SPEAKER RECOGNITION EMPLOYING WAVEFORM BASED SIGNAL REPRESENTATION IN NONORTHOGONAL MULTIPLE TRANSFORM DOMAINS

Wasfy. B. Mikhael and Pravinkumar Premakanthan
Department of Electrical Engineering, University of Central Florida, P.O. Box 162450
Orlando, FL 32816-2450, Email: mikhael@mail.ucf.edu, Phone: 407-823-3210,Fax: 407-823-5835

ABSTRACT

Automatic Speaker Recognition (ASR) technique employing Split Vector Quantized speech representation in multiple transform domains is presented. In this approach, a set of appropriate transform domains are selected and a vector quantized codebook is generated in each of these selected transform domains for the signal waveform. For each speaker, each signal vector is represented from the codebooks that yield the highest accuracy of representation. The algorithm is given and a performance measure is developed and used to evaluate the algorithm performance. Improved speech recognition accuracy was consistently obtained employing the proposed technique in comparison with vector quantization employing single transform VQ representations. Sample results for 10 speakers are presented to illustrate the considerable performance improvement for ASR.

I INTRODUCTION

Automatic Speaker Recognition (ASR) has always been a challenging problem to the researchers in this area. The problem, depending on the nature of the final task can be classified into two different subdivisions namely Speaker Identification (SI) and Speaker Verification (SV). Determining a speaker from a group of speakers is SI, while in the SV process, the existence of the speaker in the database is confirmed. Thus, the fundamental difference between the SI and the SV processes is the number of the decision alternatives. Vector Quantization (VQ) has been used as an efficient means of characterizing the short time spectral features of a speaker [5]. Vector Quantized codebooks are used as templates to store the acoustical features of the speaker for template based ASR. But, as the number of the training vectors becomes larger, the memory requirements for storage and the computational complexity during the recognition stage increases. A recently reported efficient and accurate design of VQ codebooks using multiple transform domains [3] is developed for ASR. The resulting SI/SV algorithm employing this approach is presented. The algorithm thus obtained yields higher recognition accuracy compared with SI/SV that employs VQ in a single domain. The remainder of this paper is organized as follows. Section II describes the multiple transform based Split Vector Quantization briefly. The proposed Waveform Based SI/SV using Redundant Signal Representation (WCSVRSR) algorithm is developed and presented in Sec II briefly. Sample experimental results are given in Sec IV. Finally, Conclusions are contained in Sec V.

II MULTIPLE TRANSFORM BASED SPLIT VECTOR QUANTIZATION

Mixed transform techniques represent signals using combinations of basis functions chosen from two or more transform domains simultaneously, to achieve higher energy compaction than can be achieved using a single transform. For a given number of selected basis functions, the signals represented in multiple transform domains have been shown to yield better signal reconstruction quality than the single transform domain approaches. [2]. In this approach, initially, a VQ codebook is obtained for each signal in each of the transform domains. The signal is then approximated by the speech vectors that are best represented among the different transform domains to improve the performance measure for ASR.

Appropriate linear transform domain representation compact the signal information in fewer coefficients than time/space domain representation. This implies that the distribution of energy among the various transform coefficients is highly skewed and few transform coefficients represent most of the signal energy. This fact is exploited in split vector quantization [6], also referred to as partitioned vector quantization, where the transform coefficients of the signal vector are partitioned into sub vectors. Each sub vector is separately represented. This partitioning enables processing of vectors with higher dimensions in contrast with time/space direct vector quantization. Various techniques have been used to split the transformed vector into sub vectors for Vector quantization. In this paper, Energy based split vector quantization in combination with the multiple transform domain signal representation has been adopted for ASR. As will be shown

0-7803-7448-7/02/$17.00 ©2002 IEEE

later, this technique yields considerable improvement in performance measure as compared to ASR employing single domain representation techniques.

III WAVEFORM BASED SPEAKER RECOGNITION USING MULTIPLE TRANSFORM SIGNAL REPRESENTATION (WSRMT)

In this section, the WSRMT algorithm is proposed. The algorithm consists of two stages, namely, the training stage and the testing stage. In the training stage, initially, all silence parts of the speech signal X (n) is removed and the signal amplitude is normalized over the entire length. The speech feature vectors are formed from consecutive samples of speech X (n) by windowing. Overlapping trapezoidal windows are used for splitting the speech into frames with some percentage of overlap between them. Each frame is then multiplied by a hamming window. The main purpose of windowing is for energy normalization and to avoid abrupt discontinuities [4]. Thus, each of the Q speech signal $X_q(n)$ q=1,2,..Q is divided into D vectors or frames $X_q(n) = \{X_{q1}, X_{q2}, \ldots\ldots X_{qD}\}$. Each vector $X_i = \{x_i(1), x_i(2), \ldots\ldots x_i(m)\}$ of length m is transformed simultaneously into P non-orthogonal linear transform domains. The feature vector X_i representation in the j^{th} domain is given by $\varphi_i^j = \{\varphi_i^j(1), \varphi_i^j(2), \ldots\ldots \varphi^j{}_i(m)\}$. The vectors φ_i^j, are then split into L sub bands, $\varphi_i^j = \{\varphi_{i1}^j, \varphi_{i2}^j, \ldots\ldots \varphi^j{}_{iL}\}$ in such a way that the average energy in each of the L sub bands are equal. The sub bands, generally, are of different lengths, each containing approximately 1/L of the total normalized average signal energy. In the j^{th} transform domain, the l^{th} sub vector is denoted by φ_{il}^j where j=1,2...P, l=1,2.... L. The training sub vectors corresponding to φ_{il}^j are clustered using k-means clustering algorithm [4] and the codebook for the l^{th} sub vector in the jth domain is designed. The vector-quantized representation of φ_{il}^j is designated by $\hat{\varphi}_{il}^j$. Since the energy content in the sub band is nearly the same, equal number if bits is allocated to each sub band. The vector quantized sub codebooks are concatenated to form the main codebooks $\{C_i^1, C^2{}_i \ldots\ldots C^P{}_i\}$ where, i =1,2,..Q corresponds to Q speakers in each of the P domains. The process of designing the codebooks is made efficient and accurate by employing the Adaptive Codebook Accuracy Enhancement (ACAE) algorithm. [3] Using this method, the accuracy of the codebooks designed in a given domain is improved by redesigning the represented code words using the training vectors that were better represented in that domain. This method of codebook design is repeated for all the speakers who are to be enrolled in the database. Thus, each speaker has P codebooks that correspond to P domains. In the testing phase, there exists an unknown speaker whose set of feature vectors has to be identified, i.e. a match has to be searched for a match in the database. The unknown set of feature vectors $\{X_{un1}, X_{un2}, \ldots\ldots X_{unD}\}$, which corresponds to the unknown speaker, is run through each of the P codebooks for each speaker enrolled in the database. The approximate vectors corresponding to the unknown speaker, using each of the stored codebooks, are obtained. Thus, the approximate vectors obtained in the j^{th} domain is represented by $\{X_{apx_1}{}^j, X_{apx_2}{}^j, \ldots\ldots X_{apxD}{}^j\}$, where D is the total number of feature vectors or the total number of frames. The speech waveform is then reconstructed in each of the P domains. Assume $\{X_{1t}, X_{2t}, \ldots\ldots X_{Dt}\}$ are the feature vectors of the t^{th} speaker and has the corresponding codebooks $\{C^j\}$ where, j=1,2...P domains. If $\{X^j{}_{R1}, X^j{}_{R2}, \ldots\ldots X^j{}_{RD}\}$ and $\{X^h{}_{R1}, X^h{}_{R2}, \ldots\ldots X^h{}_{RD}\}$ are the reconstructed speech vectors of the unknown speaker in domains j and k, respectively, using the generated codebooks, then the best approximated vector is obtained such that the encoder chooses $\{X^j{}_{R1}\}$ if $\| X_i - X^j{}_{Ri} \| < \| X_i - X^k{}_{Ri} \|$ for i=1,2...D. In the above expression. The symbol $\| \ \|$ represents the Euclidean distance measure [4]. The Signal to Noise ratio (SNR) between the unknown speaker's original waveform $X_{unk}(n)$ and its best-approximated waveform using each of the Q speaker's codebooks $\hat{X}_{unk,q}(n)$ is calculated by

$$SNR_{unk,q} = 10\log_{10}\left[\frac{\sum X_{unk}(n)^2}{\sum \left(X_{unk}(n) - \hat{X}_{unk,q}(n)\right)^2}\right] \quad (1)$$

Where q=1,2...Q, denotes the total number of speakers in the database. Thus, the $SNR_{unk,q}$ is calculated between the unknown speaker and each of the Q speakers enrolled in the database. The decision is finally made that the unknown speaker represents the k^{th} if max($SNR_{unk,k}$)> max (($SNR_{unk,q}$) where q≠k and q=1,2..Q.

IV EXPERIMENTAL RESULTS

The WSRMT algorithm for the ASR has been experimentally verified. A sample experiment using a 10-speaker database has been given here to illustrate the performance of the proposed technique. Each of the 10 speakers has two different files $\{[X_{1,1}(n), X_{1,2}(n)], [X_{2,1}(n), X_{2,2}(n)],....[X_{Q,2}(n),X_{Q,2}(n)]\}$, of the same utterance, recorded at different instances of time and each sampled at 8000Hz. In the training phase, speech vectors of length 32 samples each with 6-sample overlap are obtained from 25500 successive samples of speech. The obtained vectors are simultaneously projected in two transform domains namely the Discrete Cosine Transform (DCT) domain and the HAAR domain. Each vector is then divided into 5 sub bands of equal energy. Then Energy based Split vector quantization is employed to obtain vector quantized codebooks for each speaker in both the domains. The average number of bits per sample is calculated as the ratio of the total number of bits used to represent the concatenation of the sub band section code words to the length of the vector. Thus, at the end of the training phase, there are two codebooks for each of the ten speakers. During the recognition stage of the simulation, an unknown speaker (known to be speaker 1) is presented to the system for recognition. The signal to Noise ratios between each of the 10 speakers using the 1st speaker's codebook, $SNR_{unk,1}$ are calculated. The experiment is also repeated by using the second speech file of speaker1 (Sp1). Fig.1 reveals the improvement in SNR when the WSRMT algorithm was used when the Sp1 is tested with same file that used to form the codebooks during the training process. Fig.2, Fig.3 and Fig.4, respectively, show the SNR curves in DCT, HAAR and the best of both domains when different files were used for each speaker during the testing phase. The plots reveal that the algorithm recognizes that the unknown speaker corresponds to speaker 1. The plots are with respect to increasing bit rates. The SNR is found to improve when the best of both domain representations is used for recognition. Two main observations could be made from the results obtained. Firstly, the discrimination between the true speaker and rest of the imposters is found to increase with increasing bit rates. Secondly, the bit rate used for this method of SI/SV process is still less than 1.8 bits/sample thus, reducing the size of storage of speech templates.

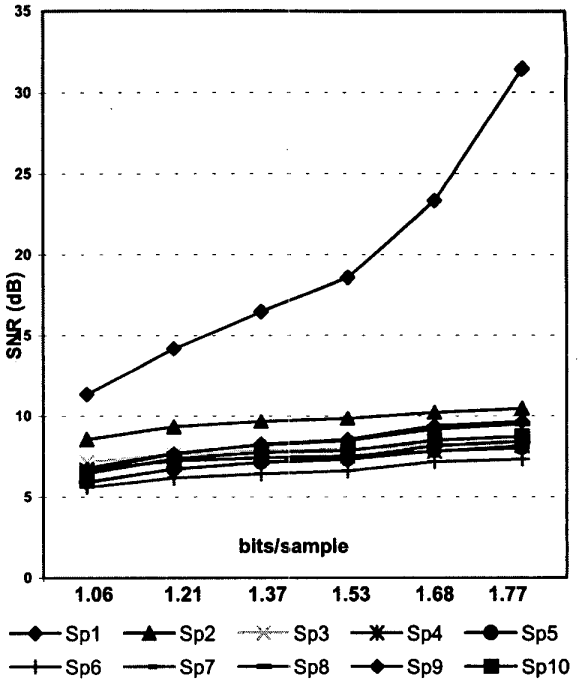

Fig 1. SNR Improvement with Sp1- same file with Sp1 Codebook (Best of Both Domains)

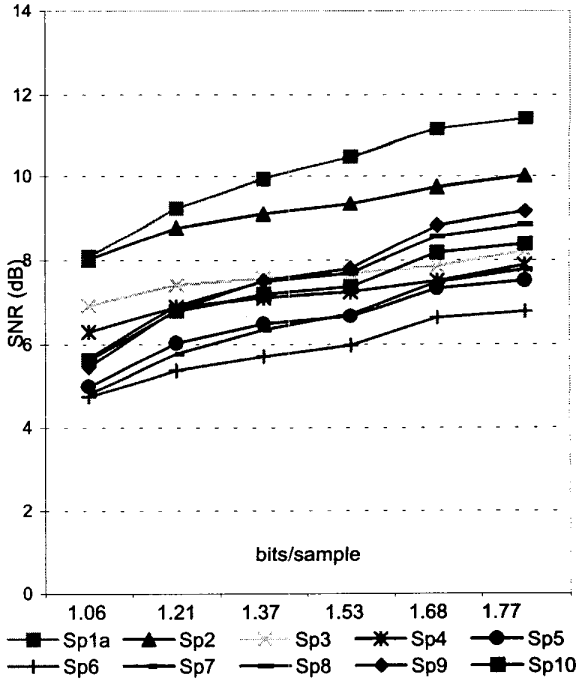

Fig 2. SNR Improvement with Bitrates -DCT Domain Sp1-Different File with Sp1 Codebook

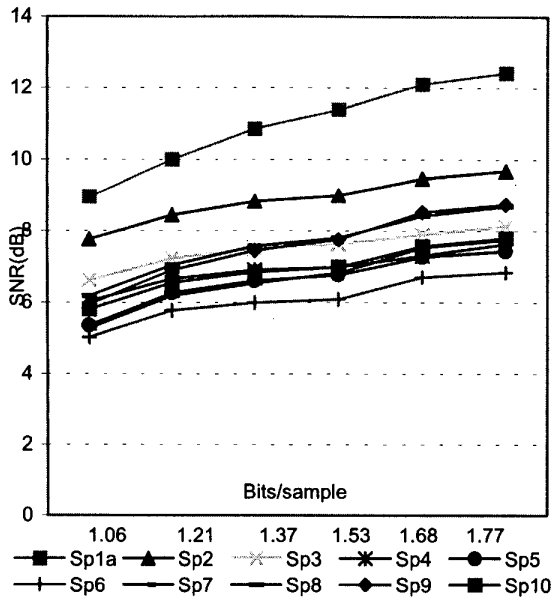

Fig 3. SNR Improvement with Bit Rates-HAAR Domain Sp1-Different File with Sp1 Codebook

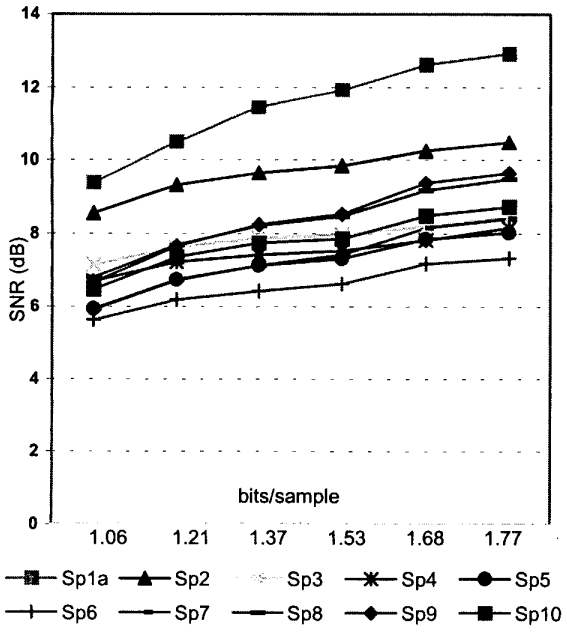

Fig 4. SNR Improvement with Bit rates (Best of Both Domains) Sp1-Different File with Sp1 Codebook

V CONCLUSIONS

In this paper, an approach for Speaker Recognition employing a recently proposed VQ in Multiple transform domains for waveform based signal characterization has been developed and presented. Sample simulation results are given which demonstrate the improvement in the signal recognition performance. The results clearly indicate that the difference between the SNR curves of the true speaker and rest of the imposters is found to increase with increasing bit rates. The performance of the proposed WSRMT is found to be superior technique (SNR is found to increase by 2.5 dB) to the existing VQ employing single transform VQ approaches at the expense of moderately increase in the computational complexity. Work is in progress to use multi-criteria approach to ASR by grouping the speakers according to certain selected criteria to reduce the searching time during the recognition process.

REFERENCES

[1] W.B.Mikhael and V.Krishnan, "Multiple transform domain split vector quantization", Electronics Letters, Volume: 37, Issue: 8, April 2001,pp: 538 –539.

[2] A.P.Berg and W.B.Mikhael,"A survey of mixed transform techniques for speech and image coding", IEEE, International Symposium on Circuits and Systems, Orlando, June 1999,pp: 106-109.

[3] W.B. Mikhael and Venkatesh Krishnan, "A Novel Adaptive Algorithm Applied to a Class of Redundant Representation Vector Quantizers for Waveform and Model Based Coding", Submitted, IEEE International Symposium on Circuits and Systems, Arizona, 2002.

[4] "Automatic Speech and Speaker Recognition, Advanced Topics", Edited by Chin-Hui Lee, Frank K. Soong, Kuldip K.Paliwal, Kluwer Academic Publishers, June 1996.

[5] F.K.Soong, A.E.Rosenberg, L.R.Rabiner and B.H.Juang, "A Vector Quantization Approach to Speaker Recognition", IEEE International Conference on Acoustics, Speech and Signal Processing, March 1985,pp: 387-390, vol.1.

[6] "Vector Quantization and Signal Compression", Allen Gersho & Robert M. Gray, Kluwer Academic Publishers, 1992.