

SPEECH FEATURE SMOOTHING FOR ROBUST ASR

Chia-Ping Chen Jeff Bilmes

Department of Electrical Engineering
University of Washington
Seattle, WA 98195-2500
{chiaping,bilmes}@ee.washington.edu

Daniel P. W. Ellis

Department of Electrical Engineering
Columbia University
New York 10027
dpwe@ee.columbia.edu

ABSTRACT

In this paper, we evaluate smoothing within the context of the MVA (mean subtraction, variance normalization, and ARMA filtering) 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 even though it is computationally expensive. Herein, MVA is applied to many acoustic feature extraction methods, and is evaluated using Aurora 2.0. We evaluate MVA post-processing on MFCCs, LPCs, PLPs, RASTA, Tandem, Modulation-filtered Spectrogram, and Modulation Cross-Correlation features. We conclude that while effectiveness does depend on the extraction method, the majority of features benefit significantly from MVA, and the smoothing ARMA filter is an important component. It appears that the effectiveness of normalization and smoothing depends on the domain in which it is applied, being most fruitfully applied just before being scored by a probabilistic model. Moreover, since it is both effective and simple, our ARMA filter should be considered a candidate method in most noise-robust speech recognition tasks.

1. INTRODUCTION

Lack of noise robustness is an obstacle that automatic speech recognition (ASR) systems must overcome to be more widely used. Indeed, one small step for noise-robustness could be one giant leap for ASR's viability. But even given the many recent and novel noise-robust techniques, there is still considerable room for improvement as indicated by the fact that human performance in noise is still far better.

Besides being accurate, ASR systems also must have tolerable computational and memory demands, especially on portable devices. MVA post-processing [1, 2] is an effective noise-robust technique on small-vocabulary ASR tasks. It achieves performance on par with most effective noise-robust techniques, but without any significant computational increase — it is therefore applicable to low-power and/or portable devices.

The main innovative idea behind MVA is a smoothing ARMA filter and the domain in which it is applied (just before the Gaussians). In this paper, the effectiveness of MVA and in particular its smoothing filter is investigated on a wide variety of speech features. Our goal is to discover the effectiveness of smoothing in particular, and MVA in general on these features. Our interest is to characterize those features that combine well with MVA. Our conclusion will be that since the ARMA filter is both effective and easy to apply, it should be a candidate technique for any

C.-P. Chen and J. Bilmes were supported by NSF grants IIS-0093430 and ITR/RC-0086032. D. Ellis was supported by NSF IIS-0238301.

noise-robust ASR system. This paper is primarily an empirical evaluation, but a detailed theoretical analysis of additive and convolutional noise under MFCCs and the MVA corrective ability is presented in [4, 12].

2. SETUP

We investigate a broad range of features in this work, namely MFCCs, LPCs, LPC-CEPSTRA, Tandem features (two types), PLP, MSG, MCG, and RASTA (references below). In each experiment, a front-end extracts "RAW" features (which might have differing dimensionality) that are evaluated as is, and are also subjected to the post-processing stages of mean subtraction (referred to as "M"), followed by variance normalization (MV), and followed by ARMA filtering (MVA) as described in [1, 2]. The back-end in each case uses whole-word HMMs (simulated using GMTK [3] 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. Results are evaluated on the Aurora 2.0 corpus under different SNR ratios and two training/testing conditions: multi-train, or *matched* training/testing, and clean-train, or *mis-matched* training/testing. The mis-matched (clean-training) condition is considered more difficult and realistic.

3. EVALUATIONS

3.1. MFCC

Our first feature set is the ubiquitous MFCC. Being the most common method, we will herein establish a performance baseline with which to compare the other methods. In our experiments, each feature vector contains 12 MFCCs enhanced by the zeroth MFCC C_0 , along with delta and double-deltas (log energy is not used [4]).

Table 1. MFCC Evaluation. The results are listed with respect to three properties: the training set used (matched or mis-matched training/testing conditions), noise SNR (clean, 0-20 dB or -5 dB), and post-processing applied (RAW, M, MV or MVA). The numbers given are *word accuracy rates*.

	matched train/test			mis-matched train/test		
	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

Results are presented in Table 1. MVA improves significantly over RAW. On the 0-20 dB noisy test data, MVA improves 44.2% relative in the matched and 67.6% in the mis-matched case. Comparing MV and MVA, the smoothing ARMA filtering improves 7.5% relative in the matched and 17.8% in the mis-matched case.¹ (Note that the “mis-matched” case, trained only on clean data, actually outperforms “matched” for the clean test condition.)

3.2. LPC

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

Table 2. Evaluation on LPC features.

	matched train/test			mis-matched train/test		
	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

LPC results are summarized in Table 2. MVA again improves significantly over the RAW features. Specifically, it improves 26.9% in the matched case and 26.6% in the mis-matched case, with the 0-20 dB test tasks. Comparing MV and MVA, the smoothing ARMA filtering improves 5.8% relative in the matched case and 4.1% in the mis-matched case. Overall, the LPCs perform worse than MFCCs (as is well known). Furthermore, MVA’s improvements over RAW is 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.

	matched train/test			mis-matched train/test		
	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

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

¹As is convention, all reported relative improvements 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.

Clearly, applying the discrete cosine transform (DCT) to the LPCs results in features better matched to the back-end Gaussian-mixture HMM. 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 mis-matched case. That is, DCT leads to a better baseline and a better relative improvement. Note, however, that while the performance of LPC-CEPSTRA is much better than LPCs, it is still not as good as the MVA-MFCCs.

3.4. Tandem-M, 3.5. Tandem-C

Tandem features [5] have performed very well on Aurora 2.0. This technique is included in our analysis to investigate how completely different features react to MVA post-processing. The feature vector is 24-dimensional corresponding to 24 phone classes, as explained in [5]. In order to extract Tandem features, two neural networks are pre-trained to map from base features (PLP and MSG respectively, see below) to phone posterior probabilities. Features are obtained by application of the trained network without the final network non-linearity, and performing additional processing. In [5], the case where the networks are trained by multi-train data only 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) leading to a double mis-match.

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

Tand-M	matched train/test			mis-matched train/test		
	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	matched train/test			mis-matched train/test		
	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

Tandem results 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. Thus, a mismatch between the data used to train the net and the test condition does indeed hurt overall performance, as expected. However, 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 [11] utilize aspects of the human auditory system such as the equal-loudness curve and the power law of hearing. Here, PLP features are based on mel-frequency filter banks.

Table 5. Evaluation on PLP features.

	matched train/test			mis-matched train/test		
	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

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

3.7. Modulation-filtered Spectrogram

The modulation-filtered spectrogram (MSG) [6] computes the 4-Hz spectral energy of filtered modulation amplitude of each critical band. The 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 [7]. Here we use the “msg3” features extracted by the SPRACHcore software release from ICSI.²

Table 6. Evaluation on MSG features.

	matched train/test			mis-matched train/test		
	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

Our MSG features are presented in Table 6. Without any post-processing, the performance level is very similar to MFCCs. With post-processing, there is no significant performance gain. Note that msg3-features already include an on-line step of mean and variance normalization.³ Furthermore, the modulation amplitude filtering is essentially *low-pass* 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 [5], e.g. in a Tandem setup as described above.

3.8. Modulation Cross-CorreloGram

The Modulation Cross-CorreloGram (MCG) [8] features are based on the cross-correlation of the magnitude sequences in different spectral channels. A two-dimensional DCT is further applied to

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

³The application of mean subtraction (and variance normalization) is not entirely redundant, however, as MVA post-processing is per-utterance, and after the msg3 variance normalization the local zero-mean property no longer holds.

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-only” feature, and are designed to be used together with static MFCCs [8] or some other base feature. Herein, however, we simply evaluate MCG features alone to assess the MVA’s effect.

Table 7. Evaluation on MCG features.

	matched train/test			mis-matched train/test		
	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 which themselves are not very good. The lack of MVA effect is presumably because MCGs are already highly normalized.

3.9. RASTA

RelAtive SpecTrA (RASTA) [9] is a filtering technique applied in a domain of the (compressed) critical-band spectral envelopes. 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.

	matched train/test			mis-matched train/test		
	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

Results are presented in Table 8. Without any post-processing, the performance level is better than that of MFCCs 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 they exist in different stages of the feature extraction procedure. Moreover, the smoothing ARMA filter of MVA introduces significant gains, particularly in the high-noise and mis-matched training/testing cases.

4. COMPARISON ACROSS FEATURE SETS

The two objectives of this paper are to show that our smoothing ARMA filter decreases error-rate in general, and also to determine characteristics of features that work well with MVA. This section therefore summarizes the above results in Figure 1.

Generally speaking, the feature sets have the same rank irrespective of the particular task (i.e. noise level, different train/test situation). We approximately linearize the ranking as: $1 \approx 3 \approx 4 \approx 5 \approx 6 \approx 9 \geq 7 > 8 \geq 2$ (using the enumerations from Figure 1 and Section 3.). With clean test data, the performance either stays the same or negligibly degrades with MVA, but under noisy and/or mismatched training/testing conditions (the more realistic settings), the performance is boosted by MVA. Moreover, Figure 1

shows a consistent improvement by the smoothing ARMA filter, again particularly in the noisy and mis-matched cases.

We divide the features into three classes based on their performance improvements under MVA:

Feature sets with substantial performance gains: MFCC, LPC-CEPSTRA, PLP and RASTA features, where on average MVA achieves a 63% relative improvement in the mis-matched and 33% in matched cases, over RAW results.

Feature sets with medium performance gains: The LPCs and Tandem features (both Tandem-M and Tandem-C), where on average MVA achieves a 35% relative improvement in the mis-matched and an 18% improvement in the matched cases.

Feature sets with minute performance gains: The MCG and MSG features, where on average MVA achieves an 8% relative improvement in mis-matched and a 1% in matched cases.

As can be seen, the features that already have a late-term built-in smoothing and normalization component (MCG, MSG) interact most poorly with MVA, while the features that in general have the least such internalization (MFCCs, LPCs) interact best with MVA. Our contention is that MVA performs well not so much because it performs normalization and smoothing, but rather because of where it does so, specifically just before the Gaussian densities. In other words, it appears that smoothing and normalization are most successfully applied just before being scored by a probabilistic model.

5. SUMMARY

In this paper, we consider speech feature smoothing for robust ASR by evaluating the MVA post-processing on many disparate speech feature representations. Our results show that MVA works well in the majority of cases, especially in highly noisy and/or mismatched (train/test) data, and that our smoothing ARMA filter almost always helps. We also show MVA's performance gain with native noise-robust features by performing smoothing in the final stage, just before the scoring of features by a Gaussian-mixture HMM. Our working hypothesis, supported by further experiments [12], is that normalization and smoothing is a fundamental property of noise-robust ASR systems, at least in small vocabulary tasks, but it is important to do such operations in the right domain, namely just before probabilistic scoring. MVA, like many other techniques, does this in a computationally simple and effective way, and is applicable to any feature extraction method.

6. REFERENCES

- [1] C.-P. Chen, J. Bilmes and K. Kirchhoff, "Low-Resource Noise-Robust Feature Post-Processing on Aurora 2.0", pp. 2445-2448, Proceedings of ICSLP 2002.
- [2] C.-P. Chen, K. Filali and J. Bilmes, "Frontend Post-Processing and Backend Model Enhancement on the Aurora 2.0/3.0 Databases", pp. 241-244, Proceedings of ICSLP 2002.
- [3] J. Bilmes and G. Zweig, "The Graphical Models Toolkit: An Open Source Software System for Speech and Time-Series Processing", Proceedings of ICASSP 2002.
- [4] C.-P. Chen and J. Bilmes, "MVA Processing of Speech Features", University of Washington Electrical Engineering Technical Report, UWEETR-2003-0024.
- [5] D. Ellis and M. Gomez, "Investigations into Tandem Acoustic Modeling for the Aurora Task", pp. 189-192, Proceedings of Eurospeech 2001.

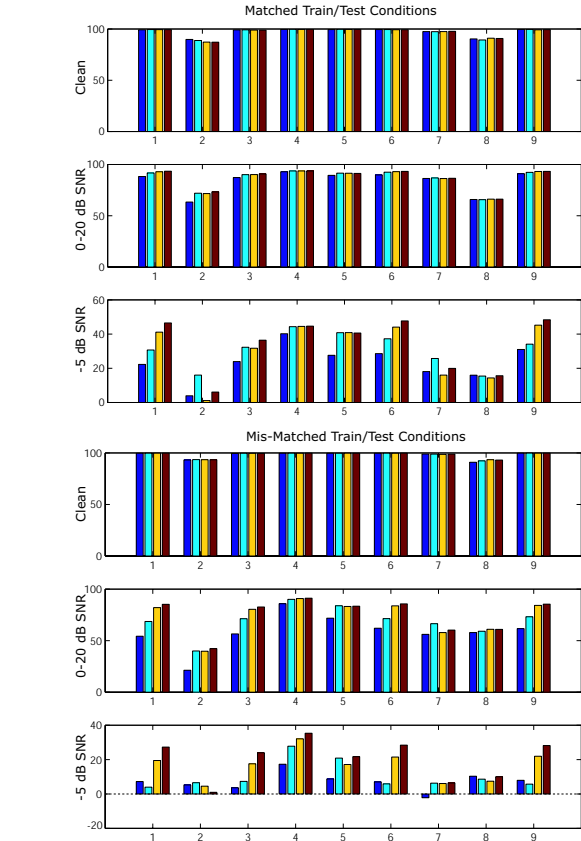


Fig. 1. The comparison of feature sets. Within each graph, the x-axis is the enumerated feature set (1 = MFCC; 2 = LPC; 3 = LPC-CEPSTRA; 4 = Tandem-M; 5 = Tandem-C; 6 = PLP; 7 = MSG; 8 = MCG; 9 = RASTA), while the y-axis is the word accuracy rates. Within each feature set, the four bars in a bar group correspond to RAW, M, MV and MVA from left to right.

- [6] S. Greenberg and B. Kingsbury, "The modulation spectrogram: in pursuit of an invariant representation of speech", pp. 1647-1650, Proceedings of ICASSP 1997.
- [7] B. E. D. Kingsbury, "Perceptually Inspired Signal-processing Strategies for Robust Speech Recognition in Reverberant Environments", PhD Thesis, University of California, Berkeley, 1998.
- [8] J. A. Bilmes, "Joint Distributional Modeling with Cross-Correlation based Features", pp. 148-155, Proceedings of IEEE ASRU Workshop 1997.
- [9] H. Hermansky, N. Morgan, "RASTA Processing of Speech", IEEE Transactions on Speech and Audio Processing, Vol. 2, No. 4, pp. 578-589, Oct. 1994.
- [10] H. Hermansky and S. Sharma, "Temporal Patters (TRAPS) in ASR of Noisy Speech", Proceedings of ICASSP 1997.
- [11] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech", JASA 1990, 87(4), April, pp. 1738-1752.
- [12] C.-P. Chen, "Noise Robustness in Automatic Speech Recognition", PhD Thesis, University of Washington, Seattle, 2004.