The CSI Journal on
Computer Science and Engineering
Vol. 1, No. 4 (b), Winter 2003
Pages 31-38
Reguler Paper

Robust Speech Recognition Based on Mapping Noisy Features to Clean Features

Babak Nasersharif Mohsen Rahmani Ahmad Akbari

Computer Engineering Department, Iran University of Science and Technology,
Tehran, Iran

Abstract

The conventional view on the problem of robustness in speech recognition is that performance degradation in ASR systems is due to mismatch between training and test conditions. If problem of robustness in ASR systems is considered as a mismatch between the training and testing conditions the solution would be to find a way to reduce it. Common approaches are: Data-Driven methods such as speech signal enhancement and using robust features and model-based methods that alter or adapt model of speech signal. In this paper, we study a model of environment and obtain a relation between noisy and clean speech features based on this model. We propose two techniques for mapping noisy features to clean features in cepstrum domain. We implement the proposed methods and some of precedent data-driven methods such as: spectral subtraction, cepstral mean normalization, cepstral mean and variance mean normalization and SNR-dependent cepstral normalization .We show that proposed methods outperform precedent methods and are effective for robust speech recognition in noisy environments.

Keywords: noise, robustness, neural network, map, data-driven methods, speech recognition

1. Introduction

The conventional view on the problem of robustness in speech recognition is that performance degradation in ASR systems is due to the differences between speech signal they receive on input (when employed in real life applications) and the speech signal used for training and estimation of parameters of their models during system construction. This is commonly refered to as mismatch between training and test conditions. Some common reasons for mismatch between training and testing speech signal are considered to be: contamination of signal with noise (additive, convolutional, reverberation), speaking style (Lombard effect, speaking rate) and inter speaker variations (voice quality, pitch, gender).

If problem of robustness in ASR systems against contamination with noise, were considered as a mismatch between the training and testing conditions the solution would be to find a way to reduce it. Common approaches are: speech signal enhancement, using robust features, using microphone arrays, using hearing properties of human ear and model alteration/adaptation.

The suggested techniques for description of effects of noisy environment can be divided in two categories: data-driven methods and model-based methods. Data-driven methods try to describe effects of environment on speech and speech features and enhance speech signal or improve its features. Some of these methods require both noisy and clean signals. Model-based methods try to change statistical model of environment so that it adapts to new properties of environment. In these methods, the observation is not changed and there is not any assumption or change for speech signal.

Model-based methods modify acoustic models instead of speech signal or its features. This has the advantage that no decisions or hypotheses about speech are necessary and observed data is unaltered. These methods don't require stereo database. Some examples of these approaches are: Hidden Markov Model decomposition [7,14,15], parallel model combination (PMC) and maximum likelihood regression (MLLR) [2,3,8].

In this paper, we limit ourselves to data-driven methods. After reviewing basic methods, we evaluate performance of them. In second section,
a model for speech recording at the input of an ASR device is discussed. This model considers acoustic noise and transmission channel effects and can explain the relation among different approaches of robust ASR. The third section is dedicated to discussing data-driven methods. Section four includes our proposed method. Our experiments and results are explained in section five. In section six, we give our conclusion.

2. A Model of Environment

In most speech recognition applications, there are two types of noise: additive noise and linear filter noise. Additive noise includes background sounds, effects of air flow and unwanted signals captured by microphone. Linear filter noise includes effects of microphone or transmission channel and reverberation. If we restrict ourselves to additive and channel noise, we can construct a model such as Figure (1), for effects of environment [1,3].

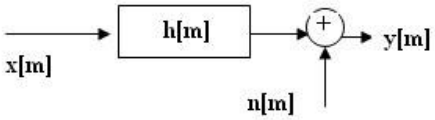


Figure 1. Model of environment

In this model, x[m] is a clean speech signal. It is filtered by a linear and time invariant filter h[m]. Output of this filter is added to the unknown noise n[m] that represents additive part of environment noise. This model was originally proposed by Acero [1] and later used by Moreno [3] and Gales [2]. It is a reasonable model and we use it in this study. In this model, n[m] and x[m] are uncorrelated and we assume n[m] is stationary and is colored in general. With these assumptions, we can obtain an expression that relates power spectral density (PSD) of output signal $P_y(w)$ to PSD of noise $P_n(w)$, transfer function amplitude of filter $|H(w)|$ and PSD of input signal $P_x(w)$:

$$P_y(w) = P_x(w)\ |H(w)|^2 + P_n(w) \tag{1}$$

According to [3], after transforming to log spectral domain, relation (1) is converted to:

$$y = x + h + \ln(1 + e^{n-x-h}) \tag{2}$$

Where y, h, n, x represent the log spectrum of y[m], h[m], n[m] and x[m] respectively. Equation 2 can be rewritten as:

$$y = x + f(x, n, h) \tag{3}$$

Where *f* can be written as:

$$f(x, n, h) = h + \ln(1 + e^{n-x-h}) \tag{4}$$

f( ) is a complex function and not so easy to compute. In effect, the noise and channel features are unknown. In following sections, we review different approaches proposed for estimating clean features from noisy features. Finally we propose two methods based on above model.

3. Background

Data-driven methods determine effects of environment on properties of speech signal. They try to improve these effects such that speech is recognized in an environment independent manner. These methods can be divided into different groups. The first group acts on noisy speech signal directly and estimates clean signal from noisy signal. Spectral subtraction and signal filtering are placed in this group [17].
The second group is based on finding and extracting robust features. This group tries to remove effects of noise in features and reduce mismatch of training and testing data. CMS[a], PLP[b] and RASTA PLP [4,5] and ZCPA[c] [6] analysis are examples of these methods.
Another group finds a map between noisy signal features and clean signal features. For estimating such mapping, we must have access to both clean and noisy signals. There are various mapping: Linear regression methods [18], nonlinear estimators such as MLP or other neural networks [18] and minimizing a cost function by MMSE criterion.
Methods of pervious groups are also combined to obtain more effective techniques. Properties of human ear are also used to improve robustness. Noise masking technique is an example of such methods [2,7].

3.1 Spectral subtraction

Spectral subtraction is a well-known method that has a long tradition in research in speech enhancement. In this technique, an estimate of the noise spectrum $\hat{P}_n(w)$ is computed and it is subtracted from noisy input spectrum $P_y(w)$, to obtain an estimation of clean speech spectrum $\hat{P}_x(w)$.

$$\hat{P}_x(w) = P_y(w) - \hat{P}_n(w) \tag{5}$$

As long as ways can be found to reliably estimate background noise spectrum, spectral subtraction is useful and can be used as a preprocessing step for ASR systems, to make them robust against quasi time-invariant noise. The general relation for spectral subtraction is given in equation (6). a is an overestimation parameter, slightly greater than 1, that controls the level of subtracted noise spectrum, while b is a noise flooring parameter normally near zero.

$$\hat{P}_x(w) = \begin{cases} P_y(w) - a\,\hat{P}_n(w) & \text{if } P_y(w) > (a+b)P_n(w) \\ b\,\hat{P}_n(w) & \text{otherwise} \end{cases} \tag{6}$$

Many different methods are developed using spectral subtraction. Adaptive spectral subtraction [16] and nonlinear spectral subtraction [20] and a-iterative spectral subtraction [21] are examples of such methods.
In speech recognition applications, features are extracted from estimated clean signal as observation vectors.

3.2 Cepstral mean subtraction

Cepstral Mean Subtraction (CMS), which belongs to second group, is perhaps one of the most effective algorithms despite its simplicity. It is applied in most large vocabulary speech recognition systems. This algorithm computes a long-term mean value of feature vectors and subtracts this mean value from each of the vectors. This helps in reducing variability of data and does a kind of normalization and this is reason of naming it as Cepstral Mean Normalization. This procedure is applied to both of training and testing data. If purely convolutional noise is present, this set of speech parameters is unaffected by changes in noise. When additive noise is present, subtracting the mean is found to aid the robustness of system. However, the system behavior is hard to predict when both forms of noise are present [2,3].

3.3 Cepstral normalization using mean and variance

This approach, like CMS, is a robust feature method. In this technique, a long-term mean value of feature vectors is computed and subtracted from each of the vectors and then each vector is divided by the variance of feature vectors. Subtraction is similar to high pass linear filtering and division is similar to Automatic Gain Controlling (AGC) [11].

A buffer with length N is considered. N is number of cepstral vectors that can be placed in this buffer. In each moment, normalized coefficients are computed according to feature vectors that are in the buffer. When half of buffer is full with cepstral vectors, normalized coefficients are computed and the first vector is normalized. The new vector is placed in buffer and then second vector of buffer is normalized and this procedure is continued until next N vectors are placed in the buffer. Practical experiments show some limitation on selecting values of N. N must be determined in such a way that 40 ms of speech is covered. This value has showed a good performance. If N was set more than a threshold, recognition rate didn't increase and its curve was saturated [11,12].

3.4 SNR-dependent cepstral normalization (SDCN)

SNR-dependent cepstral normalization operates directly in cepstral domain. It adds a compensation vector to the noisy feature vector that depends exclusively on SNR of input frame. If v, x, y represent compensation vector, cepstral vector of noisy signal and cepstral vector of clean speech respectively, we can write:

$$y = x + v \tag{7}$$

Using equation (2), we can now rewrite equation (7) as:

$$y = x + v(SNR) \tag{8}$$

Equation (2) indicates that at high SNR, $x + h >> n$ and we can assume that logarithm term is approximately equal zero and so $y \approx x + h$. At low SNR, $x + h << n$ and we suppose that logarithm term depends on n and so y depends on noise only. Hence, the SDCN algorithm performs spectral equalization at high SNR and noise suppression in low SNR and at intermediate SNR, it can be an approximation [1, 9, 10]. Estimating the compensation vector v (SNR), the goal is to transform features of noisy speech signal so that it looks like the features of clean speech signal. Based on this method, the correction vectors are estimated by computing the average difference between cepstral vectors for noisy speech signal versus cepstral vectors for clean speech signal on a frame-by-frame basis as a function of input SNR. Correction vector *v* is made discrete into 25 intervals separated 1 db each. The result is the value of *v* for k = 0,…,25 in steps of $\Delta_{SNR} = 1\ db$ as follows [1]:

$$v[j,k] = \frac{\sum_{i=0}^{N-1}(x_i[j] - y_i[j])d[SNR_i - k\Delta_{SNR}]}{\sum_{i=0}^{N-1} d[SNR_i - k\Delta_{SNR}]} \tag{9}$$

Where $x_i[j]$, $y_i[j]$ represent element j of cepstrum vectors at frame i for the clean and noisy speech signal respectively. SNR_i is the SNR of frame i in noisy speech signal. d is Kronecker delta and sum is carried out for the entire N frames in database.

SDCN requires a stereo database of environment so this algorithm is environment-dependent.

3.5 Mapping from noisy space to clean space

In mapping approach, the goal is to find a function or a transformation to transform noisy features to clean speech signal features.

Mapping can be done in different domains. Mapping in transform doamin, maps noisy signal to clean signal [13]. With respect to nonlinear adding of noise and speech in new domains, isolation must be done in a nonlinear approach. Neural networks can be used as a nonlinear isolator. For different mappings in different SNRs, an estimated value of noise or SNR is used as one of neural network input. Figure 2 shows a block diagram of mapping in transform domain.

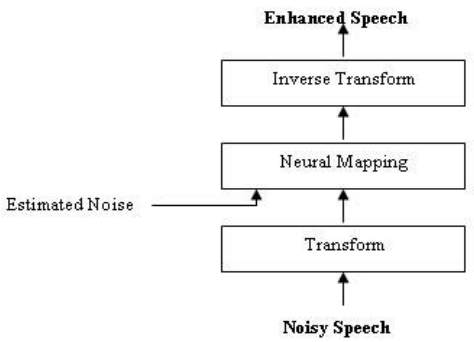


Figure 2. block diagram of mapping in transform domain

A number of researchers have performed preliminary investigations based on transform domain. We summarize two more relevant works in the following. The use of feedforward neural networks in log-spectral domain has been demonstrated in [13] for speech enhancement. Informal listening tests showed favorable acceptance for this work. Using this method, as a preprocessor on a simple HMM speech recognizer system, resulted an improvement in correct recognition rates.

In [22] a method for noise reduction in cepstral domain, based on a multi-layer neural network, is proposed and tested on a large database of isolated words contaminated with non-stationary F16 jet noise. A speech recognition system was tested that consisted of an auditory preprocessing module, a neural network for cepstral noise reduction and a neural network classifier. The recognition rate on a test database was improved, when the noise reduction network was added to the neural network classifier.
In following section, we propose an approach for mapping in cepstral domain between noisy and clean features. We justify the proposed approach using the model of environment that was studied in section 2.

4. Proposed Method

While Fourier transform domain is computationally convenient and removing noise in this domain can be done with a linear subtraction as equation (5), cepstral domain is often more desirable to work with. Feature extraction in cepstral domain is defacto standard for ASR systems and many of such systems have achieved a very high level accuracy in clean speech environment. Three methods based on this domain were discussed in pervious section. All of them try to do a simple subtraction or normalization to compensate the effects of environment. For better understanding the relation between clean features and noisy features in cepstral domain, we rewrite equation (3) in cepstral domain. Equation (3) represents feature vectors in log-spectral domain. As the relationship between cepstral vectors and log-spectral vectors is linear, we can write:

$$c_y = c_x + g(c_x, c_n, c_h) \tag{10}$$

Where c_y, c_x, c_n, c_h are cepstrum of y[m], x[m], n[m] and h[m] respectively. Function g can be computed as the inverse Fourier transform of function f(x, n, h):

$$g(c_x, c_n, c_h) = c_h + TF^{-1}\left[\ln(1 + e^{TF[c_n - c_x - c_h]})\right] \tag{11}$$

In this equation, g is an additive term that depends on c_x, c_n, c_h and represents the effects of environment on cepstrum of clean speech .
As we can see in equation (11), g( ) is a nonlinear function of variables to c_x, c_n, c_h. c_h plays 2 roles in this equation. It is an additive term and it appears in the exponent of exponential function. A simple mean subtraction in cepstral domain can only compensate the first role of c_h. This mean can be supposed as a good estimation of c_h. Errors in estimation of c_h degrade recognition rate.
If we have a good estimate of g( ) in equation (10) and subtract it from c_y, we will achieve the clean feature c_x. We propose a mapping between noisy cepstral features c_y and clean cepstral features c_x. In effect, function g( ) in (10), is a complex nonlinear function of c_h, c_n and clean features c_x. We use a neural network for learning g( ) and mapping between noisy and clean features.

In this approach, a representative training set of clean features and corresponding noisy features and an estimation of noise level, are used to train a neural network. Our network simply learns g( ) and maps noisy features c_y to clean features c_x. Distributions of noise source and speech signal affect the neural network mapping. Training with a single neural network averages across all noise sources and speech signals and assumes stationary of noise and speech signal.
In this paper, we use mapping in cepstral domain and clean cepstral coefficients are estimated from noisy cepstral coefficients by a neural network. Our neural network, as depicted in Figure 3, is a MLP with 4 layers. Each of two hidden layers has 20 nodes. Experiments have shown that in some similar cases , 20-30 nodes in hidden units are a good choice for estimation or mapping [13, 22]. Each of input and output layers have 12 nodes. Figure 4(a) shows how first cepstral coefficient of each frame is mapped to an estimated feature close to the clean feature, by MLP. Figure 4(b) shows mapping of 12 noisy cepstral features of one frame by MLP.

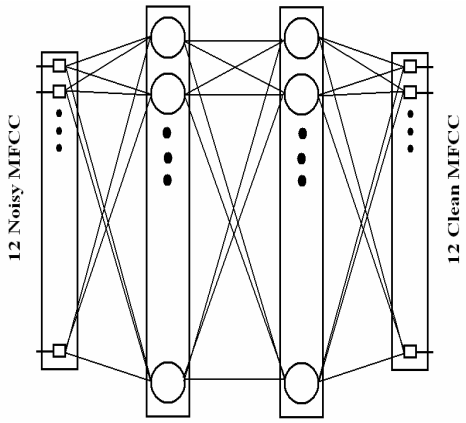
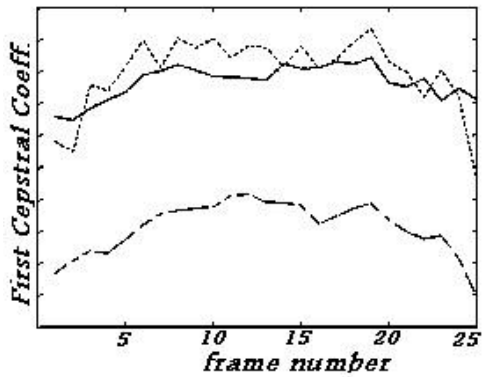


Figure 3. Architecture of MLP for estimation of clean cepstral coefficients

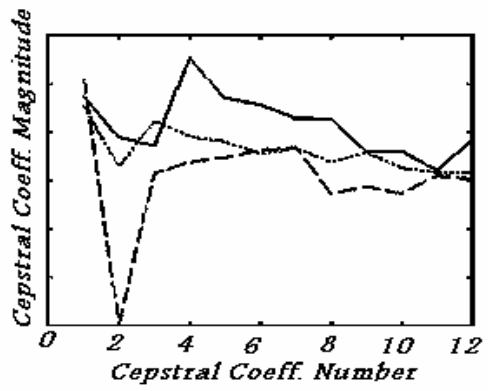
We use two approaches for mapping noisy to clean features. In the first one, called "MFCC+NN", we use noisy MEL frequency cepstral coefficients (MFCC) as input of neural network. In the second approach, called "CMS+NN", we use mean subtracted MFCC features. In effect, the second approach is a combination of CMS and mapping techniques.

5. Experiments and Results

In this section, we explain details of implementation for some of techniques that we have described in the third section and we compare them to our proposed technique.
Our recognition system is continuous density hidden Markov model (CDHMM), with 6 states and 2 Gaussian mixtures in each state. Our cepstral coefficients in one frame (solid: clean feature- dashed: noisy feature – dotted: estimated feature by MLP) Database is Persian number 1 to 10. Numbers have been recorded by sampling rate 16 kHz and 16 bits per sample, in a clean environment. A male and a female speaker have spoken these numbers. Ten utterances are used as training data and 5 utterances are used as testing



(a)



(b)

Figure 4. (a) first cepstral coefficients for 25 frame (b) 12 cepstral coefficients in one frame (solid: clean feature- dashed: noisy feature – dotted: estimated feature by MLP)

data for each speaker. Our feature vector contains 12 MFCC coefficients and their first order derivative and logarithm of energy and its first order derivative.

Hence, length of feature vector is 26. We have used 30 ms frames and 15 ms overlap and Hanning window.

We have used two types of noise: additive white Gaussian noise and F16 noise. These noises are selected from NOISEX92 database and downsampled to 16 kHz .We have used different SNRs: 0 db, 5 db, 10 db, 20 db, 30 db. Our Recognition rate for clean testing data was 98.7%. The implemented methods and their average results for two speakers and two types of noise are shown in Tables 1 and 2. The abbreviations in the tables are as follows:

MFCC: Feature vector mentioned above

MFCC+NN: mapping noisy MFCC to clean MFCC by MLP (proposed I)

CMS: cepstral mean subtraction

CMVS: cepstral normalization using mean and variance

CMS+NN: mapping mean subtracted noisy MFCC to clean ones (proposed II)

SS: spectral subtraction with overestimation 1.5 and noise flooring parameter 0.3

SDCN: SNR-dependent cepstral normalization.

In Table 1 , for SNR = 30 db, proposed method I (MFCC+NN) and spectral subtraction have the best performance. In Table 2 , for SNR = 30 db, proposed method II (CMS+NN) and spectral subtraction have the best performance. In both of tables and in SNR=20 db, 10db, 5 db, proposed method II (CMS+NN) outperforms other proposed method II has the best performance and techniques In SNR=0 db and for both two types of noise, performance of CMVS is acceptable.

Above results confirm the effectiveness of neural network mapping. In effect, the result of proposed method I is better than MFCC for all signal to noise ratios, specially for high level of noise. Proposed method II can be compared with CMS. This technique outperforms CMS for all signals to noise ratio and its performance is better than other techniques.

Table1. Average of recognition rate for 2 speakers and different data-driven robust methods and *white noise*

Method	SNR				
	30 db	**20 db**	**10 db**	**5 db**	**0 db**
MFCC	96.7%	74.7%	24%	15.3%	13.3%
MFCC+NN	98%	92.7%	69.3%	36.6%	25.3%
CMS	96.7%	93.3%	74.6%	47.3%	12%
CMVS	96.7%	85.4%	40%	27.5%	22.7%
CMS+NN	97.3%	95.3%	80.1%	52%	36%
SS	98.4%	96.2%	64.3%	37%	21.7%
SDCN	96.8%	89.9%	69.1%	35.3%	25%

Short length of utterance and insufficient number of frames can cause the unacceptable performance of CMVS in our experiments.

Increasing of distortion and musical noise in low SNR (specially 0 db) decreases performance of spectral subtraction. CMS is more effective for convolutive noise and it can cause low performance of CMS in SNR= 0 db.

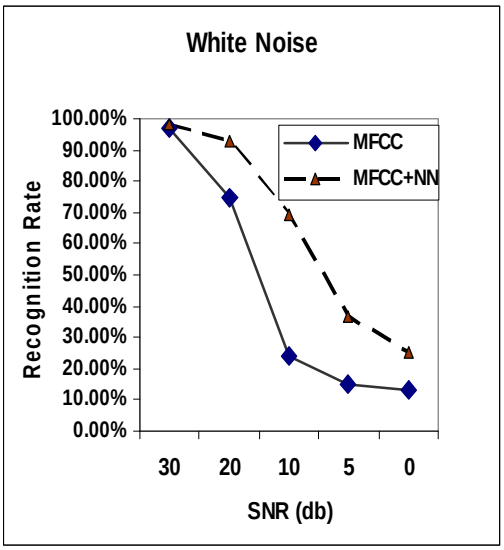
As mentioned before, SDCN compensates effects of convolutive noise on cepstral features in high SNR and effect of additive noise on cepstral features in low SNR. It has similar behavior in our experiments and in SNR=0 db, its performance is better than most of other methods.

MFCC, CMS, CMVS and SS methods don't need to stereo data and other methods need to both of clean and noisy signals and this is a limitation for their application in real word conditions.

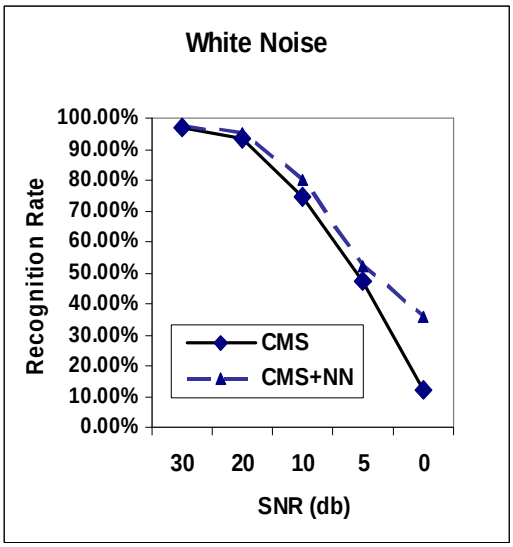
Table 2. Average of recognition rate for 2 speakers and different data-driven robust methods and *F16 noise*

Method	SNR				
	30 db	**20 db**	**10 db**	**5 db**	**0 db**
MFCC	94%	73.3%	26%	14.7%	10.7%
MFCC+NN	95.3%	90%	69.3%	34.7%	24%
CMS	96.7%	95.3%	71.3%	46.7%	11.3%
CMVS	96%	83.3%	38.3%	26%	20.3%
CMS+NN	97.3%	94%	78.7%	51.3%	34%
SS	98.7%	94.7%	63.3%	34%	20.7%
SDCN	96.7%	89.3%	69.3%	34%	24%

z



(a) MFCC and MFCC+NN



(b) CMS and CMS+NN

Figure 5. Recognition rate comparison diagram for white noise

6. Conclusions

In this paper, we have reviewed many techniques to increase the robustness of ASR systems in testing environments. We proposed a new approach for mapping noisy speech features to clean speech features. Two techniques based on this approach were suggested. In the first technique, noisy MFCC has been mapped to clean MFCC features. In the second method, noisy mean subtracted MFCC has been mapped to clean ones. Both techniques were implemented
and compared to precedent methods. They show good performances under additive white noise and F16 noise. We show that proposed methods are effective for robust speech recognition in noisy environment and give higher speech recognition rates in different SNRs.
The proposed approaches require to stereo database to train our neural network in a supervised manner. This requirement limits the application of our methods in real world condition. We try to employ a neural network with an unsupervised learning algorithm to overcome this problem.

References

[1] A. Acero, Acoustical and Environmental Robustness in Automatic Speech Recognition, Ph. D. Dissertation, Carnegie Mellon University, 1990.

[2] M. J. F.Gales, Model-based Techniques for Noise Robust Speech Recognition, Ph. D. Dissertation, Cambridge University, 1996.

[3] P.J.Moreno, Speech Recognition in Noisy Environment, Ph.D. Dissertation, Carnegie Mellon University, 1996.

[4] H. Hermansky, " Perceptual Linear Predictive Analysis of Speech," *Journal of Acoustical Society of America*, pp. 1738-1752, 1990.

[5] H. Hermansky, N. Morgan, "RASTA Processing of Speech," *IEEE Trans. Speech and Audio Processing*, vol.2, no. 4, pp. 587-589, 1994.

[6] D. Kim, S. Lee, R. M. Kil, "Auditory Processing of Speech Signals for Robust Speech Recognition in Real-World Noisy Environment," *IEEE Trans. Speech & Audio Processing* , vol. 7, no. 1, 1999.

[7] L. Josifovski, Robust Automatic Speech Recognition with Unreliable data, progress report #2, Speech and Hearing Research Group, Department of Computer Science, University of Sheffield, 1999.

[8] C. J. Legetter, P. C. Woddland, "Flexible Speaker Adaptation Using Maximum Likelihood Linear Regression," *Proc. ARPA Spoken Language Technology Workshop*, pp 104-109, 1995.

[9] A. Acero, R. M.Stern, "Environmental Robustness in Automatic Speech Recognition," *Proc. of the ICASSP*, Albuquerque, New Mexico, pp. 849-852, 1990.

[10] A. Acero, M. Stern, "Robust Speech Recognition by Normalization of the Acoustic Space," *Proc. of the ICASSP*, Toronto, Ontario, pp. 893-896, 1991.

[11] O. Vikki, K. Launilla, "Cepstral Domain Segmental Feature Vector Normalization," *Speech Communication*, no. 25, pp.133-147, 1998.

[12] J. Hakkien, J. Sountasta, R. Hariharan, M. Vasilache, K. Launilla, " Improved Feature Vector Normalization for noise Robust Connected Speech Recognition," *Proc. of EuroSpeech* , vol. 6, pp. 1833-2836, 1999.

[13] E. A. Wan, A. T. Nelson, *Handbook of Neural Networks for Speech Processing*, Boston, USA, 1998.

[14] A. P. Varga, R. K. Moore, " Hidden Markov Model Decomposition of Speech and Noise," *Proc. of ICASSP*, pp. 845-848, 1990.

[15] A. P. Varga, R. K. Moore, " Simultaneous Recognition of Concurrent Speech Signals using Hidden Markov Model Decomposition," *Proc. of EuroSpeech*, pp. 1175-1178, 1991.

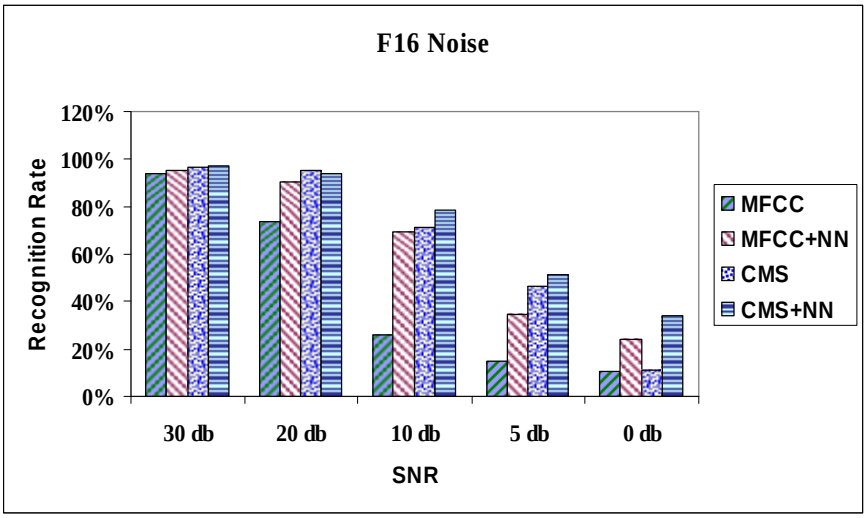


Figure 6. Recognition rate comparison diagram for different methods and SNRs when noise is F16 noise

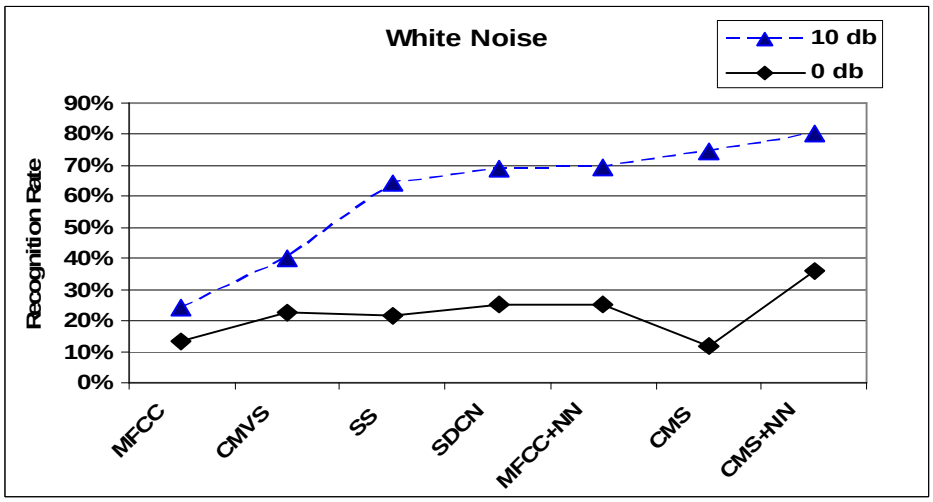


Figure 7. Comparison diagram for recognition rate of methods for white noise

[16] S. V. Vaseghi, *Advanced Signal Processing and Noise Reduction*, John Wiley & Sons Publication Company, Second Edition, 2000.

[17] S. F. Boll, "Suprression of Acoustic Noise in Speech using Spectral Subtraction," *IEEE Trans. Acoustics, Speech & Signal Processing*, 1979.

[18] C. Mokbel, L. Barbier, Y. Kerlou, G. Chollet, " Word Recognition in Car: Adapting Recognizers to New Environments," *Proc. of EuroSpeech*, vol.1, pp. 707-710, 1992.

[19] B. Nasersharif, A hybrid HMM/MLP System for Two Digits Farsi Numbers Recognition in a Speaker Dependent Method, M.S.C thesis, Iran University of Science and Industry, 2001.

[20] P. Lockwood, J. Boudy, M. Blanchet, "Non-linear spectral subtraction (NSS) and Hidden Markov Models for robust speech recognition in car noise environments," *Proc. of ICASSP*, vol. 1, pp. 265-268, 1992.

[21] R. Bouquin, " Traitements Pour la Reduction du Bruit Sur la Parole. Applications Radio-Mobile, " Thesis of university Renne, Renne, France, 1991.

[22] H. B. D. Sorensen, " A Cepstral Noise Reduction Multi-layer Neural Network," *Proc. of ICASSP*, vol.2, pp 933-936, 1991.

[23] C.Cerisara , I.Illina, "Robust Speech Recognition to Non-stationary Noise Based on Model Driven Approach," *Proc. of EuroSpeech* , vol. 6 , pp.3053-3056 , 2003.

[24] A. Morris, A.Hagen, H.Glotin, H.Bourlard, "Multi-stream adaptive evidence combination to noise robust ASR," *Speech Communication*, vol. 34, Issue. 1-2, pp. 25–40, 2001.

[25] Cerisara, C., Fohr, D.,"Multi-band automatic speech recognition," *Computer Speech and Language*, vol. 15, Issue. 2, pp. 151-174, April 2001.

[26] B.Gajic , K.K.Paliwal , "Robust Speech Recognition using Feature based on Zero Crossing with Peak Amplithude," *Proc. of ICASSP,* vol. 1, pp. 64-67, 2003.

[27] S.Molau , F.Hilger , H.Ney , "Feature Space Normalization in Adverse Acoustic Conditions," *Proc. of ICASSP,* vol. 1, pp. 656-659, 2003.

[28] J.Beh, H.ko, "A Novel Spectral Subtraction Scheme for Robust speech Recognition: Spectral Subtraction using Spectral Harmonics of Speech," *Proc. of ICASSP,* vol. 1, pp. 648-651, 2003.

[29] D.A.Raynolds, "Channel Robust Speaker Verification via Feature Mapping," *Proc. of ICASSP,* vol. 2, pp.53-56, 2003.

[30] R.Gemello, F.Mana, R.Albesano, DeMori, "Robust Multiple Resolution Analysis For Automatic Speech Recognition," *Proc. of EuroSpeech* , vol. 6 , pp.3033-3036 , 2003.

[31] P.J.Moreno, B.Raj, R.M.Stern, "A Vector Taylor Series Approach for Environment Independent Speech Recognition," *Proc. of ICASSP*, pp. 733-736, Atlanta , 1996.

[32] A.Acero, L.Deng, T.Kristjansson, and J.Zhang. "HMM Adaptation using Vector Taylor Series for Noisy Speech Recognition," *Proc .of ICSLP*, Vol.3, pp. 869-872, 2000.

[33] H.Attias, L.Deng, A.Acero, and J.Platt. "A New Method for Speech Denoising and Robust Speech Recognition using Probabilistic Models for Clean Speech," *Proc. of Eurospeech*, pp. 1903-1906, 2001.

[34] M.T.Hagan, H.B.Demuth, M.H.Beale, *Neural Network Design*, Boston, PWS Publishing, 1996.

[1] Cepstral Mean Subtraction
[2] Perceptual Linear Predictive analysis of speech
[3] Zero Crossing with Peak Amplitude

Babak Nasersharif was born in Tehran in 1974. He received the B.S degree in computer engineering from Amir Kabir university of technology, Tehran, Iran in 1997. He received the M.S degree in computer engineering from Iran university of science and technology in February 2001 and is presently a Ph.D. student in computer engineering at Iran university of science and technology. His research interests include artificial intelligence, speech recognition, robust speech recognition, digital signal processing and wavelet.

Mohsen Rahmani received the B.S degree in computer engineering from Shiraz university, Iran, in 1999. He received the M.S degree in computer engineering from Iran university of science and technology, in 2002 and is presently a Ph.D. student in computer engineering at Iran university of science and technology. His research interests include computer architecture, robust speech recognition and speech enhancement.

Ahmad Akbari received the BSc. degree in electronics engineering and the MSc. degree in communications engineering from Isfahan university of technology (IUT) in 1986 and 1989 respectively. He received the DHET degree in computer networks from ENSEEIHT, Toulouse, France in 1991. He also received the DEA and Ph.D. degrees in signal processing and telecommunications from university of Rennes 1, Rennes , France in 1992 and 1995 respectively. In 1996 he joined the computer engineering department at Iran university of science and technology (IUST) as an assistant professor, where he is now director of research center for information technology (RCIT). His research interests include acoustic modeling of speech, robust speech recognition, speech enhancement, implementation of signal processing algorithms, voice applications and interfaces and web technologies. Dr. Akbari is a member of board of the Computer Society of Iran (CSI) since 1999.