Noise Robust Features Based on MVA Post-processing

Mohamed Korba, Djemil Messadeg, Houcine Bourouba, Rafik Djemili

Noise Robust Features Based on MVA Post-Processing

Mohamed Cherif Amara Korba[1,2], Djemil Messadeg[3],
Houcine Bourouba[2], Rafik Djemili[4]

[1] Mohammed Cherif Messaadia University, Souk-Ahras, Algeria
[2] PI:MIS Laboratory, May 8, 1945 University, Guelma, Algeria
[3] LASA Laboratory, Badji Mokthar University, Annaba, Algeria
[4] August 20, 1955 university, Skikda, Algeria
{amara_korba_cherif, messadeg, bourouba2004
rafik_djemili}@yahoo.fr

Abstract. In this paper we present effective technique to improve the performance of the automatic speech recognition (ASR) system. This technique consisting mean subtraction, variance normalization and application of temporal auto regression moving average (ARMA) filtering. This technique is called MVA. We applied MVA as post-processing stage to Mel frequency cespstral coefficients (MFCC) features and Perceptual Linear Prediction (RASTA-PLP) features, to improve automatic speech recognition (ASR) system.

We evaluate MVA post-processing scheme with aurora 2 database, in presence of various additive noise (subway, babble because, exhibition hall, restaurant, street, airport, train station). Experimental results demonstrate that our method provides substantial improvements in recognition accuracy for speech in the clean training case. We have completed study by comparing MFCC and RSTA-PLP After MVA post processing.

1 Introduction

Most speech recognition systems are sensitive to the nature of the acoustical environments within which they are deployed. The performance of ASR systems decreased dramatically when the input speech is corrupted by various kinds of noise sources. It is quite significant when the test environment is different from the training environment.
In the last two decades, substantial efforts have been made and also number of techniques have been presented to cope with this issue improve the ASR performance. Unfortunately these same algorithms frequently do not provide significant improvements in more difficult environments.
MFCC and RASTA-PLP have served as very successful front-ends for the Hidden Markov Model (HMM) based speech recognition. Many speech recognition systems based on these front-ends have achieved a very high level of accuracy in clean speech environment [14], [15]. However, it is well-known that MFCC is not robust enough in noisy environments, which suggests that the MFCC still has insufficient sound representation capability, especially at low signal-to-noise-ratio (SNR).
This paper presents noise-robust technique that is simple and effective. The technique post-processing speech features using MVA [10],[11],[12]. The advantage of this technique, it makes no change to the recognition system, it does not change the size of

the space, it can be applied on any acoustic feature. it has been shown in [10] and [11] the efficacy of this technique on the database Aurora 2.0 and Aurora 3.0.

This paper is organized as follows: in section 2, we describe MVA post-processing technique, in section3, we show a graphical comparison between different features, in section 4, we present experimental result and in section 5 the work is concluded.

2 Definition and Analyze of MVA Post-Processing Technique

2.1 Definition of MVA Post-Processing Technique

In this part, we describe different steps of development of MVA post-processing technique, Figure 1 provided a block diagram.

For a given utterance, we represent the data by matrix *C* whose element $C_d(t)$ is the dth component of the feature vector at time t, $t = 1 \dots T$, the number of frames in the utterence and $\mathrm{d} = 1 \dots \mathrm{D}$, the dimension of the feature space, in other words, each column of C represents a time sequence.

$$\begin{bmatrix} C_1(1) & \cdots & C_1(T) \\ \vdots & \ddots & \vdots \\ C_d(1) & \cdots & C_d(T) \end{bmatrix} \quad (1)$$

The first step we application mean subtraction (MS) [6], [7] defined by:

$$\bar{C}_d = C_d(t) - \mu_d \quad (2)$$

Where μ_d is mean vector estimated from data and $\bar{\mathrm{C}}_\mathrm{d}$ is the subtracted feature.

$$\mu_d = \frac{1}{T}\sum_{t=1}^{T} C_d(t) \quad (3)$$

MS is an alternate way to high-pass filter cepstral coefficients, it force the average values of cepstral coefficients to be zero in both the training and testing domains. it also removes time-invariant distortions introduced by the transmission channel and recording device.

The second step is Variance normalization (VN) [8], [9] defined by:

$$\tilde{C}_d = \frac{\bar{C}_d(t)}{\sqrt{\sigma_d}} \quad (4)$$

Where σ_d is variance vector estimated from data.

$$\sigma_d = \frac{1}{T}\sum_{t=1}^{T}(C_d(t) - \mu_d)^2 \quad (5)$$

The third step is processing by a mixed auto-regression moving average (ARMA) filtering. In this study we have used two types of ARMA filters: Non Causal ARMA Filter defined by

$$\check{C}_d(t) = \begin{cases} \frac{\sum_{i=1}^{M}\tilde{C}_d(t-i) + \sum_{j=0}^{M}\tilde{C}_d(t+j)}{2M+1} & if\ M < t \le T - M \\ \tilde{C}_d(t) & Otherwise \end{cases} \quad (6)$$

and Causal ARMA Filter defined by :

$$\check{C}_d(t) = \begin{cases} \frac{\sum_{i=1}^{M} \check{C}_d(t-i) + \sum_{j=0}^{M} \tilde{C}_d(t+j)}{2M+1} & if\ M < t \leq T \\ \tilde{C}_d(t) & Otherwise \end{cases} \quad (7)$$

where M is the order of ARMA filter.

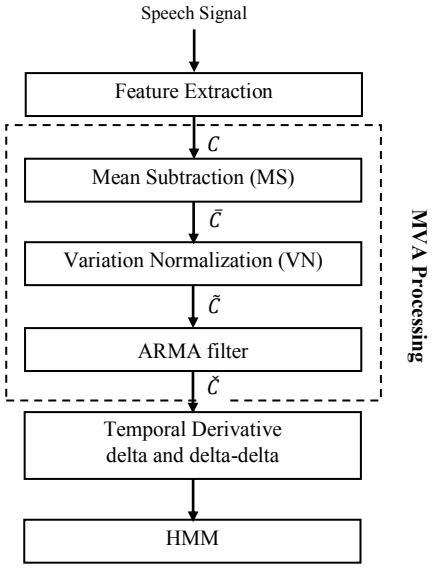


Fig. 1. Block diagram of MVA post-processing technique

In all our experiments, the performances of ASR system are enhanced by adding time derivatives to the basic static parameters for different features. The delta coefficients are computed using the following regression formula:

$$\Delta(t) = \frac{\sum_{b=1}^{B} d\left(\check{C}_d(b+1) - \check{C}_d(b-1)\right)}{2\sum_{b=1}^{B} b^2} \quad (8)$$

Where $\Delta(t)$ is the delta coefficient computed in terms of the corresponding static coefficients $\tilde{C}_d(t-B)$ to $\tilde{C}_d(t+B)$. The same formula is applied to the delta to obtain acceleration coefficients.

2.2 Effect of normalization and ARMA filter on acoustic features

In Fig. 2 and fig. 3 the time sequences of C0 and C1 are plotted for both features RASTA-PLP and MFCC of the utterance of digit string "98Z7437" corrupted by different levels of additive subway noise from the Aurora 2.0 database. For both RASTA-PLP and MFCC features, we see enormous differences between the plots of

the clean case and the more noisy case. In particular, the clean and noisy plots have quite a different average value and dynamic range.

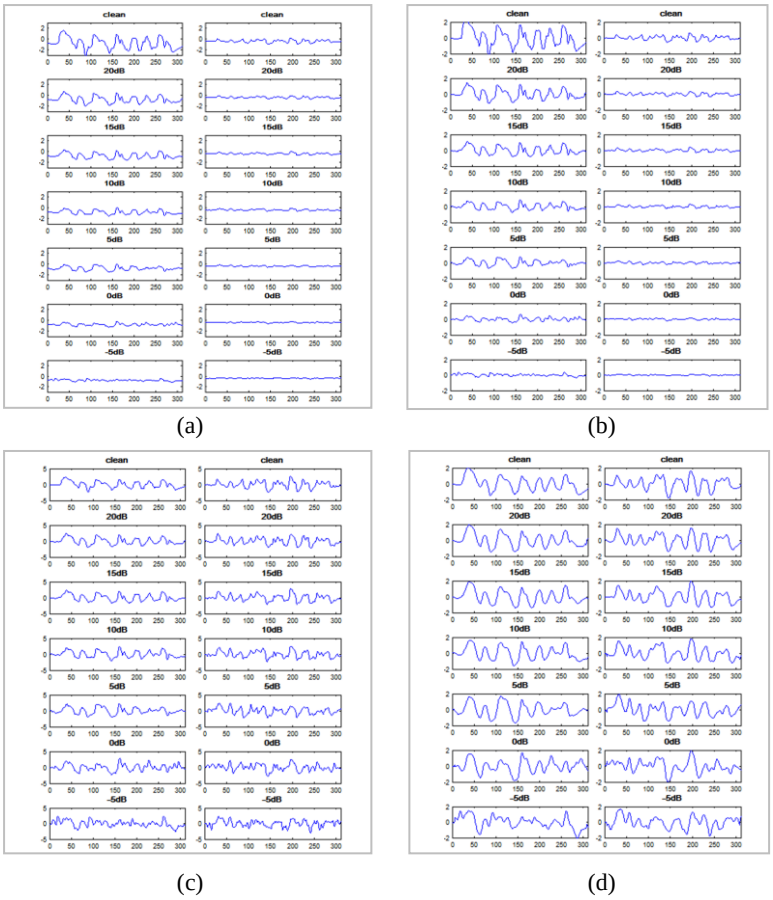
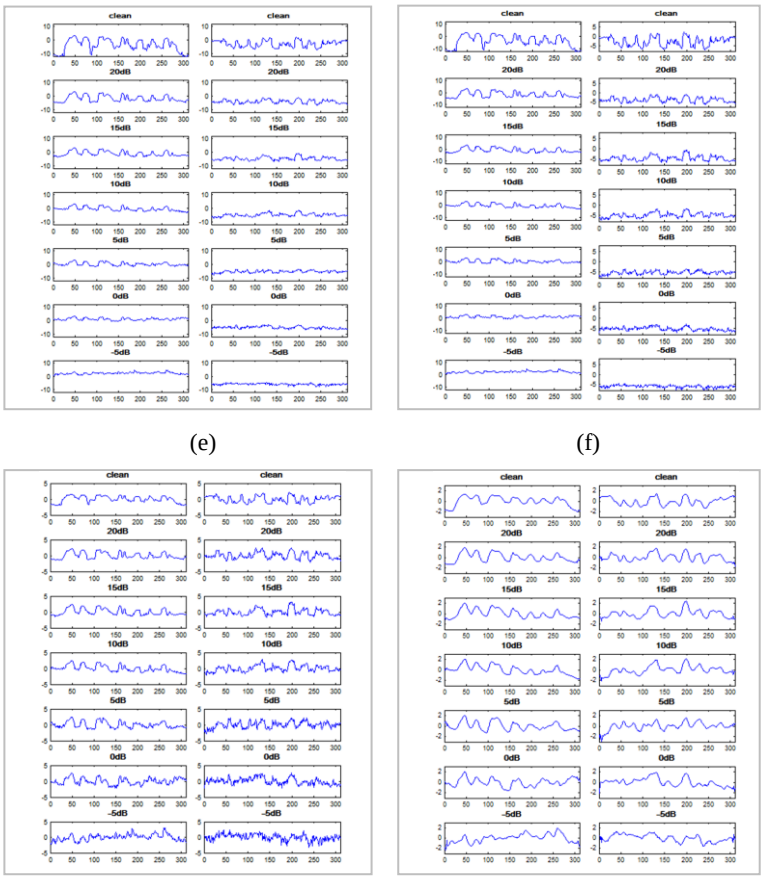


Fig. 2. The time sequence of C0 and C1 coefficients of RASTA-PLP features for the digit string "98Z7437" corrupted by additive subway noise, (a) RASTA-PLP features, (b) time sequence of RASTA-PLP + MS, (c) time sequence of RASTA-PLP + MS + VN, (d) time sequence of RASTA-PLP + MVA.

After MS and VN is applied, the difference between the clean and noisy cases are made much less severe After MS and VN is applied, the differences between the clean and noisy cases are made much less severe. Still, however, some differences remain between the clean and noisy cases. We notice in particular the case of C1, that after the application of MS and VN, the time sequences in noisy speech show spurious spikes relative to the clean case. In order to further reduce differences, we apply ARMA filtering which smoothes out the sequences thus making them more similar to each other. We remark that, the effects of noise on the MVA features are less severe for both MFCC and RASTA-PLP features.



(e) (f)

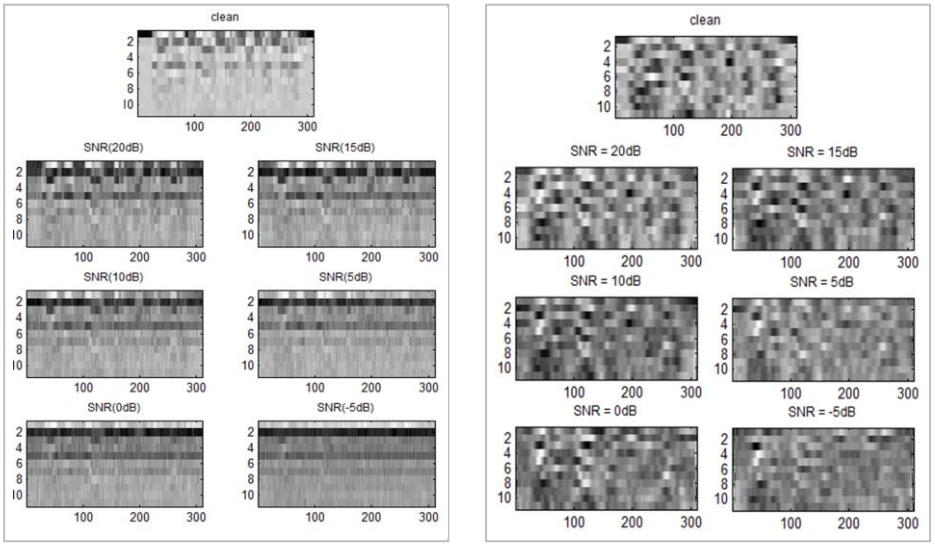
(g) (h)

Fig. 3. The time sequence of C0 and C1 coefficients of MFCC features for the digit string "98Z7437" corrupted by additive subway noise, (e) MFCC features, (f) time sequence of MFCC + MS, (g) time sequence of MFCC + MS + VN, (h) time sequence of MFCC + MVA

3 Graphical Comparison between the Different Features

Fig. 4 shows a sample comparison between baseline MFCC features and corresponding MFCC MVA post-processing features for the digit string "98Z7437" corrupted with Subway noise at different levels of noise (clean, 20dB, 15dB, 10dB, 5dB, 0dB, -5dB). As standard in MFCC, a window size of 25 ms with an overlap of 10 ms was chosen, and Cepstral features were obtained from DCT of log-energy over 23 Mel-scale filter banks.

The degradation of spectral features for baseline MFCC features in the presence of noise is evident; whereas MFCC with MVA post-processing features obtained with No Causal ARMA filter prevail at elevated noise levels. For SNR $\leq$ 0dB we can see clearly that MFCC with MVA is better noise robustness than MFCC baseline features.



(a) (b)

Fig. 4. **(a)** Baseline MFCC features for the digit string "98Z7437" corrupted by subway noise, **(b)** MFCC with MVA post-processing features for the digit string "98Z7437" corrupted by subway noise. (No causal ARMA filter used, filter order = 5)

3.1 Speech Features Description

This part contains a short description of the most widely used acoustic features in automatic speech recognition. Many of current ASRs are based on Mel frequency cepstral coefficients MFCC [5] or RASTA-PLP coefficients [3],[4]. They operate efficiently in the clean environment, by against the performances of ASR decreases dramatically in presence of noise. To remedy this problem, we introduced a post-processing stage to improve their performances without bringing changes in their structures. Table 1 shows the configuration of MFCC and RASTA-PLP features used for experiences.

4 Experiments

We first describe in detail the Aurora 2 database, then, we present experimental results that are intended to show the contribution of MVA post-processing technique for both acoustic features MFCC and RASTA-PLP in the presence of large variety of additive noise. We determine type and order of ARMA filter that gives the best speech accuracy for acoustic features used (MFCC and RASTA-PLP).

4.1 Description of Aurora 2 Database

Our speech recognition experiments were conducted using the Aurora 2 database and task [2]. The Aurora task [2] has been defined by the European Telecommunications Standards (ETSI) to standardize a robust feature extraction technique for a distributed speech recognition framework.

The Aurora 2 database is a subset of the TIDigits, which contains a set of connected digit utterances spoken in English; while the task consists of the recognition of the connected digit utterances interfered with real noise artificially added in a wide range of SNRs (-5dB, 0dB, 5dB, 10dB, 15dB, 20dB and Clean) and the channel distortion is additionally included in Set C. Noise signals are recorded at different places including suburban train, babble, car, exhibition hall, restaurant, street, airport and train station.

Two training modes are defined, training on clean data only and training on clean as well as noisy data (multi-condition). For the first mode, training data contain 8440 clean utterances produced by 55 male and 55 female adults. For the multi-condition training, 8440 utterances from TIDigits training parts are equally split into 20 subsets with 422 utterances in each subset. Four types of noise, Suburban train, babble, car, and exhibition hall noises are added to 20 subsets at 5 different SNRs (5dB, 10dB, 15dB, 20 dB and Clean).

The testing data consist of 4004 utterances from 52 male and 52 female speakers in the TIDigits test part are divided into four subsets with 1001 utterances in each. One noise is added to each subset at SNRs of 20 to -5 dB in decreasing steps of 5 dB after speech and noise are being filtered with the G. 712. Three test sets are defined as below:

Test Set A: four types of noise, babble, car, suburban train, and exhibition hall are added to the four subsets of utterances to produce 28028 utterances (4x7x1001 utterances). This set leads to a high match of training and test data as it contains the same noises as used for the multi-condition training mode.

Test Set B: the other type of noise, street, restaurant, airport and train station, are added to the four subsets of utterances to produce 28028 utterances (4x7x1001 utterances), similar to test A.

Test Set C: two types of noise, suburban train and street, are individually added to two of the four subsets of utterances to produce 14014 utterances (2x7x1001 utterances). Speech and noise are filtered with the MIRS frequency characteristic before adding.

In this study we used two sets of tests, Test Set A and Test Set B. for all experiments HMM baseline system is trained in clean condition.

4.2 The HTK Recognizer

For the baseline system, the training and recognition tests used the HTK recognition toolkit [1], which followed the setup originally defined for the ETSI Aurora evaluations.

Each digit was modeled as a left to right continuous density HMM with 16 states

with each state having 3 mixtures. Two pause models, silence "sil" and short pause "sp", were defined. The "sil" model had three states with six Gaussian mixtures per state. The "sp" model had one state with six Gaussian mixtures.

Script files provided with the Aurora 2 database for the purpose of training and testing a HTK based recognizer were used in the evaluation of the front-ends. The version of HTK used was HTK 3.3. We used the RASTA-PLP implementation that is valuable at [13], we used the version of conventional MFCC processing implemented as part of HTK platform. Configurations of RASTA-PLP and MFCC features used in our experiments are given by the table 1.

Table 1. Features parameters used for experimental analysis

Configuration features	MFCC	RASTA-PLP
Frame length (ms)	25	25
Frame shift (ms)	10	10
Pre-emphasis coefficient	0.97	NO
Analyses window	Hamming	NO
frequency range	64 – 4000 Hz	0 – 4000 Hz
No. Mel filterbanks	23	/
LPC Model order	/	11
Rasta filter	/	do
Appended log frame energy	yes	yes
Appended features	Δ + Δ Δ	
Δ window (frames)	±4	±4
ΔΔ window (frames)	±1	±1
Feature dimension	39	39

4.3 Analyses

The tables below were done to determine type and order of ARMA filter that gives the best recognition accuracy for each acoustic feature. Tables show the contribution importance of order of filter on the performances of ASR system.

For all our experiments, best results have been obtained with the non-causal ARMA filter for both acoustic features. We varied the order of the filter until 9, the best performances of the system have been obtained with order M = 6.

Table 2. Comparison of different type and order ARMA filters, word accuracy Rasta-PLP, Test speech average over SNR (clean, 20, 15, 10, 5, 0, -5dB)

	Filter Order				
Filter Type	2	3	4	5	6
Causal ARMA filter	67.52	69.09	69.06	68.94	**68.98**
Non Causal ARMA filter	68.33	68.63	69.74	69.66	**70.46**

Table 3. Comparison of different type and order ARMA filters, word accuracy MFCC, Test speech average over SNR (clean, 20, 15, 10, 5, 0 -5dB)

	Filter Order				
Filter Type	2	3	4	5	6
Causal ARMA filter	67.57	66.84	67.74	69.35	**70.55**
Non Causal ARMA filter	67.67	69.10	69.53	70.05	**71.40**

4.4 Performance of MVA post-processing

In this section we describe the recognition accuracy obtained using MVA post-processing for MFCC and RASTA-PLP features, under various noise conditions at different SNR levels (Clean, 20, 15, 10, 5, 0, -5dB).

In figure 5, remarkably improvements have been achieved up to 20% compared to RASTA-PLP features without any normalization, up to 10% to features with MSVN normalization and up to 5% to features with MS normalization.

In figure 6, Substantial improvements have been achieved up to 25% compared to MFCC features without any normalization, up to 15% to features with MS + VN normalization and up to 8% to features with MS normalization.

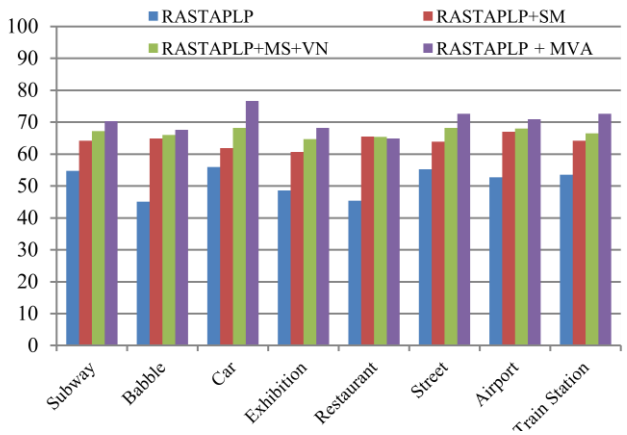


Fig. 5. Comparison of recognition accuracy for different RASTA-PLP features configuration (MVA: use non causal ARMA filter, M = 6), the recognition accuracy is calculated on an average of 7 SNR levels. (Clean, 20dB, 15dB, 10dB, 5dB, 10dB, 5dB, 0dB, -5dB).

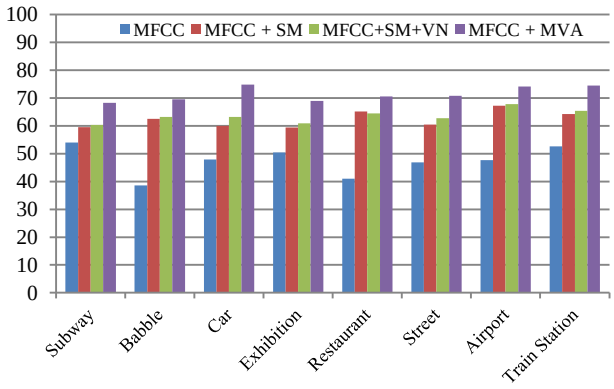


Fig. 6. Comparison of recognition accuracy for different MFCC features configuration (MVA: use non causal ARMA filter, M = 6), the recognition accuracy is calculated on an average of 7 SNR levels. (Clean, 20dB, 15dB, 10dB, 5dB, 10dB, 5dB, 0dB, -5dB).

Figure 7 shows a comparison between MFCC features and Rasta-PLP features, in the presence of stationary noise subway, street and car the RASTA-PLP features are more efficient compared to MFCC features, but in the presence of noises majority babble, suburban train, exhibition hall, restaurant, airport and the train station the MFCC coefficients provide best performance to ASR system.

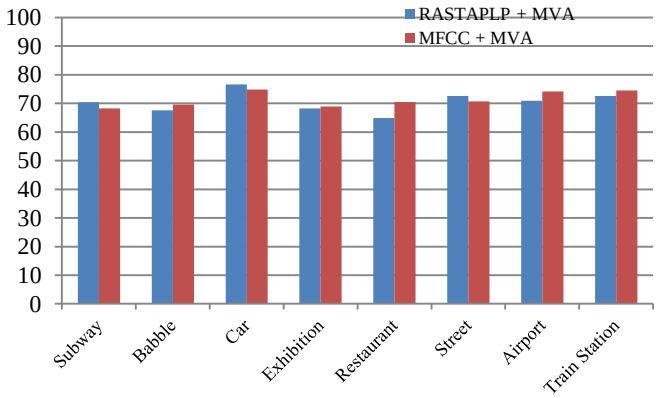


Fig. 7. Comparison of recognition accuracy for MFCC + MVA features with RASTA-PLP + MVA features (MVA: use non causal ARMA filter, M = 6 for both types of features), the recognition accuracy is calculated on an average of 7 SNR levels. (Clean, 20dB, 15dB, 10dB, 5dB, 10dB, 5dB, 0dB, -5dB).

5 Conclusions

In this paper, we introduce MVA technique to MFCC and RASTA-PLP features to improve the noise robustness of speech features. We have shown that normalization techniques followed by ARMA filter are vital for conditions with major mismatch between training and test condition.

The experimental results show that application of MVA to the Aurora 2 database can provide further robustness to noise for various types of features, and higher accuracy rates can be thereby achieved.

Acknowledgment. This work was supported by PI:MIS laboratory of Guelma University. The authors would like to thank Professor A. Boukrouche and H. Doghmane for Helpful discussion.

References

1. S. Young et al., "The HTK Book Version 3.3," 2005
2. H. G. Hirsch, D. Pearce, "The AURORA Experimental Framework for the Performance Evaluations of Speech Recognition Systems under Noisy Conditions," in Proc. ISCA ITRW ASR 2000.
3. H. Hermansky, "Perceptual linear prediction analysis of speech," J. Acoust. Soc. Am., vol. 87, no. 4, pp. 1738–1752, Apr. 1990.
4. H. Hermansky and N. Morgan, "RASTA processing of speech," IEEE. Trans. Speech Audio Process., vol. 2, no. 4, pp. 578–589, Oct. 1994.
5. S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," IEEE Trans. Acoust., Speech, and Signal Processing, vol. 28, no. 4, pp. 357–366, Aug. 1980.
6. B. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," Journal of the Acoustical Society of America, vol. 55, pp. 1304–1312, 1974.
7. S. Furui, "Cepstral analysis technique for automatic speaker verification," IEEE Trans. Acoust., Speech, Signal Process., vol. 29, no. 2, pp. 254–272, Apr. 1981.
8. P. Jain and H. Hermansky, "Improved mean and variance normalization for robust speech recognition," in IEEE Int. Conf. Acoust., Speech and Signal Processing, May. 2001.
9. G. D. Cook, D. J. Kershaw, J. D. M. Christie, C. W. Seymour, and S. R.Waterhouse, "Transcription of broadcast television and radio news: the 1996 abbot system," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), Munich, Germany, 1997
10. C.-P. Chen, J. Bilmes, and K. Kirchhoff, "Low-resource noise-robust feature post-processing on Aurora 2.0," in Proc. Int. Conf. Spoken Lang. Process. (ICSLP), 2002, pp. 2445–2448.
11. C.-P. Chen, K. Filali, and J. Bilmes, "Frontend post-processing and backend model enhancement on the Aurora 2.0/3.0 databases," in Proc. Int. Conf. Spoken Lang. Process. (ICSLP), 2002, pp. 241–244.
12. C.-P. Chen and J. Bilmes, MVA processing of speech features Dept. Elect. Eng., Univ. Washington, Seattle, WA, Tech. Rep. UWEETR- 2003-0024, 2003 [Online]. Available: http://www.ee.washington.edu/ techsite/papers
13. D. Ellis. (2006) PLP and RASTA (and MFCC, and inversion) in MATLAB using melfcc.m and invmelfcc.m. [Online]. Available: http://labrosa.ee.columbia.edu/matlab/rastamat/
14. M. N. Stuttle, M.J.F. Gales, "A Mixture of Gaussians Front End for Speech Recognition," Eurospeech 2001, pp. 675-678, Scandinavia, 2001.

15. J. Potamifis, N. Fakotakis, G. Kokkinakis, " Improving the robustness of noisy MFCC features using minimal recurrent neural networks", Neural Networks, IJCNN 2000, Proceedings of the IEEE-INNS-ENNS International Joint Conference on,vo1.5, pp. 271-276, 2000.