(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
29 January 2009 (29.01.2009)

PCT

(10) International Publication Number
WO 2009/014496 A1

(51) International Patent Classification:
G10L 15/00 (2006.01) *G10L 11/00* (2006.01)

(21) International Application Number:
PCT/SG2008/000213

(22) International Filing Date: 16 June 2008 (16.06.2008)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
11/829,031 26 July 2007 (26.07.2007) US

(71) Applicant *(for all designated States except US)*: **CREATIVE TECHNOLOGY LTD.** [SG/SG]; 31 International Business Park, Creative Resource, Singapore 609921 (SG).

(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: **XU, Jun** [SG/SG]; Blk 607, Choa Chu Kang St. 62, #06-119, Singapore 680607 (SG). **ZHANG, Huayun** [CN/SG]; Block 686, Jurong West Central 1, #12-138, Singapore 642686 (SG).

(74) Agent: **POH, Chee Kian, Daniel**; LIoyd Wise, Tanjong Pagar, P O Box 636, Singapore 910816 (SG).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, SV, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MT, NL, NO, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:
— *with international search report*

(54) **Title:** A METHOD OF DERIVING A COMPRESSED ACOUSTIC MODEL FOR SPEECH RECOGNITION

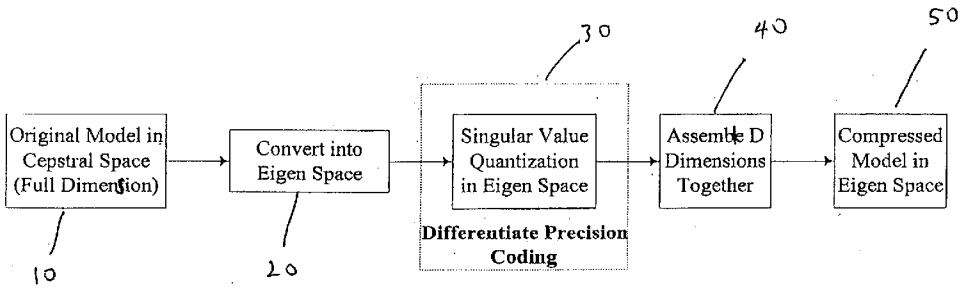

Figure 1

(57) **Abstract:** A method of deriving a compressed acoustic model for speech recognition is disclosed herein. In a described embodiment, the method comprises transforming an acoustic model into an eigenspace at step (20), determining eigenvectors of the eigenspace and their eigenvalues, and selectively encoding dimensions of the eigenvectors based on values of the eigenspace at step (30) to obtain a compressed acoustic model at steps (40 and 50).

WO 2009/014496 A1

A Method of Deriving A Compressed Acoustic Model for Speech Recognition

Background and Field of the Invention

This invention relates to a method of deriving a compressed acoustic model for speech recognition.

Speech recognition, or more commonly called automatic speech recognition has many applications such as automatic voice response, voice dialing and data entry etc. The performance of a speech recognition system is usually based on accuracy and processing speed and a challenge is to design speech recognition systems with lower processing power and smaller memory size without affecting accuracy or processing speed. In recent years, this challenge is greater with smaller and more compact devices also demanding some form of speech recognition application.

In the paper "Subspace Distribution Clustering Hidden Markov Model" by Enrico Bocchieri and Brian Kan-Wing Mak, IEEE transactions on Speech and Audio Processing, Vol. 9, No. 3, March 2001, a method was proposed which reduces the parameter space of acoustic models, thus resulting in savings in memory and computation. However, the proposed method still requires a relative large amount of memory.

It is an object of the present invention to provide a method of deriving a compressed acoustic model for speech recognition which provides the public with a useful choice and/or alleviates at least one of the disadvantages of the prior art.

Summary of the Invention

This invention provides a method of deriving a compressed acoustic model for speech recognition. The method comprises: (i) transforming an acoustic model into eigenspace to obtain eigenvectors of the acoustic model and their eigenvalues, (ii) determining predominant characteristics based on the eigenvalues of every dimension of each eigenvector; and (iii) selectively encoding the dimensions based on the predominant characteristics to obtain the compressed acoustic model.

Through the use of eigenvalues, this provides means for determining the importance of each dimension of the acoustic model which forms the basis for the selective encoding. In this way, this creates a compressed acoustic model having a much reduced size, than in cepstral space.

Scalar quantization is preferred for the encoding since such quantizing is "lossless".

Preferably, determining the predominant characteristics includes identifying eigenvalues that are above a threshold. The dimensions corresponding to eigenvalues above the threshold may be coded with a higher quantization size than dimensions with eigenvalues below the threshold.

Advantageously, prior to the selectively encoding, the method includes normalising the transformed acoustic model to convert every dimension into a standard distribution. The selectively encoding may then include coding each normalised dimension based on a uniform quantization code book. Preferably, the code book has a one byte size, although this is not absolutely necessary and depends on the application.

If one byte code book is used, then preferably, the normalised dimensions having an importance characteristic higher than an importance threshold is coded using one byte code word. On the other hand, the normalised dimensions having an importance characterise lower than an importance threshold may then be coded using a code word of less than 1 byte.

The invention further provides an apparatus/system for deriving a compressed acoustic model for speech recognition. The apparatus comprises means for transforming an acoustic model into eigenspace to obtain eigenvectors of the acoustic model and their eigenvalues, means for determining predominant characteristics based on the eigenvalues of every dimension of each eigenvector; and means for selectively encoding the dimensions based on the predominant characteristics to obtain the compressed acoustic model.

Brief Description of the Drawings

An embodiment of the invention will now be described, by way of example, with reference to the accompanying drawings in which,

Figure 1 is a block diagram showing a broad overview of a process for deriving a compressed acoustic model in eigenspace for speech recognition;

Figure 2 is a block diagram showing the process of Figure 1 in greater detail and also including decoding and decompression steps;

Figure 3 is a graphical representation of linear transformation of an uncompressed acoustic model;

Figure 4 including Figures 4a to 4c are graphs showing standard normal distribution of dimensions of eigenvectors after normalisation;

Figure 5 illustrates the different coding techniques with and without discriminant analysis; and

Figure 6 is a table showing different model compression efficiencies.

Detailed Description of the Preferred Embodiment

Figure 1 is a block diagram showing a broad overview of a preferred process for deriving a compressed acoustic model of this invention. At step 10, an original uncompressed acoustic model is first translated and represented in cepstral space and at step 20, the cepstral acoustic model is converted into eigenspace to determine what parameters of the cepstral acoustic model are

important/useful. At step 30, parameters of the acoustic model are coded based on the importance/usefulness characteristics and thereafter, the coded acoustic features are assembled together as a compressed model in eigenspace at steps 40 and 50.

Each of the above steps will now be described in greater detail by referring to Figure 2.

At step 110, the uncompressed original signal model such as, for example, speech input is represented in cepstral space. A sampling of the uncompressed original signal model is taken to form a model in cepstral space 112. The model in cepstral space 112 forms a reference for subsequent data input. The cepstral acoustic model data is then subjected to discriminant analysis at step 120. A Linear Discriminant Analysis (LDA) matrix is employed to the uncompressed original signal model (and sampling) to transform the uncompressed original signal model (and sampling) in cepstral space into data in eigenspace. It should be noted that the uncompressed original signal model is a vector quantity, and thus includes a quantity and a direction.

A. Discriminant Analysis

Through linear discriminant analysis, the most predominant information in the sense of acoustic classification is explored, evaluated and filtered. This is based on the realisation that in speech recognition, it is important that the speech received is processed accurately, but it may not be necessary to code all

features of the speech since some may not be necessary and would not contribute to the accuracy of the recognition.

Let's assume R^n is the original feature space, which is a n-dimension hyperspace. Each $\mathbf{x} \in R^n$ has a class label that is meaningful in ASR systems. Next, at step 130, an aim is to find a linear transformation (LDA matrix) $\mathbf{A}$, by converting into eigenspace, that optimize the classification performance in the transformed space $\mathbf{y} \in R^p$, which is a p-dimension hyperspace (normally, $p \leq n$), where

$$\mathbf{y} = \mathbf{A}\mathbf{x}$$

with **y** being a vector in eigenspace and **x** being data in cepstral space.

In LDA (Linear Discriminant Analysis) theory, $\mathbf{A}$ can be found from

$$\mathbf{\Sigma}_{WC}^{-1}\mathbf{\Sigma}_{BC}\mathbf{\Phi} = \mathbf{\Phi}\mathbf{\Lambda}$$

where $\mathbf{\Sigma}_{WC}$ and $\mathbf{\Sigma}_{BC}$ are the within class (*WC*) and across class (*BC*) covariance matrix respectively, and $\mathbf{\Phi}$ and $\mathbf{\Lambda}$ are $n \cdot n$ matrix of eigenvectors and eigenvalues of $\mathbf{M}_{WC}^{-1}\mathbf{M}_{BC}$, respectively.

$\mathbf{A}$ is constructed by choosing p eigenvectors corresponding to p largest eigenvalues. When **A** is derived correctly from **y** and **x**, an LDA matrix that optimises acoustic classification is derived which aids in exploring, evaluating and filtering the uncompressed original signal model.

Figure 3 shows graphically the end result of the linear transformation to reveal two classes of data along a useful dimension (Dim) and one nuisance dimension (Dim) which has no useful information. The classes of data may be, for example, phoneme, biphoneme, triphoneme and so forth. A first ellipse 114 and a second ellipse 116 both represent regions of data resulting from Gaussian distributions. A first bell curve 115 results from a projection of points from within the first ellipse 114 onto a first sub-axis 118. Similarly, a second bell curve 117 results from a projection of points from within the second ellipse 116 onto the first sub-axis 118. The first sub-axis 118 is derived using LDA on the regions of data shown in the first ellipse 114 and the second ellipse 116. A second sub-axis 119 which is orthogonal to the first sub-axis 118 is inserted at the point of intersection between the first ellipse 114 and the second ellipse 116. The second sub-axis 119 clearly separates data points into separate classes as the first ellipse 114 and the second ellipse 116 are merely approximate regions of separate classes. Thus, the classes present in the uncompressed original signal model are ascertained from the relative positions of the separated data regions. This technique may be employed primarily for the separation of two classes of data. Each class of data may also be known as a feature of the acoustic signal.

As it would be appreciated, from the data distribution of the two classes, and through LDA, it is possible to determine the eigenvalues of corresponding eigenvectors defined in order of dominance or importance based on the eigenvalues. In other words, with LDA, higher eigenvalues represents more

discriminative information whereas lower eigenvalues represent lesser discriminative information.

After each feature of the acoustic signal is classified based on their predominant characteristics in the speech recognition, the acoustic data is normalised at 140.

B. Normalisation in eigenspace

Mean estimation in eigen-space:

$$\boldsymbol{\mu} = E(\mathbf{y}_t) = \frac{1}{T}\sum_{t=1}^{T}\mathbf{y}_t$$

Standard Variance estimation in eigen-space:

$$\boldsymbol{\Sigma} = E\left((\mathbf{y}_t - E(\mathbf{y}_t))(\mathbf{y}_t - E(\mathbf{y}_t))^T\right) = E\left(\mathbf{y}_t\mathbf{y}_t^T\right) - E(\mathbf{y}_t)E(\mathbf{y}_t)^T$$

$$\boldsymbol{\Sigma}_{diag} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{y}_t^T\mathbf{y}_t - \boldsymbol{\mu}^T\boldsymbol{\mu}$$

Normalization:

$$\hat{\mathbf{y}}_t = sqrt(\boldsymbol{\Sigma}_{diag})\cdot(\mathbf{y}_t - \boldsymbol{\mu})$$

where y_t = eigenspace vector, $E(y_t)$ = expectation of y_t, Σ_{diag} = covariance matrix of elements on diagonal of variance, and T = time.

Speech feature is assumed as Gaussian distributions, this normalization converts every dimension into a standard normal distribution $N(\mu,\sigma)\ with\ \mu = 0\ and\ \sigma = 1$ (see Figures 4a to 4c).

This normalization provides two advantages for the model compression:

Firstly, since all the dimensions share the same statistics, a uniform singular codebook can be employed for model coding-decoding at every dimension. There is no need to design different codebooks for different dimensions or use other kinds of vector codebooks. This could save memory space for model storing. If the size of the codebook is defined as $2^8 = 256$, one byte is enough to represent a code word.

Secondly, since the dynamic range of a codebook is limited compared to floating point representation, model coding-decoding may bring serious problems when floating point data falls outside the range of the codebook, such as overflow, truncation and saturation, which will eventually result in ASR performance degradation. With this normalization, this conversion loss can be effectively controlled. For example, if the fix-point range is set as $\pm 3\sigma$ confidence interval, the data percentage that causes saturation problem in coding-decoding would be:

$$\int_{-\infty}^{\mu-3\sigma} N_{y_i}(\mu,\sigma)dy_i + \int_{\mu-3\sigma}^{+\infty} N_{y_i}(\mu,\sigma)dy_i \approx 0.26\%$$

It has been found that this minor coding-decoding error/loss is unobservable in ASR performance.

C. Different Coding-Decoding Precision Based on discriminant capability.

After the model is normalised, it is subjected to discriminant or selective coding at 150 of the mean vectors and covariance matrices of the acoustic model based on the quantization code book size of 1 byte. The LDA projection on the eigenvector corresponding to larger eigenvalues is considered to be more important to classification. The larger the eigenvalue, the higher importance of its corresponding direction in the sense of ASR. Thus, the maximum code word size is used to represent the class.

A threshold to segregate the "larger eigenvalues" and the other eigenvalues is determined through cross validation experiments. Firstly, a part of training data and training model is set aside. The ASR performance is then evaluated based on the set-aside data. This process of training and evaluating the ASR performance is repeated for different thresholds until a threshold value is found that provides the best recognition performance.

Since dimensions in eigenspace have different importance characteristics for voice classification, different compression strategies with different precisions are employed without affecting ASR performance. Also, since all the parameters of the acoustic model are multidimensional vectors or matrices, scalar coding is implemented on every dimension of each model parameter. This is particularly advantageous since scalar coding is "lossless". In this instance, scalar coding is "lossless" compared with ubiquitous vector quantization (VQ). VQ is a lossy compression method. The size of VQ codebook has to be increased in order to reduce quantization error. However, a larger codebook results in larger compressed model size and slower decoding

process. Furthermore, it's difficult to "train" a large VQ codebook robustly with limited training data. This difficulty would reduce the accuracy for speech recognition. It should be noted that the size of a scalar codebook is significantly less. This correspondingly helps to improve decoding speed. A small scalar code book may also be estimated more robustly than a large VQ code book with limited training data. Using the small scalar code book may also help avoid additional accuracy loss introduced by quantization error. Thus, scalar quantization outperforms VQ in relation to speech recognition with limited training data.

The selective coding is illustrated in Figure 5 in which dimensions having higher eigenvalues are coded using the maximum 8 bits (1 byte) whereas dimensions having lower eigenvalues are coded using lower bits. Through this selective coding, it would be appreciated that a reduction in memory size can be achieved.

After the selective coding, a compressed model in eigenspace is derived at 160. The compressed model in eigenspace is significantly smaller than data in cepstral space.

Figure 2 also illustrates decoding steps 170 and 180 where, if necessary, the compressed model are decoded in a discriminant manner and the compressed model decompressed to obtain the original uncompressed model.

An example of the of the compression efficiency is shown in Figure 6 which is a table depicting compression ratios of equal compression techniques compared with selective compression techniques as proposed by this invention. It can be seen that the selective compression technique can achieve a higher compression ratio.

Having now fully described the invention, it should be apparent to one of ordinary skill in the art that many modifications can be made hereto without departing from the scope as claimed.

CLAIMS

1. A method of deriving a compressed acoustic model for speech recognition, the method comprising

(i) transforming an acoustic model into eigen space to obtain eigenvectors of the acoustic model and their eigenvalues,

(ii) determining predominant characteristics based on the eigenvalues of every dimension of each eigenvector; and

(iii) selectively encoding the dimensions based on the predominant characteristics to obtain the compressed acoustic model.

2. A method according to claim 1, wherein coding the dimensions includes scalar quantizing of the dimensions in eigenspace.

3. A method according to claim 1, wherein determining the predominant characteristics includes identifying eigenvalues that are above a threshold.

4. A method according to claim 3, wherein dimensions corresponding to eigenvalues above the threshold are coded with a higher quantization size than dimensions with eigenvalues below the threshold.

5. A method according to claim 1, further comprising, prior to the selectively encoding, normalising the transformed acoustic model to convert every dimension into a standard distribution.

6. A method according to claim 5, wherein the selectively encoding includes coding each normalised dimension based on a uniform quantization code book.

7. A method according to claim 5, wherein the code book has a one byte size.

8. A method according to claim 6, wherein the normalised dimensions having an importance characteristic higher than an importance threshold is coded using one byte code word.

9. A method according to claim 6, wherein normalised dimensions having an importance characterise lower than an importance threshold is coded using a code word of less than 1 byte.

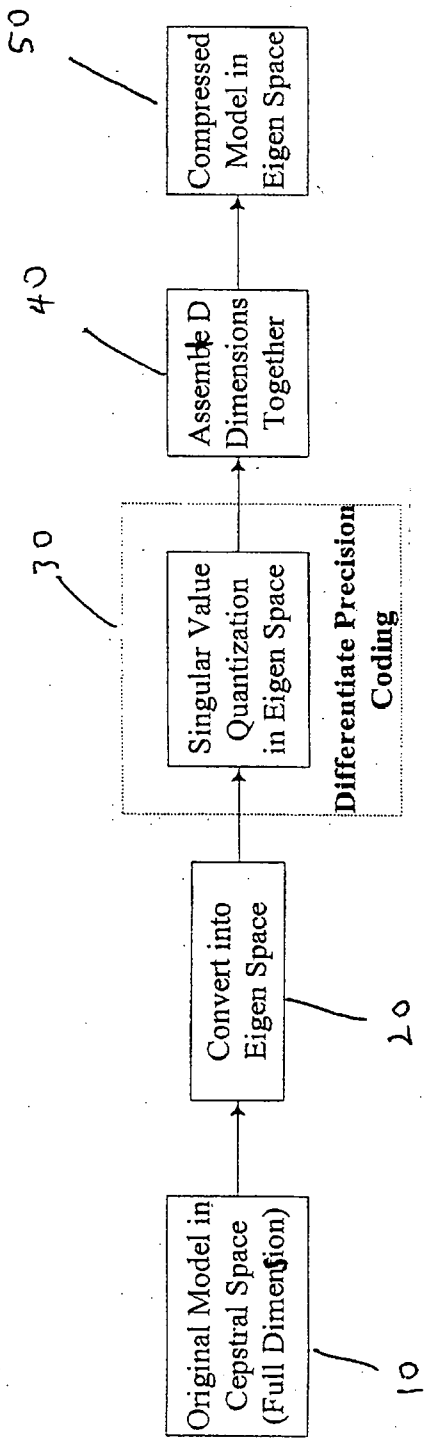
Original Model in Cepstral Space (Full Dimension)
10
Convert into Eigen Space
20
Differentiate Precision Coding
30
Singular Value Quantization in Eigen Space
Assemble D Dimensions Together
40
Compressed Model in Eigen Space
50

Figure 1

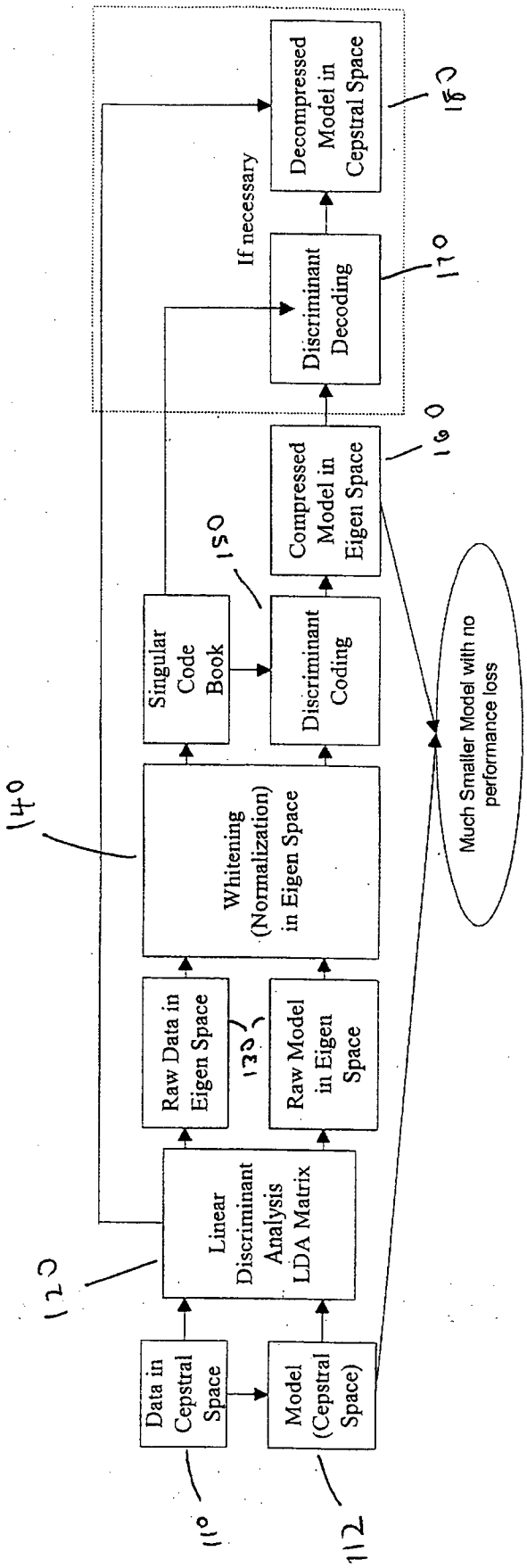
Data in Cepstral Space
Model (Cepstral Space)
Linear Discriminant Analysis LDA Matrix
Raw Data in Eigen Space
Raw Model in Eigen Space
Whitening (Normalization) in Eigen Space
Singular Code Book
Discriminant Coding
Compressed Model in Eigen Space
Discriminant Decoding
Decompressed Model in Cepstral Space
If necessary
Much Smaller Model with no performance loss
110
112
120
130
140
150
160
170
180

Figure 2

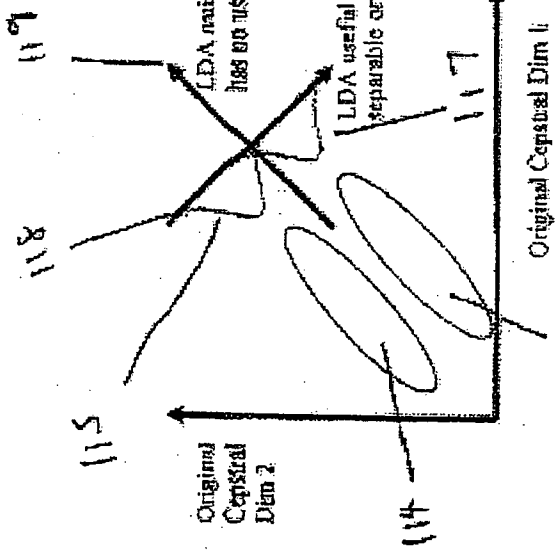
LDA nuisance Dim, which has no useful information
LDA useful Dim, 2 classes are most separable on this direction.
Original Cepstral Dim 2
Original Cepstral Dim 1
119
118
115
117
114
116

Figure 3

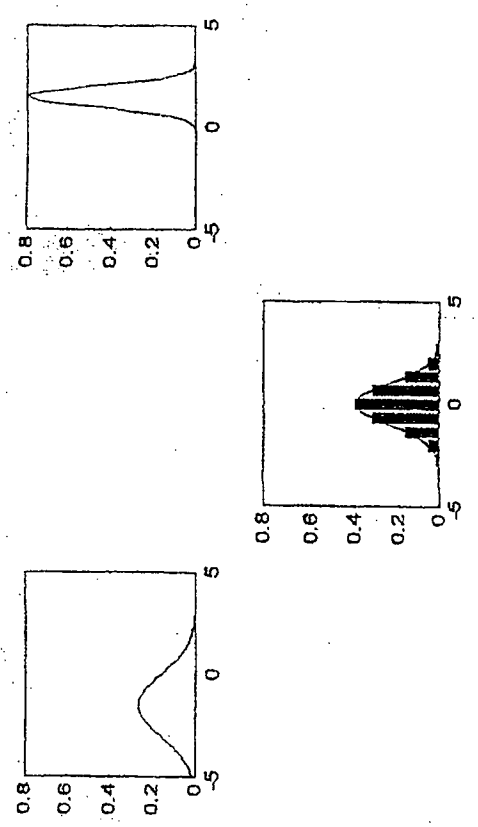

Figure 4

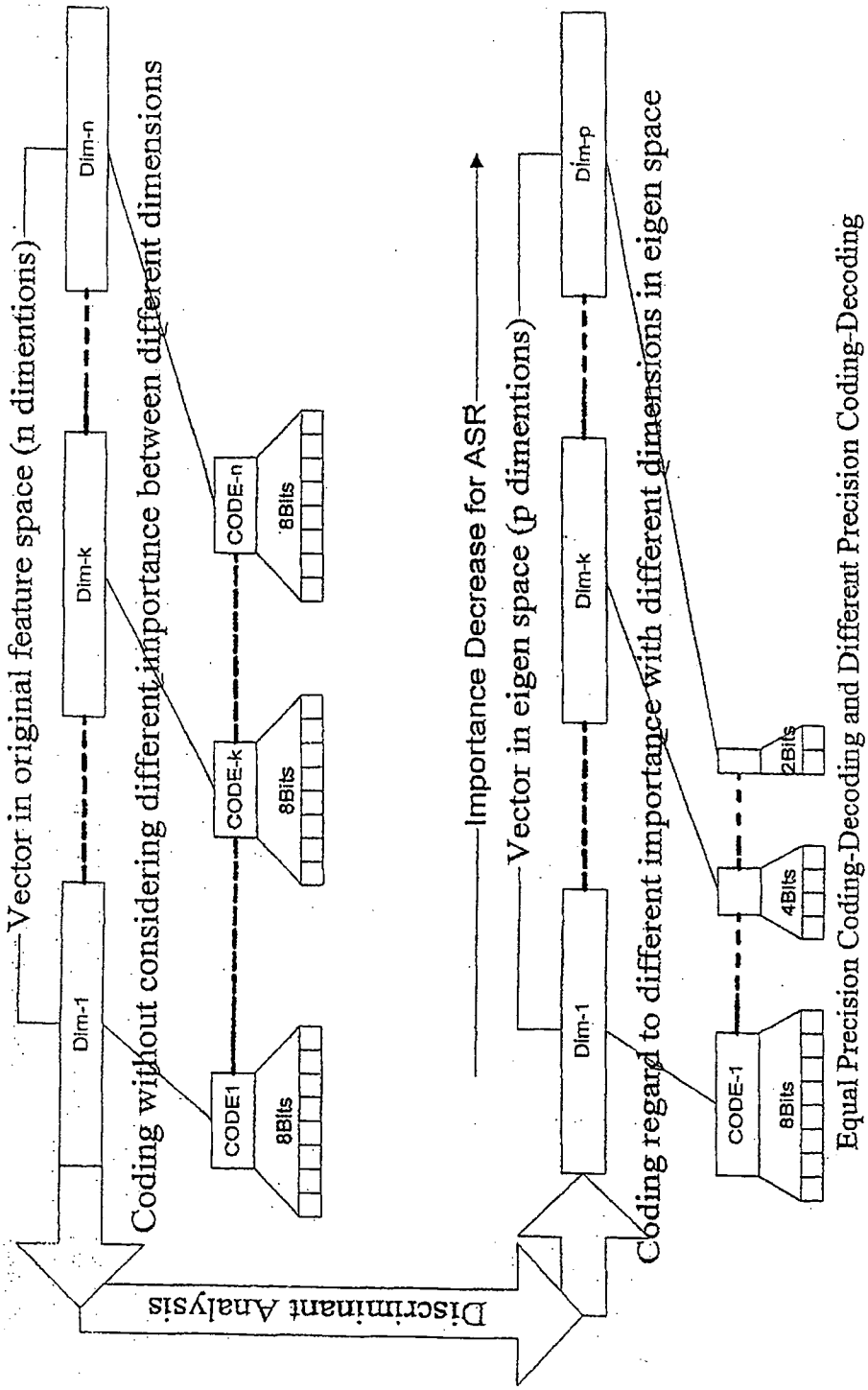
Vector in original feature space (n dimentions)
Dim-1
Dim-k
Dim-n
Coding without considering different importance between different dimensions
CODE1
CODE-k
CODE-n
8Bits
8Bits
8Bits
Discriminant Analysis
Importance Decrease for ASR
Vector in eigen space (p dimentions)
Dim-1
Dim-k
Dim-p
Coding regard to different importance with different dimensions in eigen space
CODE-1
8Bits
4Bits
2Bits
Equal Precision Coding-Decoding and Different Precision Coding-Decoding

Figure 5

Figure 6

Coding-Decoding Method	DataFormat	Model Size Before Compression	Model Size After Compression	Compression Ratio
Equal Precision Compression	FloatPoint(float32)	$M*n*4*2$	$256*4*2+M*n*2$	≈ 4
	FixPoint (32bits)	$M*n*4*2$	$256*4*2+M*n*2$	≈ 4
	FixPoint (16bits)	$M*n*2*2$	$256*2*2+M*n*2$	≈ 2
Different Precision Compression	FloatPoint(float32)	$M*n*4*2$	$n*n+256*4*2+M*\frac{7}{12}n*2$	≈ 7
	FixPoint (32bits)	$M*n*4*2$	$n*n+256*4*2+M*\frac{7}{12}n*2$	≈ 7
	FixPoint (16bits)	$M*n*2*2$	$n*n+256*2*2+M*\frac{7}{12}n*2$	≈ 3.5

Notes:

1. M is the number of Gaussions contained in an acoustic model. Commonly, M varies from tens of thousands to millions.
2. n is dimension of speech observations.
3. The codebook size is 256. So each code word is one byte.
4. Obviously, compare to $M*n$, the size of codebook and the size of matrix, $n*n$, can be both ignored.
5. Only mean vector and diagonal covariance parameters are compared.
6. For Different Precision Compression, we set $p=n$.
7. For Different Precision Compression, 3 different coding-decoding schemes are employed:
   For the first $n/3$, 8 bits are used;
   For the second $n/3$, 4 Bits are used;
   For the last $n/3$, 2Bits are used. (This strategy almost brings no loss for ASR performance)

INTERNATIONAL SEARCH REPORT

International application No.
PCT/SG2008/000213

A. CLASSIFICATION OF SUBJECT MATTER

Int. Cl.

G10L 15/00 (2006.01) ***G10L 11/00*** (2006.01)

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
DWPI: IPC G10L, Keywords- Speech, voice, recog+, identi+, acoustic, eigen+ and like words.
E-Space and USPTO Databases: Similar Keywords.
IEEE Databases: Keywords- voice, speech and eigen+

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
	US 2002/0143539 A1 (BOTTERWECK) 3October 2002	
X	See Page 1, paragraph 1; Page 3, paragraph 14.	1-5
Y		6-9
	US 2003/0046068 A1 (PERRONNIN et al.) 6 March 2003	
X	See Page 1, paragraph 4; Page 3, paragraph 33, 37.	1-5
Y		6-9
	US 5,890,110 A (GERSHO et al.) 30 March 1999	
Y	See Abstract.	6-9
	US 6,571,208 B1 (KUHN et al.) 27 May 2003	
A	See whole document.	
	US 6,460,017 B1 (BUB et al.) 1 October 2002	
A	See Abstract.	

☐ Further documents are listed in the continuation of Box C ☒ See patent family annex

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
07 August 2008	20 AUG 2008
Name and mailing address of the ISA/AU AUSTRALIAN PATENT OFFICE PO BOX 200, WODEN ACT 2606, AUSTRALIA E-mail address: pct@ipaustralia.gov.au Facsimile No. +61 2 6283 7999	Authorized officer **Debasish Dasgupta** AUSTRALIAN PATENT OFFICE (ISO 9001 Quality Certified Service) Telephone No : +61 2 6283 2587

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No.
PCT/SG2008/000213

This Annex lists the known "A" publication level patent family members relating to the patent documents cited in the above-mentioned international search report. The Australian Patent Office is in no way liable for these particulars which are merely given for the purpose of information.

Patent Document Cited in Search Report		Patent Family Member					
US	2002143539	DE	10047724	EP	1193689	JP	2002156993
US	2003046068	US	6895376				
US	5890110		NONE				
US	6571208	CN	1298172	EP	1103952	JP	2001195084
US	6460017	BR	9712979	CN	1230277	CN	1237259
		EP	0925461	EP	0925579	US	6212500
		WO	9811534	WO	9811537		

Due to data integration issues this family listing may not include 10 digit Australian applications filed since May 2001.

END OF ANNEX