$W=\text{diag}\{w_1,w_2,\ldots,w_i,\ldots\}$ and each element w_i denotes the weight of each log-spectra feature. The superscript c denotes the cepstral domain. The whole procedure can be interpreted as follows: the cepstral difference vector is first reverted to the log-spectral domain, where it is weighted and then transformed back to the cepstral domain.

MMSE-LA scheme description: First, MMSE-based speech enhancement is adopted in the front-end stage to suppress the corrupted additive noise. Meanwhile, weight factors of log-spectra features are calculated and voice activity detection (VAD) results are reserved. Secondly, MFCC features are extracted from those enhanced speech. Thirdly, the residual noise is modelled with a single Gaussian mixture state. As only the mean of the model is required, we estimate it by calculating the mean of all MFCC feature vectors taken from voiceless speech segments. Fourthly, a residual noise compensated speech model is acquired using the LA algorithm [3]. Finally, the compensated speech model, MFCC features of enhanced speech and weight factors of log-spectra features enter a Viterbi decoder adapted for the FW algorithm and recognition results are acquired.

Evaluations and conclusion: Speaker-independent TI-digits recognition experiments were carried out with an FW adapted Viterbi recogniser to evaluate the MMSE-FW-LA scheme. The contaminated speech for test was generated by artificially adding different levels of noise to the clean speech. All noise signals came from a Noisex-92 database. The model we used is continuous density HMM (CDHMM) with left-to-right structure. 500 connected digits utterances from 15 speakers and 100 connected digits utterances from four speakers unseen in the training set are used for training and testing, respectively. The features are 13-dimension static MFCC features with their delta parameters. Only static features are weighted.

Table 1: Recognition accuracy [%] of baseline, using MMSE, FW, LA separately in white noise

	−5 dB	0 dB	5 dB	10 dB	15 dB
Baseline	6	8	13	30	65.33
MMSE	14.67	24	54	80.33	91
FW	24.33	46.33	65	76.67	82
LA	22.33	34.33	67.67	84.67	92.67

Table 1 shows the results of the baseline recogniser and those using FW, MMSE and LA in noisy speech recognition. It can be seen that all these approaches improve recognition accuracy, while at low SNRs (<10 dB), the FW algorithm improves the recognition accuracy more markedly. Table 2 shows the recognition performance combining the FW algorithm with the front-end MMSE-based speech enhancement and LA model compensation, respectively. Referring to Table 1, it is distinct that the MMSE-FW and FW-LA schemes outperform either of the three algorithms used alone, and the lower the SNR, the more remarkable the improvement. Table 3 compares the MMSE-FW-LA scheme with the MMSE-LA scheme proposed in [3]. Apparently, at very low SNRs (<5 dB), the MMSE-FW-LA scheme is superior to the MMSE-LA scheme, and maintains high recognition accuracy over 80% at −5 dB SNR.

Table 2: Recognition accuracy [%] of MMSE-FW, FW-LA in white noise

	−5 dB	0 dB	5 dB	10 dB	15 dB
MMSE-FW	46.33	76	85	91.67	92.67
FW-LA	32	57.33	77.67	87.33	94.67
MMSE-LA	65.33	79.67	89.33	93	93.33

Table 3: Recognition accuracy [%] of MMSE-FW, FW-LA in white noise

	−5 dB	0 dB	5 dB	10 dB	15 dB
MMSE-LA	65.33	79.67	89.33	93	93.33
MMSE-FW-LA	81	86	89	94.33	94

Acknowledgments: This work is supported by the National Natural Science Foundation of China (No. 60072011) and the Foundation of Information School of Tsinghua University.

12 May 2002
Electronics Letters Online No: 20020940
DOI: 10.1049/el:20020940

Tao Xu and Zhigang Cao (*The State Key Laboratory on Microwave and Digital Communications, Department of Electronic Engineering, Tsinghua University, Beijing 100084, People's Republic of China*)

E-mail: xutao@sat.mdc.tsinghua.edu.cn

References

1 COOKE, M., GREEN, P., JOSIFOVSKI, L., and VIZINHO, A.: 'Robust automatic speech recognition with missing and unreliable acoustic data', *Speech Commun.*, 2001, **34**, (3), pp. 267–285
2 BARKER, J., JOSIFOVSKI, L., COOKE, M., and GREEN, P.: 'Soft decision in missing data techniques for robust automatic speech recognition'. Proc, ICSLP, Sydney, Australia, 2000, pp. 373–376
3 DING, P., and CAO, Z.G.: 'Combining MMSE enhancement with LA model adaptation for robust automatic speech recognition', *Electron. Lett.*, 2001, **37**, (8), pp. 539–540
4 EPHRAIM, Y., and MALAH, D.: 'Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator', *IEEE Trans. Acoust. Speech Signal Process.*, 1984, **ASSP-32**, pp. 1109–1121
5 HAKKINEN, J., and HAVERINEN, H.: 'On the use of missing feature theory with cepstral features', *Consistent & Reliable Acoustic Cues CRAC*, 2001, (2)

Speaker identification employing redundant vector quantisers

W.B. Mikhael and P. Premakanthan

A novel approach for automatic speaker identification employing a recently proposed signal compression technique, namely, the multiple transform domain split vector quantiser, is developed and presented. Using a normalised matching accuracy measure, the proposed technique consistently yields enhanced identification performance in comparison with existing single domain vector quantiser approaches.

Introduction: Automatic speaker recognition (ASR) has always been a challenging task. ASR can be accomplished in two stages, namely, automatic speaker identification (ASI), and automatic speaker verification (ASV). ASI is achieved by determining a speaker from a group of speakers, while in the ASV process, by setting thresholds, the existence of the speaker in the database is confirmed.

Transform coding of speech vectors compacts the signal information into fewer coefficients [1]. For speaker modelling, vector quantisation (VQ) has been used as an efficient means of characterising the short time spectral features of a speaker [2]. Various techniques that combine the advantages of VQ and transform coding for speech signal compression have been proposed. In this Letter, a novel algorithm for ASI which employs the recently reported multiple transform domain split vector quantiser (MTSVQ) signal compression technique is developed and presented. Sample results are presented which confirm the improved accuracy of this algorithm at the expense of increased computational complexity.

Multiple transform domain split vector quantiser (MTSVQ): Multiple transform domain signal representation techniques represent signals using combinations of basis functions chosen from two or more transform domains. For most practical signals, each transform domain representation from an appropriately selected set of domains of a stationary signal segment yields different energy compaction characteristics. The domain that gives the best energy compaction is selected to represent this particular segment. For a given compression ratio, the signals represented in multiple transform domains have been shown to yield better signal reconstruction quality than the single transform domain approaches. Multiple transform representation in conjunction with VQ has been shown to exploit the improved

accuracy of the first and the enhanced coding efficiency of the latter [3].

Proposed waveform-based ASI (WASI) algorithm: The algorithm has two modes, namely, the training mode, and the running mode. Initially, the silence parts of the speech signal are removed and the signal amplitude is normalised over the signal record. During the training mode t, for the kth speaker, the speech record X_t^k is segmented into N vectors (segments), each of length S samples, using overlapping trapezoidal windows, i.e. $X_t^k\{x^1, x^2, \ldots, x^i, \ldots, x^N\}$. Each x^i, $i=1$ to N, is transformed into P transform domains. This yields, in the jth domain, $\phi_{t,j}^i$, $j=1$ to P. In the jth domain, the $\phi_{t,j}^i$s, $i=1$ to N, are split into M subbands, $\phi_{t,j}^i\{L_{j1}^i, L_{j2}^i, \ldots, L_{jM}^i\}$, such that, averaged over all i, $i=1$ to N, the energy in each of the M subbands is approximately equal. In general, the subbands are of different lengths. Thus, L_{jl}^i, $l=1$ to M, denotes the lth subband in the jth transform domain of the ith segment $\phi_{t,j}^i\{L_{j1}^i, L_{j2}^i, \ldots, L_{jM}^i\}$. In the jth transform domain, for the lth subband, each set of the subvectors L_{jl}^i, $i=1$ to N, is collected and clustered using a suitable vector-quantising algorithm [2] to yield the codebook C_{jl}^k. This is repeated for all l, $l=1$ to M. Since the energy content in each of the subbands is nearly the same, an equal number of bits is allocated to each C_{jl}^k, $l=1$ to M. The resulting composite codebook in the jth domain for the kth speaker is denoted by $C_j^k(C_{j1}^k, C_{j2}^k, \ldots, C_{jM}^k)$. Similarly, codebooks C_j^k are obtained for all j, $j=1$ to P. This process is repeated for all the speech records from the Q speakers, i.e. for $k=1$ to Q speakers, who are enrolled in the database. Thus, at the end of the training stage, the codebooks C_j^k ($j=1$ to P transforms, and $k=1$ to Q speakers) are obtained, which form the speaker database.

During the running mode r, the system is presented with the speech record X_r^u, from the uth speaker. The system is required to identify u. As described in the training mode, X_r^u, is segmented into N segments, $X_r^u\{x_r^1, x_r^2, \ldots, x_r^N\}$. Each, x_r^i, $i=1$ to N, is transformed into the jth domain and divided into M subbands to yield $\phi_{r,j}^i\{L_{j1}^i, L_{j2}^i, \ldots, L_{jM}^i\}$. $\phi_{r,j}^i$s are obtained for $i=1$ to N and $j=1$ to P. These transform vectors are mapped using the corresponding codebooks for the kth speaker in the database by employing the MTSVQ technique proposed in [3], where the domain that best represents each segment, x_r^i, is determined and used to represent that particular segment $\hat{x}_r^i$. The reconstructed signal vectors $\hat{x}_r^i$, $i=1$ to N, are concatenated to form the best-reconstructed signal waveform, $\hat{X}_r^{u\to k}$. The best representation thus obtained for X_r^u using the codebooks of the kth speaker in the database is denoted by $\hat{X}_r^{u\to k}$. The process is repeated for all k, $k=1$ to Q. Let the normalised matching accuracy measure A, between the original signal waveform of the unknown speaker X_r^u and the signal waveform reconstructed using the kth speaker codebooks in the database, $\hat{X}_r^{u\to k}$ be defined as

$$A_{u,k,B} = 10\log_{10}\left[\frac{\sum_n [X_r^u(n)]^2}{\sum_n [X_r^u(n) - \hat{X}_r^{u\to k}(n)]^2}\right] \quad (1)$$

where the subscript B denotes that A is evaluated using the proposed technique. $A_{u,k,B}$ is computed for $k=1$ to Q, and k that yields the highest $A_{u,k,B}$ is identified to be the match for u.

Experimental results: A five ($Q=5$)-speaker database is used to illustrate the performance of the proposed algorithm. Each speaker k has two different speech recordings of the same utterance X_t^k, used for the training mode, and X_r^u, used for the running (identification) mode, for $k=1$ to 5, and $u=1$ to 5. The speech record, which is sampled at 8000 Hz, consists of 25508 samples. Vector (segment) length $S=32$, and trapezoidal windows with 10-sample overlap, are used. In the training mode, the vectors are projected onto three domains ($P=3$), namely, the discrete cosine transform (DCT), the HAAR transform, and the discrete wavelet transform (db1 basis) [4]. As described earlier, the transformed vector is then divided into five subbands ($M=5$) of approximate equal energy. MTSVQ is then applied to obtain vector-quantised codebooks C_j^k, $j=1$ to 3, and $k=1$ to 5. During the running mode, the unknown speakers are presented to the system for identification and the normalised matching accuracy measure $A_{u,k,B}$, $u=1$ to 5, and $k=1$ to 5, is calculated. On average, it is found that the quantiser selects 30, 25 and 45% of vectors from the DCT, the HAAR, and the wavelet domain codebooks, respectively. Due to lack of space, representative results are given. Fig. 1 gives, for X_r^u, $u=2$, $A_{u,k,B}$, ($u=2$, and $k=1$ to 5), against the number of bits per sample (bps) used. Fig. 1 shows the proposed algorithm's matching accuracy, and consequently identification ability, when an unknown speaker record is reconstructed using each of the five unknown speaker codebooks. In Fig. 2, for X_r^u $u=1$ to 5, $A_{u,k,B}$, ($u=1$ to 5, and $k=5$), are obtained against bps used. Fig. 2 shows the ability of the new technique to identify an unknown speaker by reconstructing each of the five speaker records using a particular speaker set of codebooks. From the results in Figs. 1 and 2, the approach presented uniquely identifies the unknown speaker. Figs. 3*a* and *b* show the performance improvement of the WASI over single transform vector quantisation (VQ) representation. In Figs. 3*a* and *b*, $A_{u,k,D}$, and $A_{u,k,H}$ against bps are given. $A_{u,k,D}$, and $A_{u,k,H}$, represent the normalised matching accuracy measure when single transform DCT, or the HAAR VQ representation is used, respectively. From Fig. 3, it can be seen that the single transform approach is not successful in identifying the unknown speaker. In addition, the values of $A_{u,k,D}$, and $A_{u,k,H}$ are found to be lower than those obtained in Fig. 1.

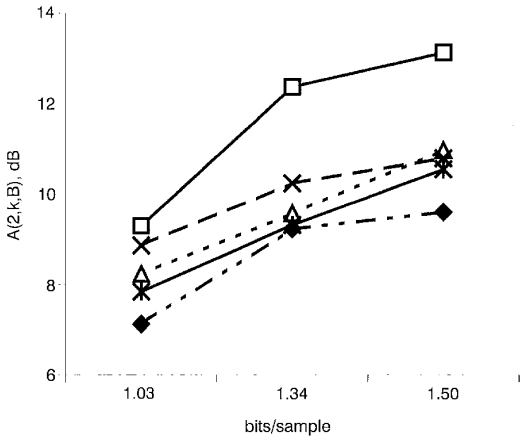

Fig. 1 *Normalised matching accuracy measure, $A_{u,k,B}$ against bps employing proposed WASI: $u=2$, $k=1$ to 5*
——◆—— A(2,1,B)—□—A(2,2,B) · · · △ · · · A(2,3,B)
—— × — —A(2,4,B)—*—A(2,5,B)

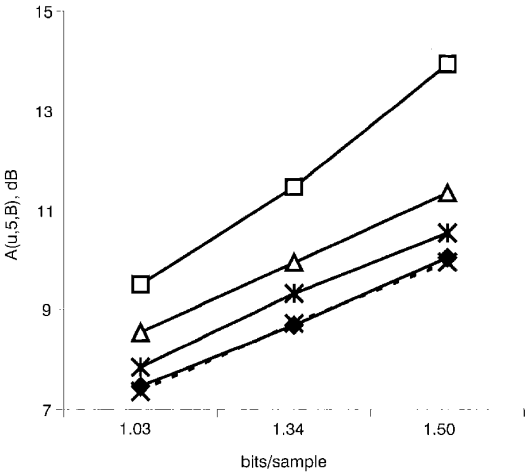

Fig. 2 *Normalised matching accuracy measure, $A_{u,k,B}$ against bps employing proposed WASI: $u=1$ to 5, $k=5$*
– – –◆– —— A(1,5,B)—*—A(2,5,B)—△—A(3,5,B)
· · · × · · · A(4,5,B) –□—A(5,5,B)

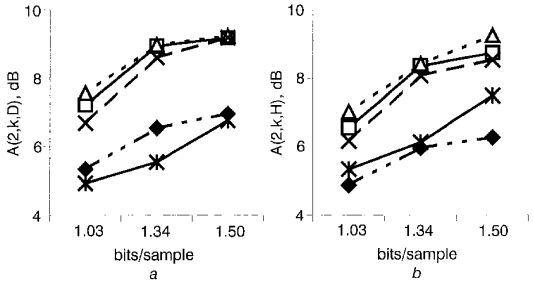

Fig. 3 *Normalised matching accuracy measure*

$A_{u,k,B}$ against bps employing single transform (DCT) VQ, $u=2$, $k=1$ to 5
—◆— A(2,1,D)—□—A(2,2,D) · · · △ · · · A(2,3,D)
— × —A(2,4,D)—*—A(2,5,D)

$bA_{u,k,H}$ against bps employing single transform (HAAR) VQ, $u=2$, $k=1$ to 5
—◆— A(2,1,H)—□—A(2,2,H) · · · △ · · · A(2,3,H)
— × — A(2,4,H)—*—A(2,5,H)

Conclusions: In this Letter, a waveform-based ASI algorithm is developed. Sample simulation results are given which successfully demonstrate the improvement in the speaker identification accuracy employing the proposed WASI technique. The algorithm differentiates between the true speaker and the imposters at low bit rates. The discrimination ability improves as the bit rates increase. The performance of the proposed WASI is found to be experimentally superior by at least 3 dB in comparison with the existing single transform domain VQ approaches. This is achieved at the expense of an increase in the computational complexity.

26 April 2002
Electronics Letters Online No: 20020948
DOI: 10.1049/el:20020948

W.B. Mikhael and P. Premakanthan (*School of Electrical Engineering and Computer Science, University of Central Florida, Orlando, FL 32816, USA*)

E-mail: mikhael@mail.ucf.edu

References

1 BERG, A.P., and MIKHAEL, W.B.: 'A survey of mixed transform techniques for speech and image coding'. IEEE, Int. Symp. on Circuits and Systems, Orlando, FL, USA, June 1999, pp. 106–109
2 GERSHO, A., and GRAY, R.M.: 'Vector quantisation and signal compression' (Kluwer Academic Publishers, 1992)
3 MIKHAEL, W.B., and KRISHNAN, V.: 'Multiple transform domain split vector quantisation', *Electron. Lett.*, 2001, **37**, pp. 538–539
4 TUFEKCI, Z., and GOWDY, J.N.: 'Feature extraction using discrete wavelet transform for speech recognition'. IEEE Proceedings, Southeastcon, Nashville, TN, USA, April 2000, pp. 116–123

Effective algorithm for multilevel converters with very low computational cost

M.M. Prats, J.M. Carrasco and L.G. Franquelo

An effective and fast modulation algorithm for high power voltage source multilevel converters is presented. This approach drastically reduces the computational load maintained, permitting the on-line computation of the switching sequence and the on-state durations of the respective switching state vectors. It has been satisfactorily implemented in very low-cost microcontrollers.

Introduction: Although the advantages of multilevel converters were known since Nabae proposed the topology neutral point clamped (NPC) inverter in 1981 [1], its implementation was limited owing to the complexity of the switching control. In recent years, multilevel voltage source inverters have been used in medium and high power applications. They present the capability of increasing the output voltage magnitude and reducing the output voltage and current harmonic content, the switching frequency and the voltage supported by each power semiconductor. Owing to these attractive characteristics, several control algorithms of multilevel converters have been recently proposed [2, 3]. However, in this Letter, an effective approach that drastically reduces the computational load using a decision-making algorithm is presented.

Description of modulation technique: This method is based on the decision-based pulse width modulation developed by Holtz [4].

Three-phase quantities are usually transformed into phasor representation since it simplifies the analysis of the modelled system. Since the switching of any power topology stays at discrete states, space vector modulation is used to approximate a reference voltage vector u^* calculating the time to its surrounding state vectors.

V_a, V_b and V_c are the three-phase quantities usually transformed into the phasor representation. Three vectors, u_1, u_2 and u_3, are used to approximate the desired voltage vector u^* in polar co-ordinates in a control cycle T_m. The modulation law requires the actual voltage vector u to equal its reference value $u^* \cdot u^*$ is represented in the stationary reference frame:

$$u = u^* = V_a + V_b e^{\frac{j2\pi}{3}} + V_c e^{\frac{j4\pi}{3}} = \mathrm{Re}\{u^*\} + j\,\mathrm{Img}\{u^*\} \quad (1)$$

During each modulation subcycle of duration T_m a switching sequence is generated. It is composed of three switching state vectors $u_1(t_1)$, $u_2(t_2)$ and $u_3(t_3)$, where t_1, t_2 and t_3 are the on-state durations of the active switching state vectors. The three vectors nearest to the reference vector must be identified. Referring to first sextant of the regular hexagon, the voltage space vector averaged over one subcycle T_m is:

$$u = (t_1 u_1 + t_2 u_2)/T_m \quad (2)$$

In this Letter, the problem is solved for the voltage vector in the first sextant. However, this reference vector can be located in any of the six sectors of the regular hexagon which contain the switching state vectors. This problem is easily solved rotating u^* anti-clockwise by an angle $(n-1)\pi/3$, where n is the sextant number, $n=1,\ldots,6$. This rotation displaces any reference vector to the first 60° to be studied there. The switching state vectors for the multilevel inverter control are determined by reverse rotation. The input to the modulation algorithm of the three-level converter is the normalised reference voltage vector. The normalisation depends on the number of levels of the multilevel converter and the voltage level value of the DC-link capacitors. As a result, *Va*, *Vb* and *Vc* (1) take entire values between 0 and $n-1$, where n is the number of the level of the multilevel converter. Thus, the first step consists of localising the sextant $n=1,\ldots,6$ where is located the reference voltage vector u^*. The voltage vector u^* is transformed into u_{flat}. This transformation consists of scaling an imaginary part and multiplying it by $1/\sqrt{(3)}$. The hexagon is flattened. Fig. 1 shows the regular hexagon defined by the switching state vectors before and after transformation in the complex plane. Since the transformed sextants are separated by 45° lines, the sextant can now be readily identified by comparing the real and imaginary parts of the complex transformed reference voltage vector u_{flat}. In addition, it can be easily proven [4] that, once the sextant has been determined, the numeric evaluation of the switching times is reduced to a single addition involving u_α and u_β.

The transformation of u^* into u_{flat} makes it possible to avoid on-line computations. These computations are substituted by decision making. The states space consists of a main regular hexagon. Each sextant of this hexagon is now divided into several sectors. The number of sectors depends on the number of levels of the multilevel converter.

Determination of sextant of reference voltage vector: Once the sextant n has been localised into the main regular hexagon, the identification of the sector into the sextant, the nearest switching sequence to approximate a reference voltage vector and the on-state durations are calculated by rotating u^* to the first sextant. The rotated reference voltage vector is:

$$u_g = u_{ga} + j u_{gb} = u \cdot \exp\left(-j(n-1)\frac{\pi}{3}\right); n = 1,\ldots,6 \quad (3)$$

This vector u_g is transformed in another with an identical real part and a reduced imaginary part: