

# Enhanced discriminant analysis for confusable sounds via speaker adaptation techniques

Mohsen Rashwan<sup>1</sup>

## Abstract

Speaker adaptation is important to reduce the variability of the speech signal from various speakers. Two approaches are introduced to solve this problem, Vector Quantization (VQ) and Canonical Correlation (CC). Their ability to enhance the discrimination of confused sounds are evaluated through a collected database of Arabic confusable sounds. A recognition accuracy of 73% is obtained without any speaker adaptation technique. When the VQ is applied the accuracy increases to 88%. When the CC Technique is applied the accuracy increases to 93%. This proves the ability of the Canonical Correlation analysis versus VQ.

**Keywords** VQ, Canonical Correlation, DFA, Speaker adaptation, speech recognition.

## 1. Introduction

The speech signal representation that is extracted from an acoustic wave is recorded in varying conditions. These conditions change from speaker to speaker and depend on the *environment* (e.g. microphone, ambient noise, and transmission channel), and the *speaker characteristics* (e.g. physiological differences, articulator language proficiency, and other paralinguistic factors).

All these differences cause an *intra* (Within) and *inter* (Between) speaker variability, therefore a speaker dependent Automatic Speech Recognition (ASR) system has achieved a higher performance level than a speaker independent one because the *intra* speaker variability is considerably less than *inter* speaker variability. The variability of speech due to the *inter* speaker differences result from the differences in the anatomical structure of speech production apparatus (e.g. vocal cord, vocal tract, nasal tract, and stiffness of muscles) or the differences in the manner of moving articulatory organs [1]. The *intra* speaker variability reflects the differences of speech sounds for the same speaker due to the fact that a speaker is never be in the same condition (physical, and mentally).

In general, speaker adaptation procedures should be incorporated in the automatic speech processing systems to improve their performance with new speakers and/or conditions. The objective of this work is to demonstrate the influence of the speaker adaptation techniques on the discrimination of confused Arabic sounds.

This paper is organized as follows: section two introduces Vector Quantization (VQ) based adaptation. Section three introduces the Canonical Correlation (CC) based adaptation. Section four introduces the discrimination criteria applied to the confused phonemes. Experimental results are discussed in section five.

---

<sup>1</sup> Cairo University, Faculty of Engineering, Department of Electronics and Communication, Giza , Egypt.

## 2. Proposed Adaptation approaches

### 2.1 Vector Quantization (VQ) based Adaptation

#### 2.1.1 Vector Quantization (VQ) principles

The objective of Vector Quantization (VQ) is to map a real valued continuous N-dimensional random variable into a discrete random variable with same dimensionality. This process of approximating continuous amplitude signals, by discrete amplitude signals is called quantization. The quantization process is widely applied to data compression or coding applications. Hence, it is clear that the major philosophy of VQ is to reduce both time and space requirements with limited degradation in recognition accuracy [2]. The VQ could be mathematically formulated as follows:

Let  $X = (x_1, x_2, \dots, x_N)^T$  is a real valued continuous random variable.

By partitioning the X vector space into L regions or cells, Y- discrete random variable with the same dimensionality of X- could be defined. I.e. Y is a set of finite set of values.

Let  $Y = (y_1, y_2, \dots, y_L)$  and  $y_i = (y_{i1}, y_{i2}, \dots, y_{iN})^T$

The quantization is achieved by applying a quantization operator  $q$ ; hence the  $x$  value is quantized to  $y$  i.e.

$y_i = q(x)$  if  $x$  vector belongs to  $y_i$  region or cell.

Our objective to minimize the quantization noise. A distortion measure is defined to evaluate the quantization process. A common distortion measure is the squared-error distortion.

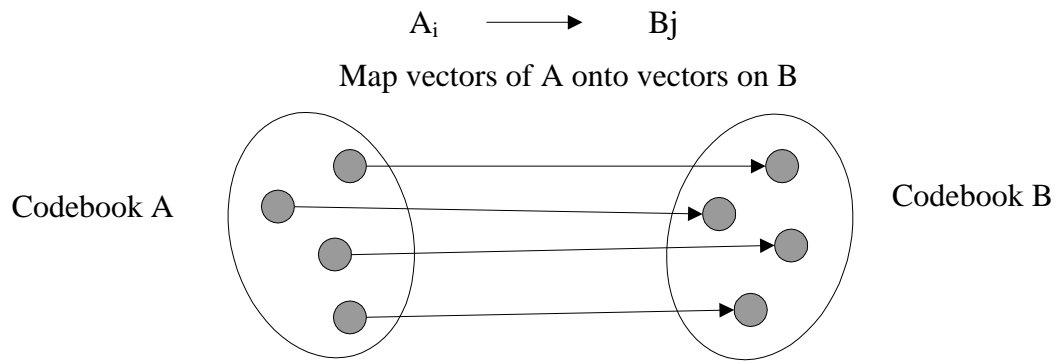
$$d(x, y) = \sum_{i=0}^{N-1} |x_i - y_i|^2$$

The optimal performance of a quantizer is achieved if the average distortion is minimized. Many approaches could be applied to develop a quantizer. Iterative clustering K-mean algorithm is the most common quantizer design algorithm. Also, Self-Organizing Maps (SOM) artificial neural network could be used as a quantizer [10].

#### 2.1.2 Vector Quantization (VQ) based adaptation

Speaker adaptation algorithms based on VQ codebooks have been proposed by Shikano, Lee & Reddy [3], and Goatche & Mason [4]. The speaker adaptation algorithms use two codebooks. One codebook is generated from a speech data of an input speaker and the other codebook is generated from a speech data of a reference speaker. Substituting vectors in one codebook by vectors of another codebook carries out speaker adaptation. This is based on the assumption that a mapping is made from the codebook entries of templates to the codebook entries of the new speaker through dynamic time warping algorithm as shown in Figure (1).

## Substitution table



## Mapped codebook

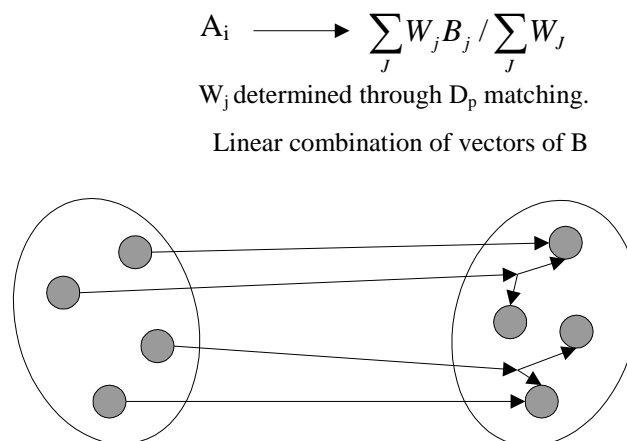


Figure (1) Speaker adaptation Principles using Vector quantization.

## 2. 2 Canonical Correlation (CC) based adaptation

### 2.2.1 Canonical Correlation (CC) principles

The study of the relationship between a set of predictor variables and a set of response measures is known as Canonical Correlation (CC) analysis. CC analysis is a multivariate statistical technique that investigates the relationship between two sets of variables and we seek two linear combinations, one for the predictor set and one for the criterion set, such that their ordinary product moment correlation is as large as possible. The CC could be mathematically formulated as follows:

Let  $\mathbf{m}$  be the number of predictors and  $P$  be the number of criterion variables, and assume that  $\mathbf{m} \geq p$ .

Let  $X^T = (x_1, x_2, \dots, x_m)$  is a  $\mathbf{m}$  dimensional vector of predictor variables.

Let  $Y^T = (y_1, y_2, \dots, y_p)$  is a  $\mathbf{p}$  dimensional vector of criterion measures. Letting  $\mu_x$  and  $\mu_y$  denote the respective mean vector associated with the set variables  $X$  and  $Y$ , the population variance covariance matrices are defined as follows:

$$\begin{aligned}\sum_{xx} &= E\{(x - \mu_x)(x - \mu_x)^T\} \\ \sum_{yy} &= E\{(y - \mu_y)(y - \mu_y)^T\} \\ \sum_{xy} &= E\{(x - \mu_x)(y - \mu_y)^T\} \\ \sum_{yx} &= E\{(y - \mu_y)(x - \mu_x)^T\}\end{aligned}$$

Note that  $\Sigma_{xx}$  and  $\Sigma_{yy}$  are within set variance. Covariance matrices, whereas  $\Sigma_{xy}$  and  $\Sigma_{yx}$  is a between - set covariance matrices. If an  $(\mathbf{m} + \mathbf{p})$  dimensional variable by  $Z = (X, Y)$  is defined, the problem could be viewed in terms of the partitioned variance - covariance matrix  $\Sigma_{zz}$ .

The objective of canonical correlation analysis is to find a linear combination of the  $\mathbf{m}$  predictors that maximally correlates with a linear combination of the Y's. We will denote the respective linear combinations by

$$\hat{X} = a^T X$$

$$\hat{Y} = b^T Y$$

The correlation between X and Y (as function of a and b) is given by

$$\rho(a,b) = \frac{\left( a^T \sum_{xy} b \right)}{\left[ \left( a^T \sum_{xz} a \right) \left( b^T \sum_{yy} b \right) \right]^{1/2}}$$

Out of the infinite number of linear combinations between the X's and the Y's we find that set which maximizes the correlation  $\rho(a,b)$ , this is equivalent to solving the following canonical equations:

$$\left( \sum_{xx}^{-1} \sum_{xy} \sum_{yy}^{-1} \sum_{yx} - \lambda I \right) a = 0$$

and

$$\left( \sum_{yy}^{-1} \sum_{yx} \sum_{xx}^{-1} \sum_{xy} - \lambda I \right) b = 0$$

Where  $I$  is the identity matrix and  $\lambda$  is the largest eigenvalue for the characteristic equations.

$$\left| \begin{array}{ccc} \sum_{xx}^{-1} & \sum_{xy} & \sum_{yy}^{-1} \\ \sum_{xy} & & \sum_{yx}^{-1} \end{array} - \lambda I \right| = 0$$

and

$$\left| \begin{array}{ccc} \sum_{yy}^{-1} & \sum_{yx} & \sum_{xx}^{-1} \\ \sum_{yx} & & \sum_{xy}^{-1} \end{array} - \lambda I \right| = 0$$

This value is the squared canonical correlation coefficient. The eigenvector associated with the eigenvalue  $\lambda$  becomes the vector of coefficients  $a$  and  $b$ . It can be shown that [6].

$$a = \frac{\sum_{xx}^{-1} \sum_{xy} b}{\sqrt{\lambda}}$$

and

$$b = \frac{\sum_{yy}^{-1} \sum_{yx} a}{\sqrt{\lambda}}$$

Which means that it is not necessary to solve for both characteristic equations, since the eigenvectors  $a$  and  $b$  are themselves defined interchangeable.

To summarize, the  $\mathbf{p}$  canonical variates associated with the  $Y$ 's are all uncorrelated with each other, as are the  $\mathbf{m}$  canonical variate for the  $X$ 's, and the correlation between the  $J^{\text{th}}$  canonical variate for the  $X$ 's and the  $K^{\text{th}}$  canonical variate for the  $Y$ 's where  $j \neq k$  is likewise zero.

### 2.2.2 Canonical Correlation (CC) based Adaptation

In this method, adaptation is a way to improve the system for the new speaker by transforming both, the new speaker and the reference speaker features, to a new parametric space where they are supposed to be identical [1]. Hence, The transformation process as shown in figure (2) is an optimal for the new speaker.

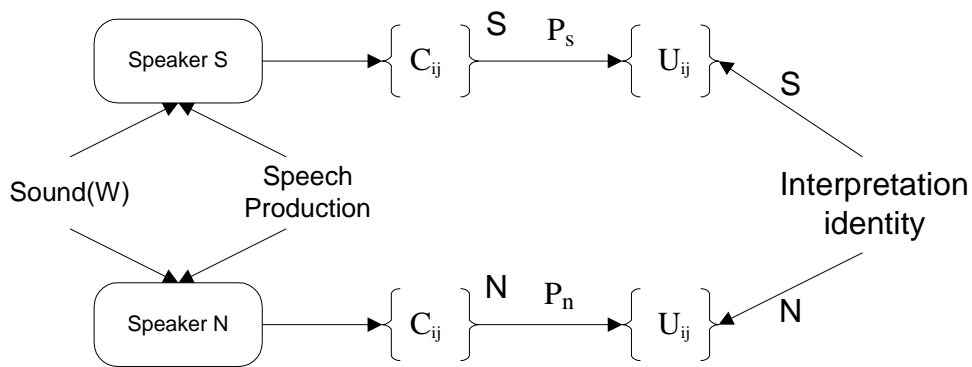


Figure (2) Speaker adaptation Principles using Canonical Correlation.

Therefore, the best linear estimation of the transformations that project the reference and new speaker spectra into a new space, so that corresponding spectra coincide in a least mean square sense is minimized. Canonical Correlation analysis is a technique which allows the construction of the new spectral space and the computation of the two projection operators  $P_s$  and  $P_n$  [5]. The projection of the representations of the two speakers (respectively  $X_i$ ,  $Y_i$ ) is given by:

$$X_i = P(X_i)$$

$$Y_i = P(Y_i)$$

Where  $X_i$  and  $Y_i$  are the feature matrices for the same sound uttered by standard (reference) and the new speaker respectively.

By assuming the distance between projections is minimized, so that error expression  $J$  should be minimized

$$J = \sum_i (x_i - y_i)^T (x_i - y_i)$$

It can be shown that variance of the distance between corresponding spectra is minimum or that corresponding spectra are maximally correlated.

### 3. Linear Discrimination analysis applied to the confusable sounds

Feature evaluation and dimensionality reductions are certainly not new areas of investigation in speech researches. One of the most important tools to do that is the discriminant analysis. It is used to judge how a feature affects the separation between classes. The main objective of the discriminant analysis is to maximize the ratio of the between classes variability to the within classes variability. In general, with K classes and P predictor variables, there are, in total min (P, K-1) possible discriminant axes (i.e., linear composites). In most applications, since the number of predictor variables far exceeds the number of classes under study, at most (K-1) discriminant axes will be considered. However, not all of these axes may show statistically significant variation among the groups, and fewer than (k-1) discriminant function may actually be needed. This is usually done by statistical evaluation on a body of test data [6]. Discriminant analysis mathematical description is summarized as follows [7]:

Definitions:

$x_{ij}$  P- dimension feature vector corresponding to the  $j^{\text{th}}$  sample from the  $i^{\text{th}}$  class of  $n_i$  sample.

$\bar{x}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}$  the  $i^{\text{th}}$  class mean.

$\bar{x} = \frac{1}{k} \sum_{j=1}^k \bar{x}_i$  the overall mean.

$W_i = \frac{1}{n_i} \sum_{j=1}^{n_i} [x_{ij} - \bar{x}_i][x_{ij} - \bar{x}_i]^T$  the covariance matrix of  $i^{\text{th}}$  class.

$W = \frac{1}{k} \sum_{j=1}^k W_i$  the pooled within-class covariance matrix.

$B = \frac{1}{k} \sum_{j=1}^k [\bar{x}_i - \bar{x}][\bar{x}_i - \bar{x}]^T$  the between class covariance matrix.

In terms of the above definitions, the equation to be discriminant analysis is given as follows:

$$[B - \lambda W]b = 0$$

Where  $\lambda$ , b is the eigenvalues and eigenvectors to determined respectively. If W is non-singular matrix, this equation could be transformed into an ordinary eigenvalue problem as follows:

$$[W^{-1}B - \lambda I]b = 0$$

The relative magnitudes of the respective  $\lambda_j$  give a descriptive index of the importance of each discriminant axis.

Using the linear discriminant analysis, classification rule of the linear discriminant function is modified. Linear discriminant function classification score of an object (X) of the P independent variables and K classes is given by:

$$S_i = \left| b^T (X - \bar{X}_i) \right| \quad 1 \leq i \leq K$$

The unknown object X is assigned to a class, which achieve the minimum value of the above score.

## 4. Experimental results

### 4.1 Experiment I

For validation purposes, our experiments concern isolated words (see Appendix A), although the method could be applied to a more general pattern matching problems. The signal acquisition is made in quiet conditions. The marking of beginning and end of words is performed with manual checking. The spectrum is parameterized using Mel Frequency Cepstral Coefficients (MFCC) [8].

A VQ codebook of 64 vectors is generated for each speaker. Also, two CC projection operators, X and Y of cepstral coefficients representing some words pronounced by a standard (reference) speaker, and by a new (test) speaker are needed. A Dynamic Time Warping (DTW) algorithm [9] realizes the one-to-one correspondence and produces two matrices. Hence, the projection operators  $P_s$  and  $P_n$  are computed. Projecting the reference talker dictionary therefore provides the adapted dictionary.

Table (1) illustrates the effect of speaker adaptation using canonical correlation method, from which we indicate that the accuracy of success increases from 70% to 87%.

Case of adaptation	Type of adaptation	
	Canonical correlation	Vector Quantization
Non – adapted	70%	70%
Adapted	87%	79%

Table (1) Results of speaker adaptation techniques.

### 4.2 Experiment II

Also, Table (2) illustrates the effect of speaker adaptation on the discriminant analysis of Arabic confused sounds. As a percentage, the error rates during classification the common confusion sounds. For example the sound (ت), is classified (ط), with error rate 54.7% without adaptation which decreases to 19.6% after adaptation using vector quantization and decreases to 12.1% after adaptation using canonical correlation. Also the sound (ط) is classified as (ت) with error rate 46.7% without adaptation which decreases to 24.2% after adaptation using vector quantization and so on. The average error rate of the discriminant analysis before speaker adaptation is 27%, which decreases to 7% after speaker adaptation via CC and to 18% after speaker adaptation via VQ.

	ت	ط
ت	-	54.7
ط	46.7	-

	ت	ط
ت	-	19.6
ط	24.2	-

	ت	ط
ت	-	12.1
ط	16.2	-

	ق	ك
ق	-	29.3
ك	16.1	-

	ق	ك
ق	-	17.4
ك	10.6	-

	ق	ك
ق	-	9.3
ك	5.2	-

	د	ض
د	-	33.3
ض	26.7	-

	د	ض
د	-	19.7
ض	13.6	-

	د	ض
د	-	10.7
ض	2.7	-

	ث	س	ص
ث	-	21.8	2.3
س	16.7	-	18.7
ص	1.3	17.3	-

	ث	س	ص
ث	-	10.6	1.4
س	7.3	-	9.7
ص	1.1	9.8	-

	ث	س	ص
ث	-	6.67	0.52
س	3.3	-	1.3
ص	0.6	1.3	-

	ذ	ز	ظ
ذ	-	20.2	23.4
ز	18.7	-	21.6
ظ	21.6	19.7	-

	ذ	ز	ظ
ذ	-	10.2	13.4
ز	14.2	-	11.3
ظ	12.7	8.4	-

	ذ	ز	ظ
ذ	-	8.7	9.1
ز	7.7	-	8.3
ظ	8.9	7.2	-

A

B

C

Table (2) speaker adaptation effect on the performance of discriminant analysis.

Where:

A - Discriminant analysis without speaker adaptation

B - Discriminant analysis VQ based adaptation.

C - Discriminant analysis CC based adaptation.

## 5. Conclusions

Two approaches are introduced in order to overcome the speaker adaptation problem, Vector Quantization (VQ) and the Canonical Correlation (CC) analysis. Arabic language contains some sounds where there is a common confusion in their pronunciation such as (ض، - (ظ، ز، ذ) - (ك، ق) (ط، ت) - (ص، س، ث) - (د). A database of Arabic confused sounds is collected. A discrimination accuracy of 73% is obtained without any speaker adaptation technique. When the VQ is applied the accuracy is increased to 88%. When the CC technique is applied, the accuracy is increased to 93 %.

## References

- [1] K. Choukri and G. Chollet, "Adaptation of automatic recognizers to new speakers using Canonical correlation analysis techniques", *Comp. Speech and Language*, Vol. 1, PP. [95-107].
- [2] R. M. Gray, "Vector Quantization", *IEEE ASSP Magazine*, April 1984.
- [3] K. Shikano, K. F. Lee, and R. Reddy, "Speaker Adaptation through vector quantization", *Proc. ICASSP*, PP. [2643-2646], 1986.
- [4] J. K. Goatcher and J.S. Mason, "An adaptive approach to a speaker independent isolated word system with short training", *Int. Con. Speech Inp./Out. Tech and Application*, PP. [67-70], 1986.
- [5] K. Choukri, G. Chollet, and Y. Grenier, "Spectral transformations through Canonical Correlation Analysis for speaker adaptation in ASR.", *Proc. ICASSP*, PP. [2659-2662], 1986.
- [6] W. R. Dillon, and M. Goldstein, "Multivariate analysis method and applications", *Jon Wiley & Sons*, New York, 1984.
- [7] W. S. Mohn, "Two statistical Feature evaluation techniques applied to speaker identification", *IEEE trans. Computers*, Vol. C-20, No. 9, PP[979-987], Sep. 1971.
- [8] M. A. Rashwan, "new Approaches for Isolated word Recognition". Ph.D. thesis, Queen's Univ., Canada, 1987.
- [9] C. Myers, L.R. Rabiner, et al, "Performance Trade-offs in Dynamic Time warping algorithm for Isolated Word Recognition", *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. ASSP-28, No. 6, PP. [623-632], Dec. 1980.
- [10] T. Kohonen, "Self-Organizing Maps", Second Extended Edition, *Springer Series in Information Sciences*, Vol. 30, Springer, Berlin, Heidelberg, New York, 1995, 1997.

## Appendix A

### Arabic words confusable sounds database

The words used for training and testing the discriminant analysis and speaker adaptation algorithms.

Word		Confused word	
Arabic	English Pronunciation	Arabic	English Pronunciation
ثمر	THMR	سمر	SMR
ثقل	THQL	صقل	SSQL
سال	SAL	صال	SSAL
سام	SAM	صام	SSAM
سار	SAR	صار	SSAR
تل	TL	ظل	TTL
تاب	TAB	طاب	TTAB
دم	DM	ضم	DDM
دع	DAA	ضع	DDAA
دال	DAL	ضال	DDAL
ذل	ZL	ظل	ZZZL
زن	ZZN	ظن	ZZZN
كد	KD	قد	QD
كل	KL	قل	QL
كال	KAL	قال	QAL