los *clusters,* el algoritmo *EM* realiza un cálculo de las probabilidades de pertenencia a cada *cluster*. Sin embargo, puede modificarse el algoritmo *EM,* realizando asignaciones "*hard"* de elementos a *clusters* para llegar a la solución *k-means* [5]

3. COMPRESIÓN DE PARÁMETROS ACÚSTICOS.

Con el objetivo de dotar de la máxima eficiencia al uso del ancho de banda requerido, se ha llevado a cabo el estudio de un conjunto de técnicas de compresión de parámetros acústicos para el reconocimiento automático del habla distribuido. Partiendo del Front-End estandarizado por ETSI ES 201 108 V1.1.2 [6] se han estudiado diferentes modos de compresión de parámetros basados en la modificación del bloque de cuantificación vectorial en él previsto.

3.1. *Front-End* para RAH distribuido ETSI ES 201 108 V1.1.2.

El estándar ETSI ES 201 108 V1.1.2 presenta un conjunto de algoritmos para la extracción de características acústicas y su posterior transmisión para sistemas distribuidos de reconocimiento automático del habla. El algoritmo de extracción de características ofrece a su salida vectores de parámetros consistentes en 13 coeficientes cepstrales junto con el coeficiente de log-energía cada 10 ms. Asimismo, define un algoritmo de compresión para reducir la tasa de transmisión. Esta compresión está basada en la cuantificación vectorial de dichos vectores tomados por parejas, dando lugar así a 7 valores cuantificados (el coeficiente C0 se cuantifica junto con el coeficiente de log-energía y el resto tomando parejas de forma correlativa).

3.2. Cuantificación vectorial diferencial (DVQ) de parámetros acústicos.

En primer lugar y para comprobar que es posible la compresión de la información de salida del algoritmo de extracción de parámetros acústicos definido en el estándar, se llevó a cabo el estudio de la Información mutua de cada una de las parejas que define dicho estándar. En la Figura 1 se muestra la Información mutua estimada haciendo uso de un subconjunto de la base de datos Albayzin [7] para las parejas de coeficientes que van desde el 1 hasta el 12. En ella puede observarse cómo los coeficientes cepstrales tomados de dos en dos presentan una información mutua no nula que permite su compresión a través de técnicas de cuantificación vectorial.

Seguidamente, para evaluar la eficiencia de los cuantificadores vectoriales propuestos por el estándar se propuso la realización de un conjunto de *codebooks* nuevo para cada una de las parejas que toma el estándar a través de la aplicación del algoritmo *k-means*. Para la obtención de estos *codebooks* se hizo uso de un subconjunto de la base de datos Speech-Dat Car en español [8]. Con el objetivo de encontrar la longitud más apropiada para dichos *codebooks* se realizó un barrido con 8, 16, 32 y 64 *codewords*.

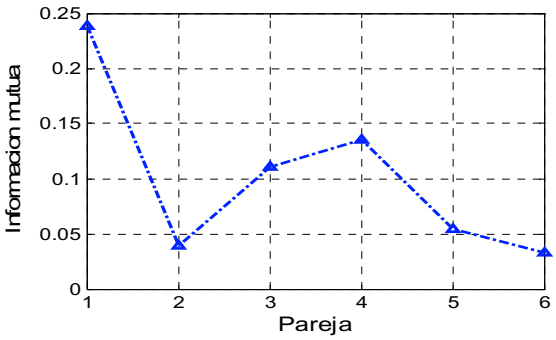

Figura 1. *Información mutua estimada para las parejas de coeficientes cepstrales.*

A continuación, se realizó una modificación sobre la estructura anterior para hacer más específicos los *codebooks* al cuantificar de manera distinta los sonidos de alta energía y los de baja energía. Para ello se optó por el uso de un umbral sobre el parámetro de energía que realizase una preclasificación de las tramas. El conjunto de datos de entrenamiento usado fue el anteriormente mencionado.

Por último, se llevó a cabo la cuantificación vectorial diferencial (DVQ), que utiliza codificación DPCM (Differential Pulse Code Modulation) empleada usualmente en compresión de audio y video digital, realizando la etapa de cuantificación de forma vectorial. Esquemas similares en cuantificación de coeficientes MFCC se habían propuesto anteriormente, como realizar predicción lineal para posteriormente hacer una cuantificación vectorial en dos etapas [9]. La diferencia fundamental de este sistema frente a DVQ reside en que a diferencia de lo que sucede en ésta, el error de cuantificación se va acumulando para tramas sucesivas lo que puede degradar las prestaciones del sistema en el caso de realizar predicción lineal adaptativa. En ese caso los coeficientes del filtro predictor no se podrán calcular con la aproximación 'Backward' en el decodificador al no disponer de las mismas señales que en transmisión y deberán ser enviados (aproximación 'Forward') con el consiguiente incremento del ancho de banda.

En la Figura 2 se muestra un esquema del cuantificador vectorial diferencial empleado, que realiza la predicción lineal de forma individual sobre cada uno de los coeficientes, mientras que la cuantificación es vectorial, con las parejas de coeficientes definidas en el

estándar ETSI. Cada pareja de coeficientes MFCC se denotada por la dupla $\mathbf{x} = (x_a, x_b)$. De esta pareja se sustraen sendas predicciones $\hat{\mathbf{x}} = (\hat{x}_a, \hat{x}_b)$ realizadas a partir de los valores cuantificados de la trama anterior, obteniéndose así la pareja de errores de predicción

$$\mathbf{d} = (d_a, d_b) = \mathbf{x} - \hat{\mathbf{x}} \quad (2)$$

posteriormente, estos errores se cuantifican dando lugar a $\tilde{\mathbf{d}} = (\tilde{d}_a, \tilde{d}_b)$, el error de predicción cuantificado

$$\tilde{\mathbf{d}} = \mathbf{d} + \mathbf{e}_q \quad (3)$$

donde $\mathbf{e}_q = (e_{q_a}, e_{q_b})$ es el error de cuantificación.

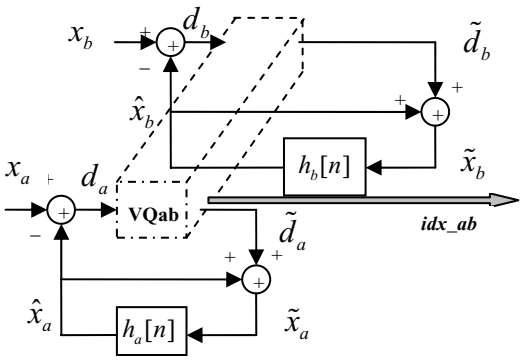

Figura 2. *Esquema del cuantificador vectorial diferencial para una pareja de coeficientes MFCC.*

Los errores de predicción de (3) serán usados para obtener la pareja de coeficientes cuantificados a través de los cuales se obtendrá la predicción de los parámetros de la trama siguiente. Los filtros de predicción lineal aparecen denotados como $h_a[n]$ y $h_b[n]$ aunque para este primer estudio se han sustituido por simples elementos de retardo de manera que el valor predicho de cada coeficiente es directamente su valor cuantificado en la trama anterior.

Uno de los principales efectos sobre los coeficientes que tiene la aplicación de técnicas de codificación diferenciales es la reducción de su varianza. Esto se ilustra en la Figura 3 dónde se representan el histograma de los coeficientes C1 y C2 (izquierda) junto con el histograma del error de predicción de los coeficientes C1 y C2 (derecha) estimados a partir de un subconjunto de la base de datos Albayzin.

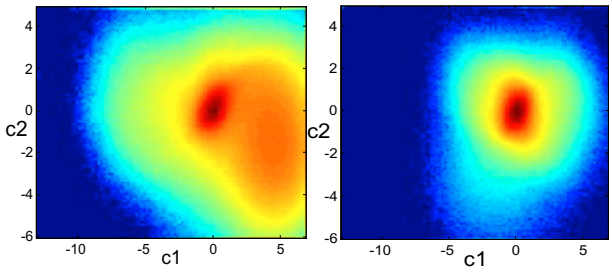

Figura 3. *Histogramas de la primera pareja de coeficientes MFCC (izquierda) y de los errores de predicción de la primera pareja (derecha).*

Esta reducción de varianza permite que con el mismo número de *codewords* se obtenga una menor distorsión media al poder tener los centroides de las mismas más próximos entre si, cubriendo la mima proporción de valores de la señal a representar.

4. EVALUACIÓN DE PRESTACIONES.

Para evaluar las prestaciones de los distintos esquemas de compresión presentados, se llevaron a cabo un conjunto de experimentos en los cuales se valoró tanto el error de cuantificación como las prestaciones del sistema completo, en términos de la tasa de error obtenida por un sistema de RAH determinado para una tarea concreta.

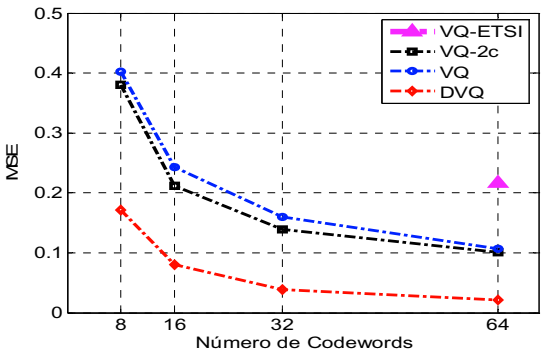

Figura 4. *Error cuadrático medio de las distintas aproximaciones de compresión en función del número de palabras del codebook .*

En cuanto a la distorsión introducida por la cuantificación, que presentan los distintos esquemas, en la Figura 4 se muestra la evolución del error cuadrático medio (MSE) para el conjunto de los 14 parámetros en función de la longitud del *codebook*. Como puede observarse, la aproximación que menor error cuadrático medio presenta es la de la cuantificación vectorial diferencial, mejor incluso que la cuantificación vectorial a partir de la definición de dos clases en función de la energía que a su vez presenta menores errores en general que un *codebook* no específico. Junto a estas representaciones se presenta el error cuadrático medio obtenido con el *codebook* propuesto en el estándar, con una distorsión superior a todas las aproximaciones presentadas, para el mismo número de *codewords*. Los valores han sido obtenidos haciendo uso de la base de datos Albayzín.

Por otro lado, se realizaron un conjunto de experimentos de reconocimiento para comprobar la validez de los esquemas de compresión propuestos. El B*ack-end* empleado hace uso de modelos continuos de palabra con 3 estados cada uno y modelos de mezclas de gaussianas como probabilidades de observación de 16 componentes cada una. La tarea tomada para la evaluación es el reconocimiento de dígitos continuos y aislados en castellano pertenecientes a la base de datos Speech-dat Car. La evaluación se llevó a cabo tanto con señal limpia (tomada de un micrófono de cercanía o Close-Talk) como con señal ruidosa (tomada de un micrófono situado en el techo del vehículo).

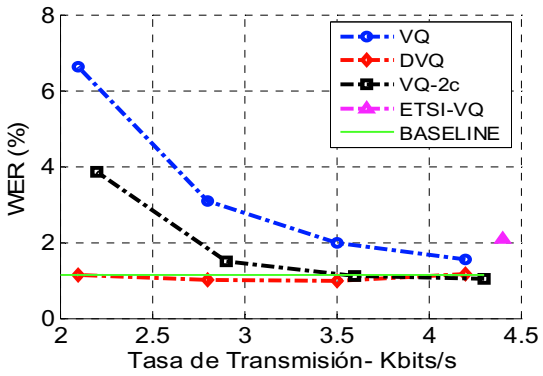

Figura 5. *Valores de Word Error Rate (WER) en función de la tasa de transmisión necesaria para los distintos esquemas de compresión y señal límpia.*

En la Figura 5 se muestra la tasa de error obtenida con señal limpia para los distintos esquemas de compresión propuestos en función del ancho de banda empleado para la transmisión junto con el error que se obtendría si no se realizase ningún tipo de compresión (*Baseline*). Como puede observarse la tasa de error desciende a medida que se incrementa el tamaño del *codebook* y por tanto se aumenta el ancho de banda necesario para la transmisión en los esquemas de cuantificación vectorial básico(VQ) y con preclasificación de la entrada en dos clases en función de la energía (VQ-2c). Sin embargo, la tasa de error se mantiene prácticamente invariante con la tasa de transmisión para el esquema propuesto (DVQ) y en valores muy similares al caso de hacer uso de ningún tipo de compresión (Baseline). Como referencia, se presenta la tasa de error obtenida con la cuantificación vectorial propuesta por el estándar (ETSI-VQ).

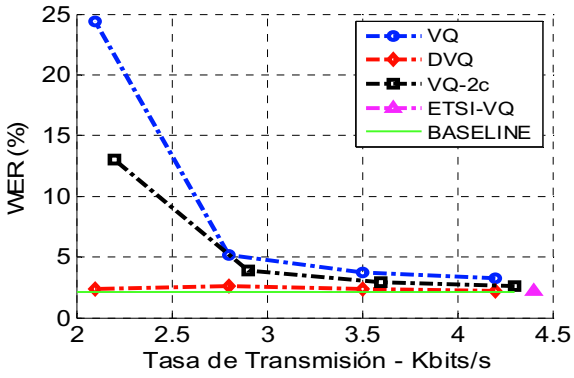

Figura 6. *Valores de Word Error Rate (WER) en función de la tasa de transmisión necesaria para los distintos esquemas de compresión y señal ruidosa.*

Por último, en la Figura 6 puede verse la evolución de la tasa de error obtenida con señal ruidosa para los distintos esquemas de compresión propuestos en función del ancho de banda empleado para la transmisión junto con el error que se obtendría si no se realizase ningún tipo de compresión (*Baseline*). En ella se aprecia, al igual que en la figura anterior, el descenso en la tasa de error cuando se aumenta la tasa de transmisión empleada para VQ y VQ-2c, hasta llegar a valores comparables con los obtenidos con el éstandar (ETSI-VQ). Sin embargo, la aproximación diferencia (DVQ) mantiene valores comparables al *Baseline* incluso con tamaños muy pequeños del *codebook*, es decir con tan sólo 8 *codewords,* 2.1 kbps, la tasa de error obtenida ya se encuentra en valores comparables a los obtenidos con el estándar con 4.4 kbps.

5. CONCLUSIONES

En el presente trabajo se ha presentado un estudio para mejorar la eficiencia en la transmisión de parámetros acústicos basado en el empleo de técnicas de cuantificación vectorial diferencial (DVQ). Dicha aumento de eficiencia tiene por objetivo la reducción del ancho de banda empleado en la transmisión de vectores de características acústicas en sistemas de reconocimiento automático del habla distribuido, sin degradar las prestaciones del sistema en términos de WER. Se han alcanzado tasas de error comparables a las que se obtendrían sin realizar ningún tipo de compresión con velocidades de transmisión tan bajas como 2.1 Kbps, lo que indica que la técnica propuesta puede ser apropiada para ser incluida en determinadas aplicaciones con interfaces orales sobre dispositivos móviles con restricciones de ancho de banda.

6. BIBLIOGRAFÍA

[1] Linde, Y., Buzo, A., Gray, R., "An Algorithm for Vector Quantizer Design", IEEE Trans on Comm., v. 28, (1980).

[2] Stuart P. Lloyd. "Least Squares Quantization in PCM." IEEE Trans. on Inf. Theory, vol. 28(2), pp. 129-137, 1982.

[3] Huang, X., Acero, A., Hon, H., "Spoken Language Processing: a guide to theory, algorithm, and system development" (2001) Prentice Hall.

[4] Dempster, A. P., Laird, N. M., and Rubin, D. B., "Maximum likelihood from incomplete data via the EM algorithm," J. R. Statist. Soc., vol. 39, (1) , pp. 1–21, 1977

[5] Qiu, D. and Ajit Tamhane, C., "A comparative study of the K-means algorithm and the normal mixture model for clustering: Univariate case" Journal of Statistical Planning and Inference. vol. 137 (11), pp. 3722-3740, 2007.

[6] ETSI standard doc. "Speech Processing, Transmission and Quality Aspects (STQ); Distributed speech recognition; Front-end Feature extraction algorithm; Compression algorithms", ETSI ES 201 108 Ver.1.1.2 (2000-04).

[7] Casacuberta, F., García, R., Llisterri, J.,Nadeu, C., Pardo, J.M., Rubio, A. "Development of Spanish Corpora for Speech Research (Albayzin)",Workshop on Standardization of Speech Databases and Speech Assessment Methods. (1991).

[8] Moreno, A., Lindberg, B., Draxler, C., Richard, G., Choukri, K., Euler, S., and Allen, J., "Speechdat-car: A large speech database for automotive environments", LREC (2000).

[9] Ramaswamy, G. N. and Gopalakrishnan, P. S., "Compression of Acoustic Features for Speech Recognition in Network Environments" in Proc of ICASSP, Seattle. 1998.