Journal of Academic and Applied Studies
Vol. 2(3) March 2012, pp. 32-42
Available online @ www.academians.org

ISSN1925-931X

The Reduction of Speech Characteristic Vector Using PSO Algorithm and the Evaluation of the Effectiveness of Different Speech Characteristic in Recognition of Persian Language State

Nasim Chagha Ghasemi, Khosro Rezaee

Department of Electronic Engineering, Islamic Azad University-South Tehran Branch , Tehran, Iran
Department of Biomedical Engineering, hakim sabzevari University-sabzevar, Iran

Abstract

Speech has a number of characteristic, the extraction of which can play an important role in the accuracy of speech recognition. In this regard, many researchers have attempted to investigate these features and provide methods which enhance the recognition and identification of speech states. These features include MFCC coefficients, energy, Formant Frequency and Pitch Frequency which are highly important in speech state recognition system. This paper explores the effect of these features on speech state recognition and four different states, i.e. angry, happy, natural and question will be tested. The study investigates a variety of speech characteristics in form of a vector contains 55-characteristics. In the next step, drawing on PSO optimization algorithm, 49, 24 and 15-characteristic vectors are achieved. The less the characteristics of a vector are, the higher the action velocity will be. After that, the mean Normalization, Cepstral variance and Cepstral gain methods are applied on these vectors and using GMM algorithm, speech state recognition is executed on normalized vectors. Finally, following the normalization of the output vectors and speech state recognition through GMM algorithm, the effect of different speech characteristics as well as different normalization methods on speech state recognition are examined.

Keywords: Speech recognition, PSO algorithms, Normalization, GMM model, Speech characteristics

I. Introduction

Languages differ from one community to another and understanding one language requires special training and education. Despite the numerous commonalities between languages, speech processing follows a relatively consistent method in many languages. Recently, there has been a growing research interest in speech recognition. This article investigates the effect of different speech characteristics and various normalization methods on speech state recognition, a subject which has not been adequately addressed in academic researches.

First, using PSO algorithm optimization, we will attempt to reduce the characteristic vectors to accelerate the process. PSO method was proposed by Eberhart and Kendedy in 1995(Kennedy and Elberhart, 1995), inspired by the behavior of fish or birds swarms on food search. Then, the resulted small vectors are normalized and subjected to state recognition test using GMM model

(Eslami and Sayadian, 2006) and the effect of removing different characteristics in speech state is examined.

Normalization strategy is used in speech recognition system to offset the effect of environmental incompatibility. Normalization techniques are very helpful in reducing the effects of certain noises in noisy environments. Characteristic normalization seeks to normalize characteristics of speech vectors which entail an application of such statistical features as mean, variance and gain to reduce the inconsistency of characteristic vector. Cepstral Mean Normalization (CMS) technique was first used by Atal, B.S in 1974 (Atal, 1974). This technique is utilized to remove the average characteristic speech in order to reduce the complex interference effects.

In 1998, O. Viikki and K. Laurila proposed Cepstral Mean Variance Normalization (CMVN), which normalizes both mean characteristics and variance (Viikki and Laurila, 1998). In 2004, Yoshizawa et al presented Cepstral Normalization Gain (CGN) (Yoshizawa et al., 2004). In 2006, Hilger, F, and Ney, H. proposed the Histogram Equalization (HEQ) to compensate for the nonlinear distortion of environment (Hilger and Ney, 2006). The aim of this method was to find a conversion which could convert the distribution of each component of speech characteristic vector to a predetermined distribution compatible with training speech. In 2006, the speech characteristic temporal structure normalization (TA) was presented by Shih–Hsiang et al. This method was applied to the components of characteristic vector to reduce the effect of the sharp peaks and valleys created due to non-static noises (Hsiang et al., 2006).

It should be noted that a large number of the proposed normalization methods have already been tested in speech recognition. This article, however, seeks to enhance speech recognition. To this purpose, Cepstral Mean, Variance and Gain Normalization methods were used for recognition of Persian language states. The results indicate the effectiveness of this method in enhancing the normalization of some speech states.

II. Methods

In this study, 24 different speakers were asked to state some sentences in Persian in four states: angry, happy, natural and question. After the extraction of such characteristics of their speech as the MFCC coefficients, energy and Formant Pitch and Frequencies, each 10ms of the statements was expressed as a frame, which given the average time of each sentence, i.e. 2 seconds, amounts to 200 frames.

Then, a vector with 55-speech characteristics was resulted and speech recognition state was tested using GMM model (Reynolds and Rose, 1995), (Styliaou et al., 1998). Gaussian Mixture Model (GMM) is an absolutely viable method to show the speaker's speech characteristics. It is a mixture of Gaussian’s models in which Gaussian’s peaks in the density range are the point where vectors associated with a particular sound are clustered. In this method, first the parameters of GMM transfer function at the learning stage are estimated and then the speech processing is performed. In Gaussian model, the aim is to estimate parameters using existing educational data to achieve the best harmony in speaker’s characteristic vectors. At that point, the model recognition is performed.

In the first section, 55-charateristic vector is analyzed with each vector containing the following characteristics:

$c_1, c_2, ...c_{12}$ of MFCC coefficients, Energy (E); $c_1^{'}, c_2^{'}, ...c_{12}^{'}$ of the first derivative of MFCC coefficients; the first derivative of energy ($E^{'}$), $c_1^{''}, c_2^{''}, ...c_{12}^{''}$ of the second derivative of MFCC coefficients; the second derivative of energy ($E^{''}$); the first Formant frequency (F$_1$); the second Formant frequency (F$_2$); the third Formant frequency (F$_3$); the first derivative of the first Formant frequency ($F_1^{'}$); the first derivative of the second Formant frequency ($F_2^{'}$); the first derivative of the third Formant frequency ($F_3^{'}$); the logarithm of first Formant frequency ($\log F_1$); the logarithm of second Formant frequency ($\log F_2$); the logarithm of third Formant frequency ($\log F_3$); the first Formant frequency normalized using Cepstral Mean normalization method (zF_1); the second Formant frequency normalized using Cepstral Mean normalization method (zF_2); the third Formant frequency normalized using Cepstral Mean normalization method (zF_3); Pitch Frequency (F_0); the logarithm of Pitch Frequency ($\log F_0$); the first derivative of Pitch Frequency ($F_0^{'}$) and the normalized Pitch Frequency using Cepstral Mean normalization method (zF_0).

In 55-speech characteristic vector, first, speech state recognition test is conducted. Table I shows the results of this recognition.

TABLE I
Results of state recognition (55-speech characteristic vector)

Speech State	Percent of recognition
Angry	68.55%
Happy	67.00%
Natural	47.77%
Question	66.70%

To accelerate the speech state recognition and remove less-effective characteristics, the size of speech characteristic vectors are reduced using PSO optimization algorithm (Hegashi and Iba, 2003),(Kennedy and Eberhart, 1993) .

The collective movement of particles (PSO) is a potential optimization technique which is based on population. This method has been inspired by the collective behavior of fish or birds on food search. In PSO algorithm, each solution, which is called a particle, resembles a bird in a bird swarm. Each particle has a fitness value which is calculated by a fitness function. The closer is each particle in search space to the target food in birds swarm model, the more its fitness value will be. Each particle has a certain velocity which helps its conduction. Following the optimal particles in the current state, each particle continues its movement in the problem space. At the outset, PSO is a group of particles (solutions) which have been formed haphazardly and seek the optimal solution by upgrading generations. In the next section, we will evaluate characteristic vectors using the above optimization algorithm.

A. Evaluation of 49-characteristic vector

In this section, an evaluation of 49- speech characteristic vector is made using the PSO algorithm. This vector contains the following properties:
$c_1, c_2, ... c_{12}$ MFCC coefficients, Energy (E); $c_1^{'}, c_2^{'}, ... c_{12}^{'}$ the first derivative of MFCC coefficients; the first derivative of energy ($E^{'}$), $c_1^{''}, c_2^{''}, ... c_{12}^{''}$ the second derivative of MFCC coefficients; the second derivative of energy ($E^{''}$); the first Formant frequency (F_1); the second Formant frequency (F_2); the third Formant frequency (F_3); the logarithm of first Formant frequency ($\log F_1$); the logarithm of second Formant frequency ($\log F_2$); the logarithm of third Formant frequency ($\log F_3$); the first Formant frequency normalized using Cepstral Mean normalization method (zF_1); Pitch frequency (F_0); Pitch frequency logarithm ($\log F_0$); the first derivative of Pitch frequency ($F_0^{'}$).
Table II presents the results of state recognition for this vector.

TABLE II
Results of state recognition (45-speech characteristic vector)

Speech State	Percent of recognition
Angry	69.04%
Happy	64.70%
Natural	52.89%
Question	61.87%

O the whole, the results of the above table suggest that the average recognition of 49-vector is 62.13 % which shows an insignificant change (0.38%) compared to 62.51% of 55-vector. It means PSO algorithm has been successful in removal of less-effective characteristics. The comparison of the results of 55-vector and 45-vector suggests that the removal of $F_1^{'}$ ، $F_2^{'}$ ، $F_3^{'}$ ، zF_2 ، zF_3 and zF_0 characteristics increases the accuracy of recognition in angry and natural states. With the removal of above characteristics, the recognition of angry and natural states increases by 0.49 and 5.12 percent respectively. The removal of these characteristics, obviously, reduces the recognition percentage of happy and question states. Given the fact that in this section only a few characteristics were removed, in what follows, we will attempt to remove a larger number of less-effective characteristics using PSO algorithm.

B. 24-characteristic vector

The 24- speech characteristic vector is composed of the following characteristics:
$c_1, c_2, c_3, c_4, c_6, c_8, c_{11}$ MFCC coefficients, Energy (E); $c_{12}^{'}$ the first derivative of MFCC coefficients; the first derivative of energy ($E^{'}$), $c_1^{''}, c_{12}^{''}$ the second derivative of MFCC coefficients; the second derivative of MFCC coefficients; the first Formant frequency (F_1); the second Formant frequency (F_2); the third Formant frequency (F_3); the first derivative of third Formant frequency $F_3^{'}$;

logarithm of the first Formant frequency ($\log F_1$);logarithm of the second Formant frequency ($\log F_2$); logarithm of the third Formant frequency ($\log F_3$); the first Formant frequency normalized using Cepstral Mean normalization method (zF_1); the second Formant frequency normalized using Cepstral Mean normalization method (zF_2); the third Formant frequency normalized using Cepstral Mean normalization method zF_3; Pitch frequency (F_0); Pitch frequency logarithm ($\log F_0$); the first derivative of Pitch frequency ($F_0^{'}$).
Table III presents the results of state recognition for this vector.

TABLE III
Results of state recognition (24-speech characteristic vector)

Speech State	Percent of recognition
Angry	69.56%
Happy	62.48%
Natural	44.37%
Question	61.79%

The averaging of above vector results yields 59.55% recognition which is 3% less than 55-vector. This decrease has been mainly due to decline of happy and question states. The recognition of angry state using 24-speech characteristic vector shows 0.52% and 1% rise compared to the results of 45 and 55 vectors respectively, suggesting the negative effect of the removed characteristics on recognition of this speech state.
Compared to 49-vector, no improvement was observed in other states of speech recognition suggesting that the removal of MFCC-related characteristics (speech characteristics have not been removed in 45-vector) has left a negative effect on the recognition of happy and natural states, though with a lesser degree on question state. In the next section, we will use PSO algorithm to obtain 15-vector.

C. Evaluation of 15- characteristics vector

The features of 15- characteristic vector in this algorithm are:
$c_1, c_2, c_3, c_6, c_{11}$ Coefficients MFCC, energy (E), the first Formant frequency (F_1), the second Formant frequency(F_2), the third Formant frequency (F_3), the first Formant Frequency logarithm ($\log F_1$), The second Formant frequency logarithm ($\log F_2$), the third Formant frequency logarithm ($\log F_3$), the first Formant Frequency normalized using Cepstral Mean Normalization (zF_1), Pitch frequency (F_0), the first derivative of Pitch Frequency ($F_0^{'}$).
Table IV presents the results of state recognition using 15-charqcteristics vector.

Table IV
Results of state recognition (15-speech characteristic vector)

Speech State	Percent of state recognition
Angry	64.77%
Happy	58.56%
Natural	47.17%
Question	49.90%

The average percentage of this vector is 55.1%, which compared to other vectors, displays a greater decrease, though it is mainly due to the results of happy and question states. On the other hand, the comparison of the results of state recognition for 15 and 55-vectors reveals that PSO algorithm has been significantly successful in natural and angry states. The comparison of the results drawn between state recognition and 24-vector shows that with the removal of MFCC-related characteristics, the recognition percentage decreases. The first section of this paper dealt with the evaluation of different speech characteristics and their recognition.

At the end of this section, the results of analyzed vectors have been shown in Figure 1. It should be noted that in this diagram, states 1 to 4 are respectively related to angry, happy, natural and question states.

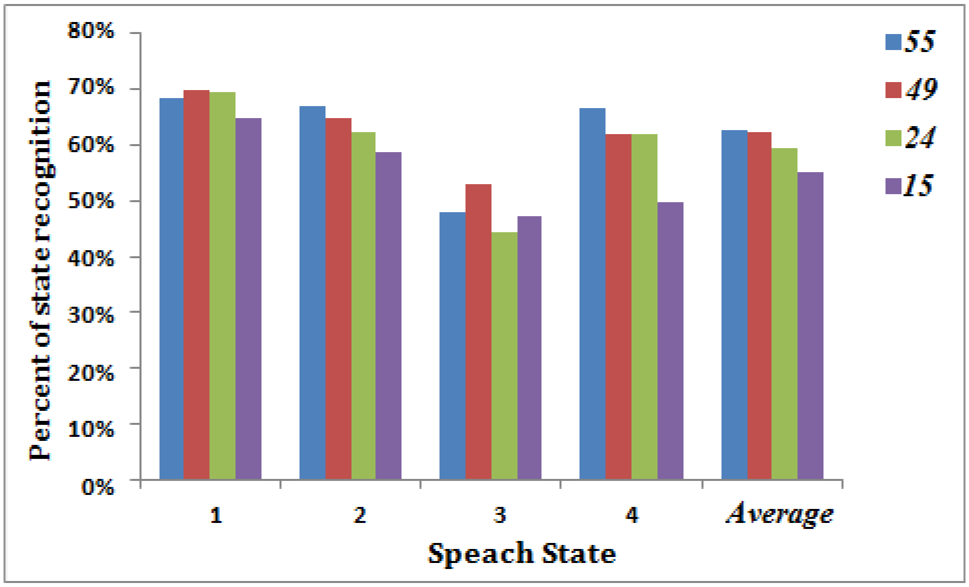

Figure. 1. The results derived from recognition of different characteristic vectors

In the next section of this paper, we will analyze the different speech characteristic normalization methods.

D. Normalization methods

D.a Cepstral Mean Normalization (CMN)

Cepstral Mean Normalization is one of common normalization which is particularly effective in reducing the channel disturbance (Furui, 1981), (Liu et al., 1994). In this algorithm, first mean is calculated for characteristic vectors, then it is subtracted from each characteristic as follows:
For a set containing T Cepstral vector x=$\left\{ x_1, x_2, ...x_T \right\}$, $\overline{x}$ is calculated as follows:

$$\overline{x}_T = E\left[x\right] = \frac{1}{T}\sum_{\tau=1}^{T} x_\tau \qquad (1)$$

The deletion of Cepstral Mean is not a rapid and easy reductive operation. It is carried out as follows:

$$\overline{x}_T = x_T - \overline{x}_T \qquad (2)$$
$$y_T = x_T + h_T \qquad (3)$$

In which $\overline{x}_T$,%x_T,h_T ,x_T ,y_T and $\overline{x}_T$ are respectively Cepstra noise, Clean Speech Cepstra, Filter Response, Speech Cepstra and Cepstra Mean.

$$\overline{y}_\tau = \frac{1}{T}\sum_{\tau=1}^{T}\left(x_\tau + h_T\right) = \overline{x}_\tau + h_\tau \qquad (4)$$

Equation (4) shows that Cepstral Mean Normalization is a linear line in terms of channel distribution.
A noticeable improvement in the accuracy of speech state recognition even in clean condition (without any noise) is the main advantage of this method (Yapanel et al., 1994).

D.b Cepstral Variance Normalization (CVN)

Another method used for normalization of speech characteristic in this study was a method known as Cepstral Variance Normalization. A common technique in speech recognition, this method is highly effective, especially in environments with increasing noise (Torre et al., 2002). The mean and variance of Cepstral coefficients have been included in this analysis.
Variance is a scale for statistical dispersion around mean distribution which is the result of mathematical expectation calculated as follows:

$$\sigma^2 = E\left[\left(x - E\left[x\right]\right)^2\right] = E\left[x^2\right] \qquad (5)$$

By delimiting Cepstral variance limit, Cepstral Variance Normalization considerably affects speech recognition (Liu et al., 1993) as follows:

$$x_{CVN} = \frac{x - \overline{x}}{\sigma} = \frac{x_{CMN}}{\sigma} \qquad (6)$$

D.c Cepstral Gain Normalization (CGN)

Cepstral Gain Normalization was another method used in this study. It is used to remove the effects of mismatch between testing and training environments, which is carried out by adjusting the gain
and removing Cepstral DC offset. With Cepstral Mean Normalization, DC offset is removed and Gain Normalization is then used to normalize noisy and clean Cepstral Gain to unit 1.
Cepstral Gain Normalization is defined by x_{CGN}

$$x_{CGN} = \frac{x_{CMN}}{\max\{x_{CMN}\} - \min\{x_{CMN}\}} \tag{7}$$

D.d Evaluation of various normalization methods

In this section, we will examine the normalization of 55 and 24 characteristic vector. Table V presents the results of state recognition in 55-characteristic vector after the application of the above normalization method.

Table V
Results of state recognition after normalization (55-speech characteristic vector)

Speech State	Percent of recognition		
	CMN	CVN	CGN
Angry	69.00%	68.00%	63.49%
Happy	50.40%	44.60%	46.27%
Natural	57.04%	57.70%	56.34%
Question	53.00%	55.90%	65.13%

The results suggest that Cepstral Mean Normalization method improve recognition in angry and natural state by 0.45 and 9.27 percent respectively. Cepstral Variance Normalization and Cepstral Gain Normalization also displays respectively 9.93% and 8.57 improvement in speech state recognition suggesting that Cepstral Variance Normalization would be a better method for natural state recognition while Cepstral Mean Normalization would be a better choice for increasing the accuracy of angry state recognition. The effectiveness of normalization methods in speech state recognition has been substantiated.
Figure (2) presents the results of different normalization methods for 55-charactersitic vector.

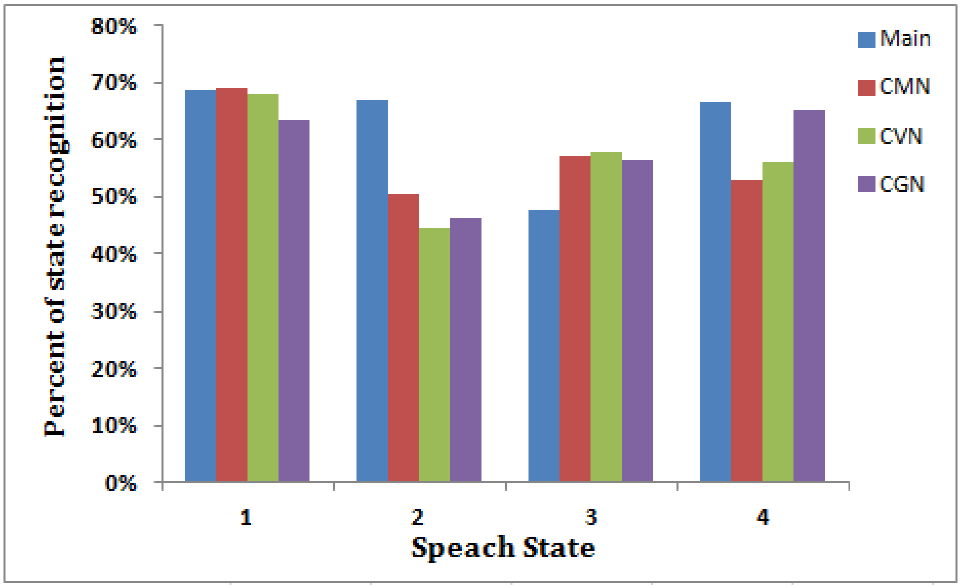

Figure. 2. State recognition results for different normalization methods (55-speech characteristic vector)

Table VI displays the results of state recognition for 24-speech characteristic vector after normalization.

Table VI
Results of state recognition after normalization (24-speech characteristic vector)

Speech State	Percent of recognition		
	CMN	CVN	CGN
Angry	42.85%	61.00%	60.1%
Happy	32.14%	40.48%	35.30%
Natural	61.70%	51.75%	50.00%
Question	55.32%	50.69%	56.80%

The results suggest that the above normalization methods have only been effective in natural state with CMN displaying the best performance by 17.33 % improvement in natural state recognition.

In other methods, normalization of 24-vector is not effective, suggesting the negative effect of Formant and Pitch normalization relative to MFCC coefficients.

Figure 3 presents the results of different normalization methods for 24-vector.

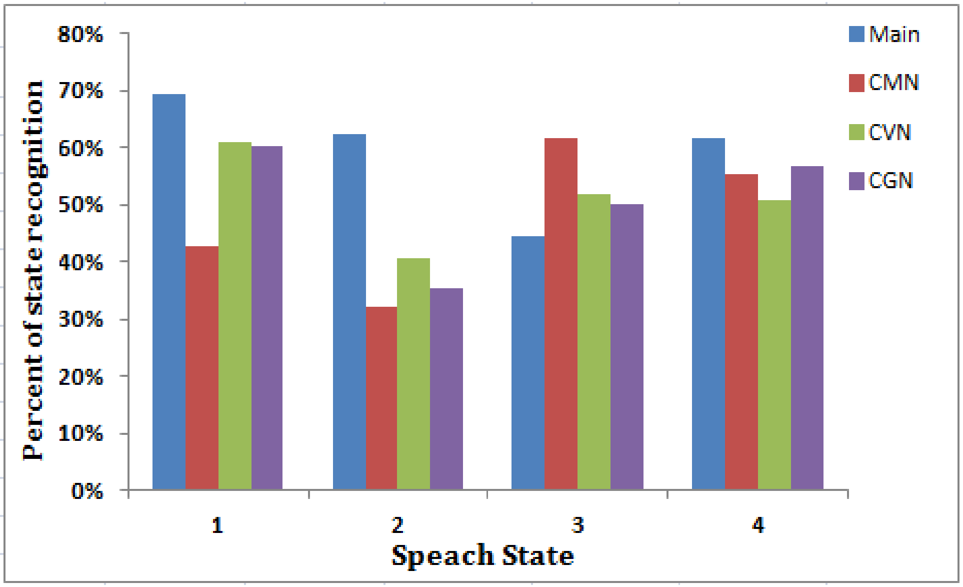

Figure. 3. State recognition results for different normalization methods (24-speech characteristic vector)

III. Conclusions

In this study, we examined the speech recognition state in Persian language. First, speech characteristics of 24 speakers in such different states as angry, happy, natural, and question were extracted and then the recognition was conducted using the GMM model. In the next step, using PSO optimization technique, less-effective speech characteristics were removed and recognition was carried out again. The results reveal the effectiveness of the above optimization method in removal of less-effective characteristics which can help the recognition speed and extraction of these characteristics. Last section addresses the effect of different normalization methods on speech state recognition for two characteristic vectors. The results suggest that the different characteristics might leave a positive or negative effect on recognition after normalization. Overall, it is better not to apply normalization to some characteristics. Furthermore, the results indicate that the above methods can improve the accuracy of speech recognition.

Acknowledgements

Thanks from Telecommunication Infrastructure Company, Iran, Kermanshah and Comkar System Company.

References

Atal B.S (June 1974). Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. J. Acoustic Soc. Amer, vol (55), pp.1304-1312.

Torre A de la, Segura J.C, and Benitez C, etc (2002). Non-linear Transformation of the Feature Space for Robust Speech Recognition. Proc. ICASSP'02, pp. 401-404.

Eslami M and Sayadian A (winter 2006). Quality improvement of speech conversion systems based on GMM. Journal of Tarbiat Modares University. Vol (22), pp. 23-36.

Furui S (April 1981). Cepstral analysis technique for automatic speaker verification. IEEE Trans. Acoustics., Speech, Signal process., vol (ASSP-29), no.2, pp.254-272.

Hagashi N and Iba H (2003). Particle Swarm Optimization with Gaussian Mutation. Proceeding of the IEEE swarm intelligence symposium, Indianapolis, Indiana, USA, pp. 72-79.

Hilger F and Ney H (2006). Quantile based histogram equalization for noise robust large vocabulary speech recognition. IEEE Trans. On Audio, Speech and Language Processing, 14(3), pp. 845-854.

Hsiang Sh, Ming Yeh Y and Berlin Chen (2006). Exploiting polynomial – fit histogram equalization and temporal average for robust speech recognition. Interspeech.

Kennedy J and Eberhart R. C (Mar 1993). A Discrete Binary Version of the Particle Swarm Algorithm. Proceedings of the International Confrance on Systems, Man, and Cybernetics, IEEE Service Center, Piscataway, NJ, pp. 69-74.

Kennedy J and Eberhart R. C (1995). Particle Swarm Optimization. Proceedings of IEEE International Confrance on Neural Networks, Piscateway, NJ, pp.1942-1948.

Liu F.-H, Stern R.M, Acero A and Moreno P.J (1994). Environment normalization for robust speech recognition using direct cepstral comparison. ICASSP94, vol (2), pp.61-64.

Liu F, Stern R, Huang X, and Acero A (Mar 1993). Efficient cepstral normalization for robust speech recognition. in Proc. ARPA Speech Natural Language Workshop, Princenton, NJ, pp. 69-74.

Reynolds D. A and Rose R.C (1995). Robust text independent Speaker identification using Gaussian mixture Speaker models. IEEE Trance. On Speech, Audio Processing, vol (3), pp. 72-83.

Styliaou Y, Cappe O and Moulines E (1998). Continuous Probabilistic Transform for Voice Conversion. Speech Communication. Vol (24), No. 2, pp. 192-200.

Viikki O and Laurila K (1998). Cepstral domain segmental feature vector normalization for noise robust speech recognition. Speech Communication, vol (25), pp. 133-147.

Yapanel U, Zhang X, and Hansen J (Sep 2002). High performance digit recognition in real car environment. In Proc. ICSLP, Denver, CO, pp. 793-796.

Yoshizawa S, Hayasaka N, Wada N, and Miyanaga Y (2004). Cepstral gain normalization for noise robust speech recognition. In proceeding of 2004 International Conference on Acoustics, Speech and Signal Processing (ICASSP 2004), I- 209- 212.