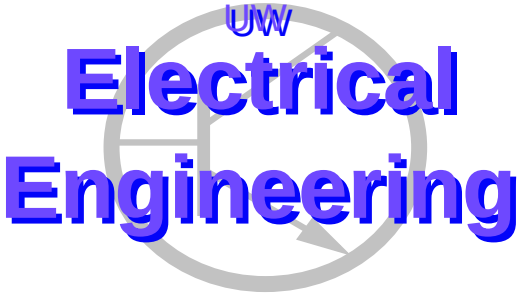
Blind MVA Speech Feature Processing on Aurora 2.0

Chia-Ping Chen, Jeff Bilmes
{chiaping,bilmes}@ssli.ee.washington.edu
Department of Electrical Engineering, University of Washington
Seattle, WA 98195-2500

Daniel P. W. Ellis
dpwe@ee.columbia.edu
Department of Electrical Engineering, Columbia University
New York 10027



UWEE Technical Report
Number UWEETR-2004-0017
June 2004

Department of Electrical Engineering
University of Washington
Box 352500
Seattle, Washington 98195-2500
PHN: (206) 543-2150
FAX: (206) 543-3842
URL: http://www.ee.washington.edu

Blind MVA Speech Feature Processing on Aurora 2.0

Chia-Ping Chen, Jeff Bilmes
{chiaping,bilmes}@ssli.ee.washington.edu
Department of Electrical Engineering, University of Washington
Seattle, WA 98195-2500

Daniel P. W. Ellis
dpwe@ee.columbia.edu
Department of Electrical Engineering, Columbia University
New York 10027

University of Washington, Dept. of EE, UWEETR-2004-0017

June 2004

Abstract

This paper is focused on the MVA (mean subtraction, variance normalization, and ARMA filtering) feature post-processing scheme for noise-robust automatic speech recognition. MVA has shown great success in the past on the Aurora 2.0 and 3.0 corpora. To test its generality, in this work MVA is blindly applied to many different acoustic feature extraction methods, and is evaluated using the Aurora 2.0 corpus. Specifically, we apply MVA post-processing to feature extraction techniques including: MFCC, LPC, PLP, RASTA, Tandem, Modulation-filtered Spectrogram and Modulation Cross-CorreloGram. We find that while effectiveness depends on the extraction method used, the majority of features benefit significantly from MVA. We conclude with a brief analysis.

1 Introduction

Vulnerability to noise is a main obstacle for automatic speech recognition (ASR) systems to become more widely used. One small step in noise-robustness improvement could be one giant leap in an ASR system's viability. Indeed, there is still considerable room for improvement, as indicated by the fact that human performance in noise is still far better than a machine.

Besides being accurate, ASR systems also need to have tolerable computational and memory demands, especially on portable devices. MVA post-processing is a very effective noise-robust technique on small-vocabulary ASR tasks [4, 5]. It achieves essentially the same performance level as the most effective noise-robust techniques without any signifi cant increase in computation, and thus is favored in the above sense.

In this paper, the effectiveness of MVA is investigated on a spectrum of speech features with the goal of discovering what characteristics do and do not combine well with MVA. The paper is organized as follows: In Section 2, the experimental setup of the various front-end features and back-end recognizer are described. In Section 3, the results are presented for each feature set. In Section 4, an explicit comparison across different feature sets is presented. Conclusions are drawn in Section 5.

2 Setup

The collection of feature sets included in this investigation is meant to cover a broad range of different feature types. The chosen feature sets are the MFCC, LPC, LPC-CEPSTRA, Tandem (two types), PLP, MSG, MCG and RASTA (see references in each subsection below).

In each experiment, a front-end extracts features. Each feature type might have differing numbers of dimensions in each case. The extracted features are both evaluated as is (referred to as RAW), and are also subject to post-processing

stages of mean subtraction (referred to as M), followed by variance normalization (MV), followed by ARMA filtering (MVA) as described in [4, 5]. The back-end in each case uses whole-word HMMs (simulated using GMTK [1] for training and decoding), with 16 emitting states for a word model, 3 states for the silence model, and 1 state for the short-pause model. The observation density is a Gaussian mixture with up to 16 components.

No attempt in this paper is made at a detailed analysis of the form of noise corruption, as this paper deals with *blind* application of MVA. A detailed analysis of MFCC feature distortions in the presence of additive and convolutional noises and how MVA post-processing corrects such corruption is presented in [3].

3 Evaluations

3.1 MFCC

Our first feature type is the ubiquitous MFCC. Being the most common feature extraction method, this will establish a performance reference (or baseline) with which to compare the other features. In our case, each feature vector contains 12 MFCCs enhanced by the zeroth MFCC, along with the delta and double-delta elements (so log energy is not used).

Table 1: Evaluation on MFCC features. The results are listed with respect to three properties: what training set is used (multi-train or clean-train), how noisy the test data is (clean, 0-20 dB or -5 dB), and what post-processing to the features is applied (RAW, M, MV or MVA). The numbers in the table are the *word accuracy rates*.

	multi-train			clean-train		
	clean	0-20dB	-5dB	clean	0-20dB	-5dB
RAW	99.21	88.25	22.27	99.65	54.28	7.23
M	99.48	91.64	30.69	99.72	68.58	3.94
MV	99.33	92.91	41.11	99.66	81.99	19.44
MVA	99.34	93.44	46.49	99.66	85.20	27.24

The MFCC results are presented in Table 1. One can see that MVA improves significantly over RAW. With the 0-20 dB noisy test data, MVA improves 44.2% relative in multi-train case and 67.6% in the clean-train (or mismatched train/test) case. Comparing MV and MVA, the ARMA filtering improves 7.5% relative in the multi-train case and 17.8% in the clean-train case.[1]

3.2 LPC

LPC features represent the quasi-stationary process in a speech analysis window based on an all-pole model of the vocal tract. MFCCs, on the other hand, are derived from a sinusoidal basis expansion of the log spectral energy. Therefore the LPCs and the MFCCs are different, and there are no immediately obvious reasons for them to be corrupted in identical ways in the presence of noise. Nevertheless, the same post-processing has a similar effect. In our experiments, the feature vector contains 12 LPC coefficients with log energy, and with delta and delta-deltas.

Table 2: Evaluation on LPC features.

	multi-train			clean-train		
	clean	0-20dB	-5dB	clean	0-20dB	-5dB
RAW	89.76	63.34	3.90	93.11	21.22	5.24
M	88.76	71.92	15.97	93.52	39.96	6.52
MV	87.19	71.56	1.10	93.39	39.69	4.53
MVA	87.14	73.21	6.06	93.33	42.17	0.94

The results with LPC features are summarized in Table 2. MVA again improves significantly over the RAW features. Specifically, it improves 26.9% in the multi-train case and 26.6% in the clean-train case, with the 0-20 dB

[1] As is convention, all relative improvements reported in this paper are with respect to the word error rate. However, the performance levels are represented using word accuracy rate, a standard for Aurora 2.0 evaluations.

test tasks. Comparing MV and MVA, the ARMA filtering improves 5.8% relative in the multi-train case and 4.1% in the clean-train case.

Overall, the performance level of the LPCs is worse than MFCCs (as is well known). Furthermore, MVA's improvements over RAW are not as significant as the MFCC case.

3.3 LPC-CEPSTRA

The observation that the LPCs perform worse than the MFCCs leads us to evaluate LPC-CEPSTRA. The feature vector contains log energy and 12 LPC cepstral coefficients derived from 12 LPCs, along with their delta and delta-deltas.

Table 3: Evaluation on LPC-Cepstral features.

	multi-train			clean-train		
	clean	0-20dB	-5dB	clean	0-20dB	-5dB
RAW	99.13	87.21	23.92	99.41	56.50	3.61
M	99.20	90.06	32.29	99.59	71.26	7.33
MV	98.88	90.12	31.70	99.52	80.39	17.52
MVA	98.86	90.71	36.43	99.46	82.61	23.97

The results with LPC-CEPSTRA are summarized in Table 3. MVA again improves significantly over the RAW features. Specifically, it improves 27.4% in the multi-train case and 60.0% in the clean-train case, with the 0-20 dB test tasks. Comparing MV and MVA, the ARMA filtering improves 6.0% relative in the multi-train case and 11.3% in the clean-train case.

Overall, the performance level of LPC-CEPSTRA is much better than LPCs, although still not as good as that of the MFCCs. Apparently, applying the discrete cosine transform (DCT) to the LPCs results in features more germane to the back-end HMM with Gaussian mixture densities. In addition, the MVA case shows a better relative improvement over the RAW case in the LPC-CEPSTRA than in the LPCs, especially in the (mismatched) clean-train case. That is, DCT leads to a better baseline and a better relative improvement.

3.4 Tandem-M, 3.5. Tandem-C

Tandem features are one of the best on the Aurora 2.0 corpus. This technique is included in the evaluation of MVA to see how a completely different feature set reacts to MVA post-processing. The feature vector is 24-dimensional corresponding to 24 phone classes, as explained in [6]. In order to extract Tandem features, two neural networks are pre-trained to map from base features (PLP and MSG respectively) to phone posterior probabilities. Features are obtained by application of the trained network without the final network non-linearity, and performing additional processing. In [6], only the case where the networks are trained by multi-train data is investigated (Tandem-M). In this paper, we also investigate the case where the networks are trained using *only* the clean-train data (Tandem-C).

Table 4: Evaluation on Tandem features. Top: nets trained on multi-train data. Bottom: nets trained on clean-train data.

Tand-M	multi-train			clean-train		
	clean	0-20dB	-5dB	clean	0-20dB	-5dB
RAW	99.50	92.87	40.19	99.63	85.84	17.28
M	99.58	93.66	44.31	99.68	90.05	27.79
MV	99.50	93.69	44.48	99.65	90.78	32.06
MVA	99.57	93.68	44.61	99.67	91.15	35.33
Tand-C	multi-train			clean-train		
	clean	0-20dB	-5dB	clean	0-20dB	-5dB
RAW	99.45	89.34	27.57	99.62	71.77	8.69
M	99.57	91.47	40.78	99.65	83.81	20.87
MV	99.51	91.39	40.87	99.68	83.15	17.14
MVA	99.55	91.25	40.60	99.64	83.41	21.68

The results with the Tandem features are summarized in Table 4. The top part shows Tandem-M and the bottom part shows Tandem-C. Overall, Tandem-M is better than Tandem-C, on the noisy test data. This shows that the mismatch in the performance level can be attributed, at least partially, to the trained networks with mismatched data. Furthermore, MVA post-processing is capable of reducing the effect of mismatch to some extent. This is evidenced by the observation that the relative difference in the two cases is larger with RAW features than with post-processed features. The improvement is mostly accounted for by M (mean subtraction). Finally, it is interesting to look at the case where the networks are trained using only the clean-train data. While such networks have not been exposed to noisy data, the discriminative power of the network's training procedure results in RAW Tandem-C features that still perform better than RAW MFCCs.

3.6 PLP

PLP features incorporate a notion of human auditory characteristics such as the equal-loudness curve and the power law of hearing. The PLP features used here are based on mel-frequency filter banks (and are generated via HTK).

Table 5: Evaluation on PLP features.

	multi-train			clean-train		
	clean	0-20dB	-5dB	clean	0-20dB	-5dB
RAW	99.40	89.81	28.53	99.65	62.05	6.98
M	99.46	92.47	37.26	99.69	71.35	5.84
MV	99.29	93.01	44.07	99.61	83.71	21.48
MVA	99.28	93.20	47.66	99.67	85.68	28.40

The results of the PLP features are summarized in Table 5. MVA improves significantly over the RAW features. Specifically, it improves 33.3% in the multi-train case and 62.3% in the clean-train case, with the 0-20 dB test tasks. Comparing MV and MVA, the ARMA filtering improves 2.7% relative in the multi-train case and 12.1% in the clean-train case. Here it is interesting to compare the MFCC and PLP features. Without any feature processing, PLP is significantly better than MFCC (89.81% vs 88.25% in multi-train and 62.05% vs 54.28% in clean-train). However, with MVA feature processing, the disadvantage of MFCC greatly decreases (93.20% vs 93.44% in multi-train and 85.68% vs 85.20% in clean-train).

3.7 Modulation-filtered Spectrogram

The modulation-filtered spectrogram (MSG) [7] computes the 4-Hz spectral energy of filtered modulation amplitude of each critical band. The main idea is to "focus on the elements in the signal encoding phonetic information", which changes at a typical rate between 0-8 Hz corresponding to articulatory gestures. The signal processing steps in MSG have evolved from their original setting for improved performance [10]. Here the features are the so-called "msg3" features extracted by the SPRACHcore software release from ICSI.[2]

Table 6: Evaluation on MSG features.

	multi-train			clean-train		
	clean	0-20dB	-5dB	clean	0-20dB	-5dB
RAW	97.42	86.37	18.05	98.90	56.06	-2.19
M	97.39	86.82	25.70	98.86	66.41	6.29
MV	97.50	86.29	15.98	98.59	57.78	6.08
MVA	97.74	86.53	19.93	98.84	60.19	6.40

The results of MSG features are presented in Table 6. Without any post-processing, the performance level is very similar to that of MFCC. With post-processing, there are no significant performance gains. Note that in the processing

[2]We thank Brian Kingsbury, the original MSG designer, for instructions and discussions.

of "msg3" features, an on-line normalization of mean and variance has already been implemented.[3] Furthermore, the modulation amplitude filtering is essentially a *hi-reject* filtering which is similar to ARMA. Also note that to be used in an HMM recognizer the MSG features are often further processed by neural networks, which is not the case here.

3.8 Modulation Cross-CorreloGram

The Modulation Cross-CorreloGram (MCG) [2] features are based on the cross-correlation of the magnitude sequences in different spectral channels. A two-dimensional DCT is further applied to the cross-correlation matrix and the lowest-order 6×6 sub-matrix of the DCT output constitutes the final 36-dimensional feature vector. MCG features are a "delta-like" feature, and are meant to be used together with MFCCs [2] or some other base feature. Herein, however, we simply evaluate MCG features alone in order to gauge the affect of MVA.

Table 7: Evaluation on MCG features.

	multi-train			clean-train		
	clean	0-20dB	-5dB	clean	0-20dB	-5dB
RAW	90.29	65.85	15.94	90.89	57.77	10.29
M	89.30	65.64	15.41	92.32	59.07	8.61
MV	91.00	66.20	14.36	93.44	60.98	7.40
MVA	90.73	66.16	15.62	93.02	60.88	9.99

The results of MCG are presented in Table 7. The post-processing introduces only minor improvements over the RAW case, presumably because MCGs are already highly normalized.

3.9 RASTA

RelAtive SpecTrA (RASTA) [8] is a filtering technique applied in a domain of the (compressed) critical-band spectral envelops. It is designed to remove the slow-varying environmental variations and the fast-varying artifacts. The 39-dimensional plain RASTA-PLP (instead of j-rasta or log-rasta) is used in this evaluation.

Table 8: Evaluation on RASTA features.

	multi-train			clean-train		
	clean	0-20dB	-5dB	clean	0-20dB	-5dB
RAW	99.50	90.95	30.97	99.72	61.59	7.93
M	99.52	92.28	34.06	99.76	73.16	5.70
MV	99.18	93.18	45.18	99.70	84.20	21.95
MVA	99.34	93.25	48.37	99.66	85.40	28.15

The results are presented in Table 8. Without any feature post-processing, the performance level is better than that of MFCC in 0-20 dB test data. With MVA post-processing, the performance levels are virtually identical, as with PLP. Specifically, the ARMA filter introduces a significant performance boost with MFCC but only a minute gain with RASTA. To a certain degree, the RASTA filtering and the ARMA filtering are somewhat redundant (but exist in different stages of the feature extraction procedure).

4 Comparison across feature sets

In Section 3, the results are presented individually. As an objective of this paper is also discover the characteristics of features that work well with MVA, comparison across feature sets is summarized in this section. A summary over all features is given in Figure 1. These figures clearly show the absolute performance levels of different feature sets.

[3] The application of mean subtraction (and variance normalization) is not entirely redundant, as the post-processing is per-utterance. Also after the variance normalization the zero-mean property no longer holds.

Generally speaking, the rankings of feature sets with respect to the performance level are quite similar in different tasks (i.e. different combinations of train/test data). The relative ranking is:

$$1 \approx 3 \approx 4 \approx 5 \approx 6 \approx 9 \succeq 7 \succ 8 \succeq 2$$

(using the same feature enumerations as Figure 1 and Section 3.).

The relative improvements of MVA over RAW are summarized in Table 9. Generally speaking, with clean test data, the performance may inappreciably degrade with MVA, but with noisy and/or mismatched data, the performance is boosted by MVA. The feature sets can be divided into three classes based on the performance gains on 0-20 dB test data when MVA is applied.

- **Feature sets with substantial performance gains.** The MFCC, LPC-CEPSTRA, PLP and RASTA features belong to this category. On average MVA achieves 63% relative improvements in clean-train and 33% in multi-train cases, on top of decent baseline RAW results.
- **Feature sets with medium performance gains.** The LPCs and Tandem features (including Tandem-M and Tandem-C) belong to this category. On average MVA achieves 35% relative improvements in clean-train and 18% in multi-train cases.
- **Feature sets with minute performance gains.** The MCG and MSG belong to this category. On average MVA achieves 8% relative improvements in clean-train and 1% in multi-train cases.

Table 9: Relative improvements of *word error rates* of MVA over RAW on all features.

	multi-train			clean-train		
	clean	0-20dB	-5dB	clean	0-20dB	-5dB
MFCC	16.46	44.17	31.16	2.86	67.63	21.57
LPC	-25.59	26.92	2.25	3.19	26.59	-4.54
LPC-C	-31.03	27.37	16.44	8.47	60.02	21.12
Tand-M	14.00	11.36	7.39	10.81	37.50	21.82
Tand-C	18.18	17.92	17.99	5.26	41.23	14.22
PLP	-20.00	33.27	26.77	5.71	62.27	23.03
MSG	12.40	1.17	2.29	-5.45	9.40	8.41
MCG	4.53	0.91	-0.38	23.38	7.36	-0.33
RASTA	-32.00	25.41	25.21	-21.43	61.99	21.96

5 Summary

In this paper, the MVA feature processing scheme is applied to feature sets of different natures. The results show that MVA works well in the majority of cases, especially in highly noisy and/or mismatched (train/test) data. It is also shown that the performance gain of certain noise-robust features can be improved by performing temporal integration in the final stage, as is done by MVA processing, effectively utilizing information from time spans longer than a typical analysis window. Our working hypothesis is that appropriately extracting information across multiple analysis windows is a fundamental property of noise-robust ASR systems, at least in small vocabulary. This agrees with a recent research trend to integrate informations at different time scales [9]. MVA, like many other features, does this in a simple and effective way, and is applicable to any feature extraction method.

In summary, MVA is an effective and low-cost processing technique for noise-robustness. Whenever a new feature is proposed, applying MVA might lead to a further performance gain!

References

[1] J. Bilmes and G. Zweig. The Graphical Models Toolkit: An open source software system for speech and time-series processing. *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2002.

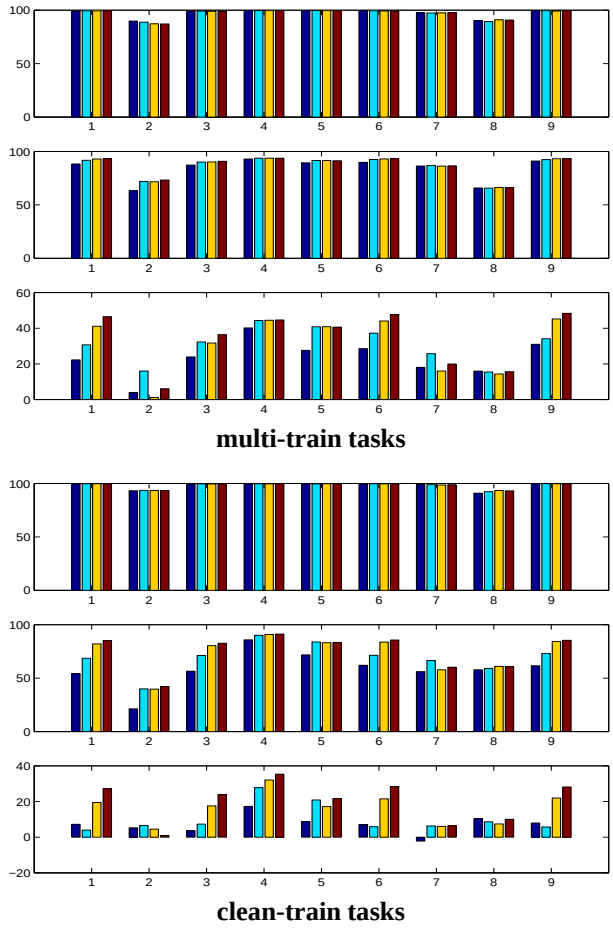


Figure 1: The comparison of feature sets. The top half summarizes the multi-train tasks while the bottom half summarizes the clean-train tasks. In each half, the three blocks correspond to clean, 0-20 dB and -5 dB test tasks. Within each block, the abscissa is the enumerated feature set (1 = MFCC; 2 = LPC; 3 = LPC-CEPSTRA; 4 = Tandem-M; 5 = Tandem-C; 6 = PLP; 7 = MSG; 8 = MCG; 9 = RASTA), while the ordinate is the word accuracy rates. Within each feature set, the four bars in a bar group correspond to RAW, M, MV and MVA from left to right.

[2] J. A. Bilmes. Joint distributional modeling with cross-correlation based features. In *Proceedings of IEEE ASRU Workshop*, pages 148–155, 1997.

[3] C.-P. Chen and J. Bilmes. Mva processing of speech features. Technical Report UWEETR-2003-0024, University of Washington, Dept. of EE, 2003.

[4] C.-P. Chen, J. Bilmes, and K. Kirchhoff. Low-resource noise-robust feature post-processing on Aurora 2.0. In *Proceedings of International Conference on Spoken Language Processing (ICSLP)*, pages 2445–2448, 2002.

[5] C.-P. Chen, K. Filali, and J. Bilmes. Frontend post-processing and backend model enhancement on the Aurora 2.0/3.0 databases. In *Proceedings of International Conference on Spoken Language Processing (ICSLP)*, pages 241–244, 2002.

[6] D. Ellis and M. Gomez. Investigations into tandem acoustic modeling for the Aurora task. In *European Conference on Speech Communication and Technology (EuroSpeech)*, pages 189–192, 2001.

[7] S. Greenberg and B. Kingsbury. The modulation spectrogram: in pursuit of an invariant representation of speech. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 1647–1650, 1997.

[8] H. Hermansky and N. Morgan. RASTA processing of speech. *IEEE Transactions on Speech and Audio Processing (SAP)*, 2(4):578–589, October 1994.

[9] H. Hermansky and S. Sharma. Temporal patters (TRAPS) in ASR of noisy speech. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1999.

[10] B. E. D. Kingsbury. *Perceptually Inspired Signal-processing Strategies fro Robust Speech Recognition in Reverberant Environments*. PhD thesis, University of California, Berkeley, 1998.