Low Bit-Rate Encoding Algorithm for Distributed Speech Recognition Based on SVD Decomposition

A. Touazi and M. Debyeche

Abstract—**The paper presents an algorithm for compression of front-end feature extracted parameters used in Distributed Speech Recognition (DSR). In the proposed method the source encoder is mainly based on truncated Singular Value Decomposition transform (SVD) with conventional vector and scalar quantizers. The system provides a compression bit-rate around 3500 bps. The experiments were carried out on the TIDigitsAurora-2 database using Hidden Makcov Model Toolkit (HTK). The simulation results show good recognition performance without dramatic change, comparing toconventional ETSI Aurorastandard front-end feature compression algorithm with quantized features at 4400 bps.**

Index Terms—**Distributed speech recognition, vector and scalar quantizers, singular value decomposition, aurora-2 database.**

I. Introduction

The increasing use of mobile and World Wide Web networks for speech communication has led to Distributed Speech Recognition (DSR) systems being developed and standardized by the European Telecommunication Standards Institute ETSI [1]. As shown in Fig. 1, the basic idea of DSR consists of using a local Front-end (FE) from which speech features are extracted and transmitted through a data channel to a remote Back-end (BE) or remote server recognizer. The speech features used for recognition are the first 12 MFCCs c1-c12, the zerothcepstral coefficient c0 and the log energy log*E*in the frame. The 14-dimentional feature vector is split into seven sub-vectors.

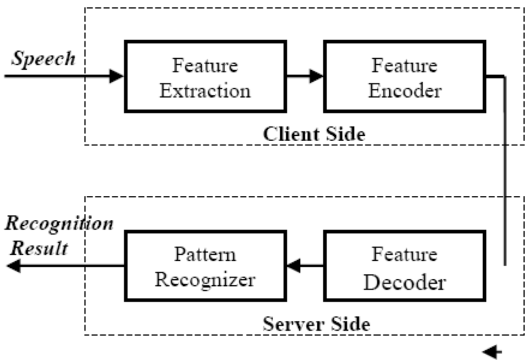

Fig. 1. A DSR block model.

Each of the sub-vectors is encoded with a different 2-dim Vector Quantizer (VQ). The standard computes a feature vector every 10ms and allocates 44 bits to each feature vector to achieve a total bit-rate of 4400 bps [1]. The number of bits allocated to the different sub-vectors is shown in Table I.

TABLE I: Bit Allocation in ETSI Aurora Standard

Sub-vector	Bits allocated
c1, c2	6
c3, c4	6
c5, c6	6
c7, c8	6
c9, c10	6
c11, c12	6
c0, log*E*	8

The Aurora-2 database [2] consists of connected digit sequences for American English Talkers. It provides speech samples and scripts to perform speaker independent speech recognition experiments in clean and noisy conditions. This database has been prepared by down-sampling to 8 kHz, filtering with the G.712 and MIRS characteristics; noise is artificially added to the filtered TIDigits at a desired SNR (20, 15, 10, 5, 0, -5dB) with including clean condition, and eight different noise conditions such as:

- Subway
- Babble
- Car
- Exhibition hall
- Restaurant
- Street
- Airport
- Train station.

Various schemes for compressing the MFCC vectors have been proposed in the literature. Among these methods there are the coding based on Discrete Cosine Transforms (DCT & 2DCT) [3], [4] and another method that exploits the mutual information measure between feature sub-vectors [5].

In this paper a truncated Singular Value Decomposition (SVD) transform [6] is used to compress feature vectors. This transform is widely used in signal processing such as image coding systems and noise reduction. In the proposed method we applied the same principle that employed in image compression by stacking a set of MFCC feature vectors to have a matrix structure. The rest of the paper is organized as follow: Section II introduces a general overview of SVD transform, a detailed description of the algorithm is provided in Section III. In Section IV we summarize the experimental results. Finally in Section V we offer our conclusion.

II. Singular Value Decomposition

Singular Value Decomposition is an extremely powerful

Manuscript received October 9, 2012; revised January 19, 2013.
The authors are with the University of Science and Technology Houari Boumediene (USTHB), Faculty of Electronics and Computer (FEI), Signal Processing and Speech Communication Laboratory (LCPTS) (e-mail: touazi.azzedine@gmail.com, mdebyeche@gmail.com; mdebyeche@usthb.dz).

DOI: 10.7763/IJIEE.2013.V3.373

and useful tool in linear algebra. Let's say we have a matrix A with m rows and n columns, then there exist orthogonal matrices U ($m \times m$) and V ($n \times n$), such that:

$$U = [u_1, u_2, \ldots, u_m] \tag{1}$$

$$V = [v_1, v_2, \ldots, v_n] \tag{2}$$

It can be proven that [7]:

$$U^T AV = diag(\sigma_1, \ldots, \sigma_p); p = \min(m, n) \tag{3}$$

where:

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p \geq 0 \tag{4}$$

The σ_i are the singular values of A and the vectors $u_i$and $v_i$are the *ith*left singular vector and the *ith* right singular vector respectively. Then A can be factorized into three matrices:

$$A = USV^T \tag{5}$$

Here, S is an $m \times n$ diagonal matrix with singular values (σ_i) on the diagonal. The SVD reveals a great deal about the structure of matrix. If the SVD of A is given by (5) and we define r by:

$$\sigma_1 \geq \cdots \geq \sigma_r > \sigma_{r+1} = \cdots = \sigma_p = 0 \tag{6}$$

Then:

$$Rank(A) = r \tag{7}$$

So we have the compact SVD defined by:

$$A = \sum_{i=1}^{r} \sigma_i u_i v_i^T \tag{8}$$

In other words, the rank of matrix A is equal to the number of its nonzero singular values [7].

A. Truncated SVD

In the truncated version, the SVD of A given by (8) can be adjusted by:

$$A^* = \sum_{i=1}^{t} \sigma_i u_i v_i^t \qquad where \quad t < r \tag{9}$$

Only the t column vectors of U and the$t$ column vectors of V corresponding to the t largest singular values are calculated. The rest of the matrix is discarded; this can be much quicker and more economical than the compact SVD if t<<r. The approximate matrix A^* is in a very useful sense the closest approximation to A that can be achieved by a matrix of rank t [7].

III. Compression Algorithm

The use of this method is motivated by the SVD energy compaction property or truncated SVD, The analysis part of the algorithm is depicted in Fig. 2. It can be seen that 12 successive MFCC vectors are stacked together to form a block of 14×12 (matrix of 14 rows and 12 columns).

By considering the high difference in magnitude between (c0, log$E$) and the rest of MFCC coefficients, theblock of 14×12 is split into two sub-blocks of 12×12 and 2×12, such that the rank of the first sub-block equals to 12 and the rank of the second sub-block equals to 2. In the next step and by applying a truncated SVD for each sub-block, various experiments have been performed to evaluate the adequaterank. Therefore the new ranks for the truncated versionswill be set to 1 and 5 respectively.

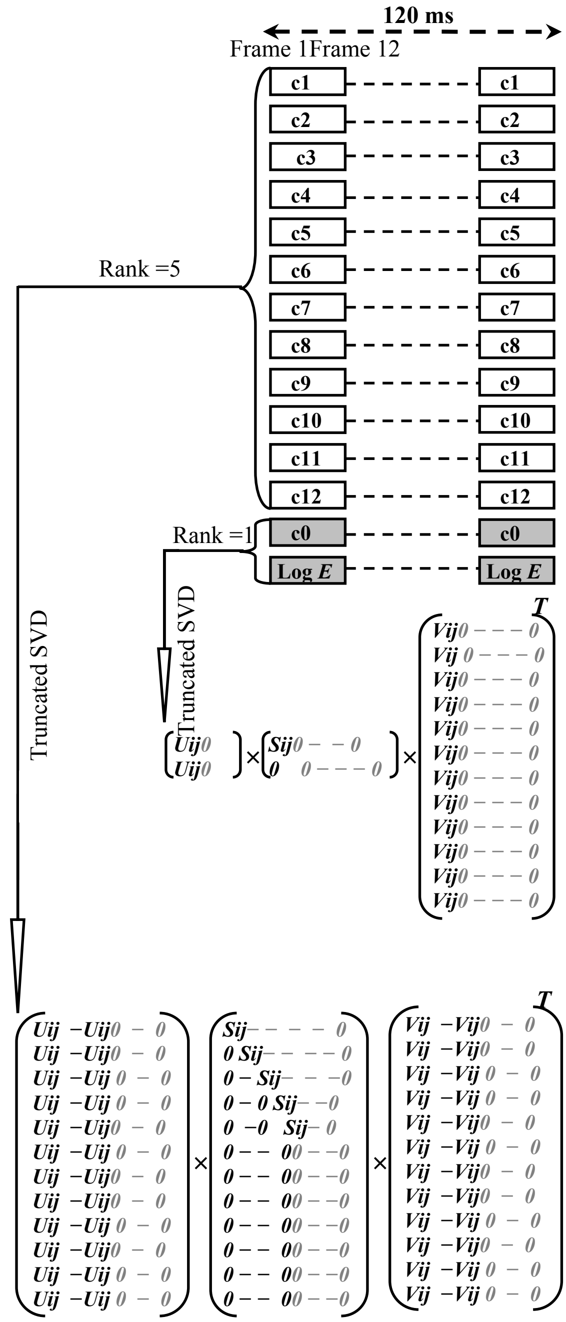

Fig. 2. SVD transforming for MFCC block.

The choice of these new ranks is approved by an experiment with comparing the SNR average (sets A, B and C) of each MFCC coefficient in the case of both Aurora encoder and truncated SVD with different ranks (1, 4, 5, 6 and 8). As shown in Fig. 3 for the first sub-block (c1-c12) it can be seen that in the truncated SVD at the rank number 5 the SNR degrees are higher than the Aurora encoder for the first five coefficients (c1-c5) and are decreasing from the coefficient c6. It is well known that the lower feature coefficients provide the greatest contribution to recognition performance [8]. Thus, a truncated SVD with a rank number 5 can lead to a minor influence in the recognition performance. Another reason to choose Rank 5 is due to the gain on computational cost that we can achieve in the quantization phase, with maintaining the recognition performance comparing with superior ranks.

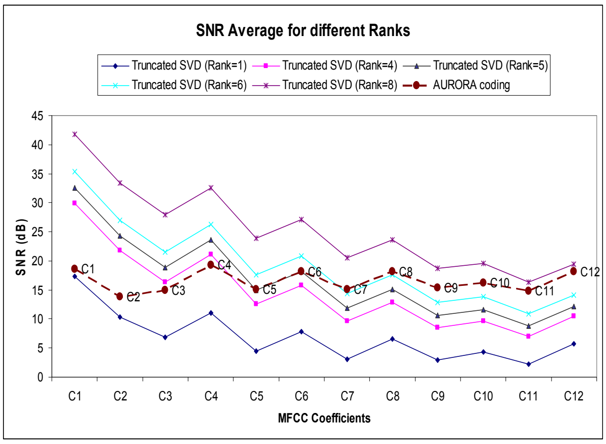

Fig. 3. SNR measurement for each MFCC coefficient (c1-c12) with different ranks.

For the second sub-block (c0, log*E*), from the results shown in Table II and comparing to ETSI Aurora encoding, for the new truncated SVD with rank 1 it is very likely that we can improve the recognition performance if we use c0 in the recognition task; unless if we use log*E* the performance will be smoothly degraded.

TABLE II: SNR MEASUREMENTFOR ENERGY COEFFICIENTS

MFCC Coefficient	Aurora coding	Truncated SVD (Rank=1)
c0	41.87	77.03
log*E*	40.44	34.01

In the quantization phase, for the first sub-block all columns vectors of both matrix *U* and *V* are encoded using Split Vector Quantizer (SVQ) with the same codebooks, in which each columnvector is split into four sub-vectors and each sub-vector is quantized using its own VQ codebook trained with LBG algorithm [9]. The first and second column vectors for matrix *U* and *V* are encoded with codebooks of size 512 each. The third and fourth column vectors are encoded with codebooks of size 256 each. The fifth vectors are encoded with codebooks of size 128 each. The five singular values of matrix *S* are encoded using uniform scalar quantization of 8, 8, 8, 8, and 7 bits respectively.

For the second sub-block, the first column vector of *V* is encoded using SVQ in which this last is split into four sub-vectors and each of them is quantized using its own VQ codebook of size 512. The first vector column of *U* is encoded using VQ with codebook of size 512. The first singular value of *S* is encoded using uniform scalar quantization of 10 bits. In order to minimize the computational costin the quantization of the first singularvalue of *S*,the 1024(for 10 bits) values are sorted anddivided into four codebooks of 256 values each, then the scalar quantization is performed through 2 stages, the first stage for determining the nearest codebook that we can use (2bits) and the second stage for the quantization (8bits).

The decoding process consists of the inverse operations of the encoding in reverse order. TheTable below shows the bits allocation for each sub-blockwith total of 422 bits by block of 120 ms.Then the resulting quantization bit-rate is around 3.51 kbps.

TABLE III: SVD ENCODER BITS ALLOCATION

	i	u_i	v_i	σ_i
	1	36	36	8
	2	36	36	8
Sub-Block 1	3	32	32	8
	4	32	32	8
	5	28	28	7
Sub-Block 2	1	9	36	10

IV. EXPERIMENTS AND RESULTS

The experiments were carried out on the TIDigitsAurora corpus (Test sets A, B, and C) with MFCC vectors extracted using the STQ-Aurora front-end algorithm [1]. In the figures 4, 5, and 6 we compared the SNR average results for the following cases:

- Aurora encoder working at 4.4 kbps [1].
- Proposed SVD encoder working at 3.5 kbps.
- Uncompressed truncated SVD (Rank =5).

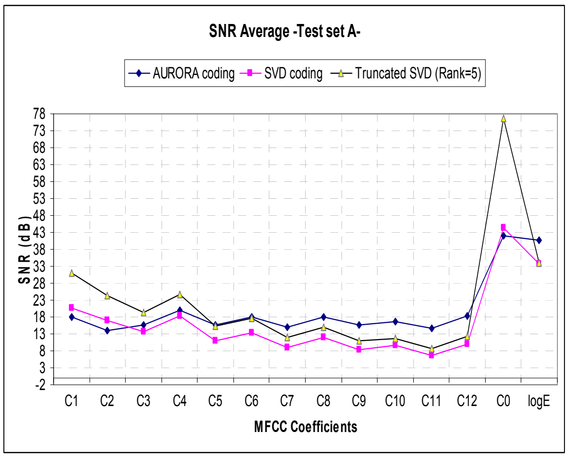

Fig. 4. SNR measurements (Test set A).

As seen in Table IV, for (c0-c12) coefficients we note degradation from SNR levels after quantization; but for the first five MFCC coefficients (c0- c5) we got acceptable SNR values when comparing toAurora encoder. Also, we observe acceptable valuesin case ofc0 andlog*E*.

The recognition were done using HTK 3.4 speech recognizer [10] to the coded MFCCs, while the c0 and logecoefficients are both used in the compression and only log*E* is used in the recognition task. However, the results are compared for both compressed and uncompressed Aurora recognition performance.

As it can be shown from Fig. 7, 8 and Table V, in the clean condition the word level accuracies for SVD encoder are slightly superior in comparison with the compressed Aurora features and slightly inferior in the case of multi-condition.

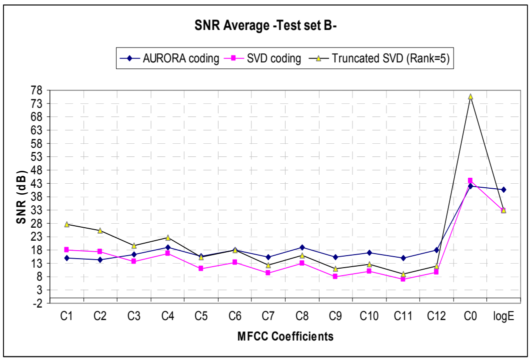

Fig. 5. SNR measurements (Test set B).

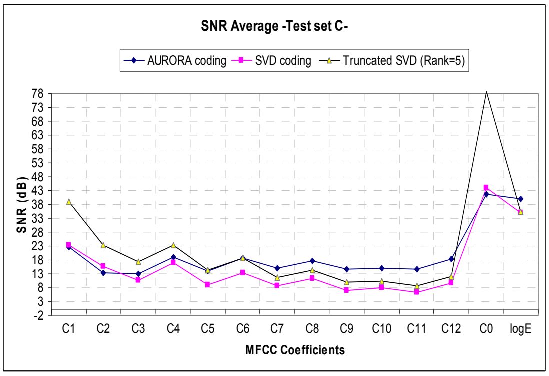

Fig. 6. SNR measurements (Test set C).

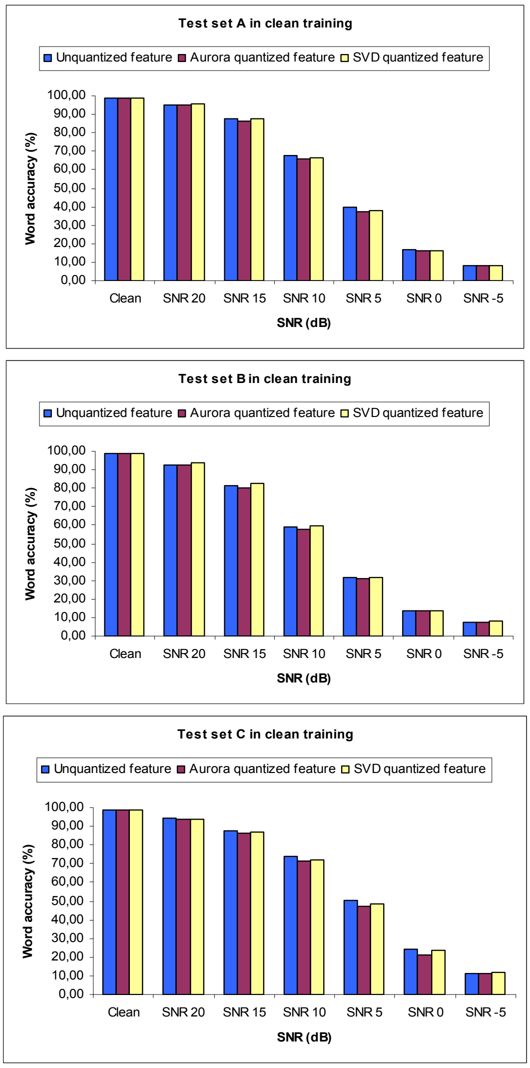

Fig. 7.Word accuracy before and after compression, in clean condition (Test sets A, B and C).

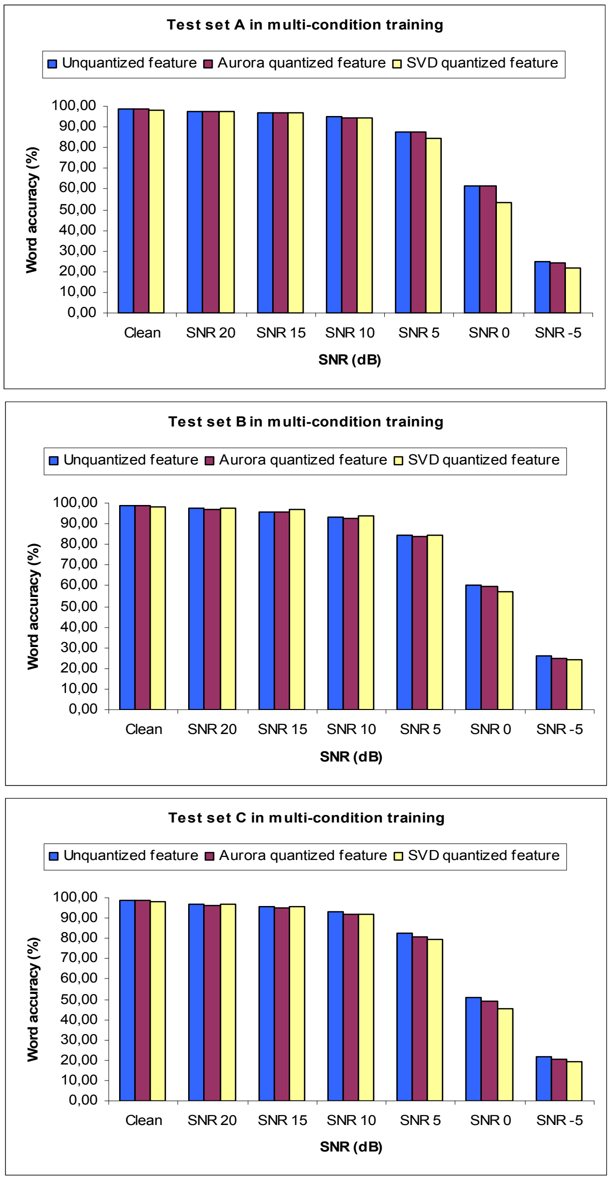

Fig. 8. Word accuracy before and aftercompression, in multi-condition (Test sets A, B and C).

TABLE IV: SNR MEASUREMENTS AVERAGE FOR TEST SETS (A, B AND C)

MFCC Coefficients	Aurora Encoding [1]	Truncated SVD (Rank=5)	SVD Encoding
c1	18.62	32.59	20.68
c2	13.78	24.32	16.8
c3	14.97	18.85	12.6
c4	19.32	23.58	17.4
c5	15.14	15.21	10.24
c6	18.21	18.19	13.37
c7	15.08	11.94	8.94
c8	18.14	15.12	12
c9	15.35	10.62	7.81
c10	16.21	11.6	9.23
c11	14.84	8.81	6.55
c12	18.14	12.19	9.82
c0	41.87	77.03	44.09
log*E*	40.44	34.01	33.73
Average (c1- c5)	16.36	22.91	15.54
Average (c0, log*E*)	41.15	55.52	38.91

TABLE V: WORD ACCURACY AVERAGE (0 – 20 DB), FOR TEST SETS (A, B, AND C)

Set	Training mode	Unquantized Aurora	Quantized Aurora	Quantized SVD
A	Clean	67.62	66.65	**67.03**
	Multi-Condition	89.60	**89.58**	87.46
B	Clean	62.96	62.29	**63.25**
	Multi-Condition	88.31	87.91	**88.00**
C	Clean	71.62	**69.80**	**70.58**
	Multi-Condition	86.24	**85.30**	84.56

V. CONCLUSION AND FURTHER WORK

In the proposed SVD algorithm we focused on reducing the bit-ratearound 3500 bps. Generally this source encoder maintains the same recognition performance comparing with the conventional ETSI Auroraencoder working at 4400 bps, with relatively more computational cost. In addition, the proposed technique can be extended to compress othertypes of parameters like LPCcoefficients. Further work will involve improving both computational cost by proposing a new quantization techniques for the SVD vectors and recognition performance.

ACKNOWLEDGMENT

The authors gratefully acknowledge the Signal Processingand Speech Communication Laboratoryteam for their contribution to carry out this research work.

REFERENCES

[1] *Front-End Feature Extraction Algorithm Compression Algorithms*, Speech Processing, Transmission, and Quality Aspects, Distributed Speech Recognition, European Telecomm Standards Inst (ETSI), ES, 2003.

[2] H. G Hirsch and D. Pearce, "The aurora experimental framework for the performance evaluations of speech recognition systems under noisy conditions," in *Proc. ASR2000 ISCA ITRW*, Paris, 2000, pp. 181-188.

[3] I. Kiss and P. Kapanen, "Robust feature vector compression algorithm for distributed speech recognition," in *Proc.1999 ICSLP Conf.*, 1999, pp. 2183-2186.

[4] Q. Zhu and A. Alwan, "An efficient and scalable 2D-DCT based feature coding scheme for remote speech recognition," in *Proc. 2001 ICASSP Conf.*, 2001, pp. 113-116.

[5] N. Srinivasamurthy, A. Ortega, and S. Narayanan, "Enhanced standard compliant distributed speech recognition (Aurora encoder) using rate allocation," in *Proc. 2004 ICASSP Conf.*, 2004, pp. 485-488.

[6] Singular Value Decomposition. [Online]. Available: http://en.wikipedia.org/wiki/Singular_value_decomposition

[7] G. H. Golub and C. F. V. Loan, *Matrix Computations*, 3rd ed. Baltimore and London: The Johns Hopkins University Press, 1996, ch. 2, pp. 69-75.

[8] Y. Zhang, "Acoustic model and pronunciation adaptation in automatic speech recognition," Ph.D. dissertation, Dept. Elect. Comp. Eng., University of Miami, 2006.

[9] Y. Linde, A. Buzo, and R.M. Gray, "An algorithm for vector quantizer design, " *IEEE Trans. Commun*, vol. 28, pp. 84-95, Jan. 1980.

[10] S. Young, G. Evermann, M. Gales, T. Hain D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book*, HTK Version 3.4:Dept. Eng., Cambridge University, 2006.

Azzedine Touazi was born in Algeria. In the year 2003 he received his Engineer degree in Electronics from University of Science and Technology HouariBoumedieneof Algiers(USTHB),and Magister degree in Signal Processing and Communication from National Polytechnic Schoolof Algiers (ENP), during the year 2007. He is pursuing Ph.D in the field of speech coding and speech recognition in GSM Network. Presently he is working as research associate in the Center for the Development of Advanced Technologies (CDTA).His research interest includes digital speech and image processing, communication over the mobile network, and multimodal pattern recognition.

Mohamed Debyeche received the Engineer degree in Electrical Engineering from the National School Polytechnic of Algiers, Algeria in 1982, the Magister degree in 1991 in signal processing and the "Doctoratd'Etat"(PhD) in Speech recognition in 2007 from HouariBoumediene University, Algiers, Algeria. He is presently Professor of Electronics at University (USTHB) Department of Telecommunication. His expertise is in the Electronics field, more specially Speech Processing and Speech recognition. His current research concentrates on the development of new methods for speech recognition, recognition over the mobile network, and multimodal pattern recognition applied to Arabic language.