



A Fast Approach to Psychoacoustic Model Compensation for Robust Speaker Recognition in Additive Noise

Ashish Panda

TCS Innovation Labs-Mumbai,
Yantra Park, Thane, Maharashtra, India, 400610.

ashish.panda@tcs.com

Abstract

This paper addresses the problem of speaker verification in the presence of additive noise. We propose a fast implementation of Psychoacoustic Model Compensation (Psy-Comp) scheme for static features along with model domain mean and variance normalization for robust speaker recognition in noisy conditions. The proposed algorithms are validated through experiments on noise corrupted NIST-2000 speaker recognition database. We show that the Psy-Comp scheme along with model domain mean and variance normalization provide significant performance gain compared to the Vector Taylor Series (VTS) scheme and feature domain cepstral mean and variance normalization scheme. Moreover, the computational cost of the proposed method is significantly less than the VTS scheme.

Index Terms: Speaker Recognition, Additive Noise, Psychoacoustic Compensation

1. Introduction

Channel and environment mismatches are among the major factors that adversely affect the performance of a speaker verification system and hence they have been the subject of a significant number of research efforts. Approaches have been proposed to deal with the mismatch conditions in various domains, such as feature domain, model domain and score domain.

Cepstral Mean and Variance Normalization (CMVN) is a simple and effective feature domain technique to deal with mismatch conditions [1]. Remarkable robustness against channel mismatch has been reported through the use of i-vector technique [2]. Spectral Subtraction [3], Parallel Model Combination (PMC) [4] and Posterior Union Model (PUM) [5] have been shown to provide robustness in environmental mismatch conditions. Efforts have also been made to impart environmental robustness in i-vector setup in [6] through the Vector Taylor Series (VTS) expansion. VTS is computationally expensive and hence simplified VTS (sVTS) has been proposed in [7]. However, even the sVTS technique is computationally demanding as compared to the PMC.

Psychoacoustic Model Compensation (Psy-Comp) method for Gaussian Mixture Model - Universal Background Model (GMM-UBM) setup, has been proposed in [8]. The PMC and the VTS use a corruption function which assumes the noisy speech in the spectral magnitude domain to be the sum of the clean speech and the noise (both in the spectral magnitude domain). The Psy-Comp, on the other hand, uses a corruption function which determines the noisy speech based on the masking thresholds of the clean speech and the noise. Psy-Comp has been shown to outperform PMC and PUM in [8]. However, it has two major limitations. First, it does not consider the channel

effect and hence its applicability is limited. Second, the complex corruption function makes it very difficult to estimate the compensated model variances accurately. Although, channel effect has been considered with Psy-Comp in [9] by integrating a model domain cepstral mean subtraction, cepstral variance normalization has not been considered.

In this paper, we propose a modified Psy-Comp technique and a model-domain CMVN, which addresses the above two limitations. We have simplified the psychoacoustic corruption function that provides for easy estimation of the compensated model variance and also greatly reduces the computational complexity of the original algorithm. While the earlier methods employed a semi data-driven approach to variance estimation, the same can be accomplished, now, with model and noise parameters, rather than resorting to actual noise observation. Since traditional feature domain CMVN is incompatible with Psy-Comp, we have also proposed a model domain CMVN, which imparts channel robustness to the system. We have demonstrated the performance of the proposed algorithms using the GMM-UBM approach of speaker recognition. Although we have not addressed the proposed algorithms' applicability to i-vector approach in this work, it can be achieved along the lines of [6] and [7] and is one of the goals of our current efforts.

The rest of the paper is organized as follows. Section 2 describes the modified psychoacoustic compensation scheme. Section 3 describes the proposed model domain CMVN operation. We describe the overall algorithm in Section 4. Section 5 deals with the experiments and Section 6 concludes this paper.

2. Modified Psychoacoustic Compensation Scheme

The Psy-Comp algorithm assumes the availability of clean training speech. This is a reasonable assumption, since training is a one-time affair and can be accomplished in a controlled environment for most of the applications, other than forensics. It also assumes availability of the noise statistics during the verification phase. This may be achieved by employing a second microphone away from the speakers' mouths. Given that many of the current mobile devices come with dual microphones, this assumption is also reasonable. One of the microphones in the mobile device can be placed in such a way that it can collect the noise samples. These same assumptions are also made for the popular PMC algorithm of model compensation. The goal of the Psy-Comp scheme is to use the noise observations during the verification phase to compensate the parameters of the clean training speech so as to reduce the mismatch between the training and test utterances. This is accomplished by a psychoacoustic corruption function, which is a function of clean model

mean and noise observations, both in the mel-filter-output (mel-spectral) domain.

We will use the following notations in this section. The clean model parameters are denoted as $\{w_i, \bar{\mu}_{xi}^c, \Sigma_{xi}^c\}$, $1 \leq i \leq M$, with the superscript c signifying the cepstral domain and subscripts x and i signifying the clean speech and the component index respectively. The observed noise parameters during the verification phase are denoted as $\{\bar{\mu}_n^c, \Sigma_n^c\}$, with the subscript n signifying the noise. We will denote the clean model parameters, noise parameters and the compensated model parameters in the mel-filter-output (mel-spectral) domain as $\{w_i, \bar{\mu}_{xi}, \Sigma_{xi}\}$, $\{\bar{\mu}_n, \Sigma_n\}$ and $\{w_i, \bar{\mu}_{yi}, \Sigma_{yi}\}$ respectively. The absence of superscript signifies the mel-spectral domain and the subscript y signifies the compensated model. The compensated model parameters in the cepstral domain are denoted as $\{w_i, \bar{\mu}_{yi}^c, \Sigma_{yi}^c\}$.

The first step in Psy-Comp is to transform the model parameters from the Mel-Frequency Cepstral Coefficient (MFCC) domain to the mel-filter output domain. The equations used for this transformation can be found in [8]. The second step is to compute the central frequencies of the mel-filters in Bark scale. The Bark value of the central frequency of mel-filter f is denoted here as h_f . For each element of the clean model mean vectors in the mel-filter-output domain, compute the masking thresholds as follows:

$$T_{xif} = 20 \log_{10}(\mu_{xif}) - 0.275 \cdot h_f - 6.025 \quad (\text{dB}) \quad (1)$$

In the above equation, μ_{xif} is an element of $\bar{\mu}_{xi}$ and f signifies the mel-filter number. The compensated model mean is computed as:

$$\mu_{yif} = \mu_{xif} + \mu_{nif} - 10^{\frac{T_{xif}}{20}} \quad (2)$$

The compensated model variance is computed as:

$$\sigma_{yifg}^2 = \omega_{if} \cdot \omega_{ig} \cdot \sigma_{xifg}^2 + \sigma_{nfg}^2 \quad (3)$$

In the above equation, σ_{yifg}^2 is an individual element of the covariance matrix Σ_{yi} , with f and g standing for the row and the column of the covariance matrix. Similarly, σ_{xifg}^2 and σ_{nfg}^2 are the individual elements of the covariance matrices Σ_{xi} and Σ_n , respectively. The ω is a weighting factor and is given by:

$$\omega_{if} = \frac{\mu_{xif} - 10^{\frac{T_{xif}}{20}}}{\mu_{xif}} \quad (4)$$

It should be noted that Equations (2) through (4) have been modified from the corruption equations suggested in [8] and [9]. Only the noise parameters have been used for the compensation instead of the observed noise vectors. In [8] and [9], a semi data-driven approach has been used to estimate the compensated model parameters. The clean model mean vectors were corrupted by individual noise observations and the compensated model statistics were derived from data obtained after individual corruption. This induced significant latency. In the new approach, as described in Equations (2) and (3), only the noise parameters are used, which is much faster. Here, we have also ignored the masking effect of the noise, which simplified the equation for the compensated model variance estimation.

Contrast the compensation scheme proposed here to the additive compensation scheme proposed by PMC and VTS algorithms. According to the PMC and VTS, the spectral magnitude of the noisy speech is the sum of spectral magnitudes of the clean speech and the noise. The addition is performed without taking into consideration the ‘‘audibility’’ of the noise signal.

The proposed algorithm, however, takes the level of the noise into consideration and adds only the ‘‘audible’’ portion of the noise to the clean speech signal. According to Equation (2), the ‘‘audible’’ noise is defined as $\mu_{nif} - 10^{\frac{T_{xif}}{20}}$.

3. Model Domain Cepstral Mean and Variance Normalization

CMVN technique has been a very popular feature normalization technique for robust speaker and speech recognition and it has traditionally been performed in the feature space. Although, feature space CMVN provides significant improvements in performance, it applies a transformation to the feature vectors, which is difficult to model in the mel-filter-output domain. Also, the transformation applied by feature space CMVN is utterance specific and not uniform across all utterances. This renders it incompatible with the Psy-Comp. As described in Section 2, Psy-Comp compares the masking threshold derived from mel-spectral magnitude of the clean model mean with the mel-spectral magnitude of the noise mean. If feature space CMVN is applied, then the model mean will be transformed according to training utterance and the noise mean will be transformed according to the test utterance. This difference in transformation will render the comparison meaningless. Therefore, we propose a model domain CMVN, which transforms the model parameters, instead of the feature vectors.

It should be noted that in this section, although we have not used superscript c , all the parameters are expressed in the MFCC domain. The proposed model domain CMVN (MCMVN) works as follows. During the training, the models are trained using unaltered MFCC vectors (i.e., no CMVN is performed on the training utterances). During verification, feature space CMVN is performed on the test utterances. For a model $\lambda = \{w_i, \bar{\mu}_i, \Sigma_i\}$, $1 \leq i \leq M$, assuming diagonal covariance matrices, we compute a mean vector \bar{m} and a standard deviation S as follows:

$$\bar{m} = \frac{1}{\sum w_i} \sum_{i=1}^M w_i \bar{\mu}_i \quad (5)$$

$$S^2 = \sum_{i=1}^M w_i (\Sigma_i + \text{diag}[\bar{\mu}_i \bar{\mu}_i^T]) - \text{diag}[\bar{m} \bar{m}^T] \quad (6)$$

In the above equation, the superscript T signifies the transpose of the matrix and diag denotes the diagonal matrix. The MCMVN transformed model parameters are then estimated as follows:

$$\bar{\mu}_i = S^{-1}[\bar{\mu}_i - \bar{m}] \quad (7)$$

$$\bar{\Sigma}_i = (S^2)^{-1}[\Sigma_i] \quad (8)$$

In the above equations, $\bar{\mu}_i$ and $\bar{\Sigma}_i$ are the mean and covariance matrix of the MCMVN transformed model. The test utterance with feature space CMVN is scored with the transformed model $\bar{\lambda} = \{w_i, \bar{\mu}_i, \bar{\Sigma}_i\}$, $1 \leq i \leq M$.

The mean \bar{m} and the standard deviation S computed above, can be intuitively viewed as the mean and standard deviation of the training utterance. Instead of computing these values from training utterance, we have computed them from the model parameters. Another way to implement the MCMVN might be to compute the mean and standard deviation from the training utterance and store in memory. The same can be retrieved during verification phase to normalize model parameters. But, this will not work if the models are trained using more than one utterance, since the mean and standard deviation are computed per

utterance. The proposed MCMVN, however, will work even if models are trained from multiple training utterances.

4. Overall Psy-Comp and MCMVN Scheme

The speaker models and UBM are trained using clean speech. MFCC feature vectors are computed and no normalizations are applied for training utterances. For training, we follow the GMM-UBM scheme described in [10]. For the recognition phase, the algorithm is as follows:

1. Read the noisy test utterance and compute MFCC followed by feature space CMVN.
2. The noise frames are processed in the same way as the speech frames and noise mean and variance are computed in the cepstral domain. No CMVN is applied to the noise frames.
3. Transform the clean model parameters from MFCC domain to mel-filter-output domain.
4. Transform the noise parameters from MFCC domain to mel-filter-output domain.
5. Compute the central frequencies of the mel-filters and convert to Bark scale.
6. Compute the masking thresholds of the clean model mean vectors using Equation (1).
7. Compute the compensated model mean and variances using Equations (2) and (3).
8. Transform the compensated parameters from mel-filter-output domain to MFCC domain.
9. Compute the model mean and standard deviation from the compensated parameters using Equations (5) and (6).
10. Compute MCMVN transformed model parameters using Equations (7) and (8).
11. Score the test utterance against the MCMVN transformed parameters.

It can be seen here that, unlike the VTS, the proposed method does not resort to the Expectation Maximization algorithm for estimation of noise and channel parameters. This makes the proposed method computationally more efficient than the VTS.

5. Experiments

Experiments were conducted on the male portion of NIST-2000 SRE database. Training speech were kept unaltered, while the test utterances were corrupted with Factory-1 noise and real-life street noise. The Factory-1 noise is from the NOISEX-92 noise database. The street noise was collected from the streets of Mumbai, India, using mobile phones over the course of a few months (The noise collection is part of our efforts to build a noise database with realistic noise from Indian scenarios and this noise database will be released at a suitable time). The street noise samples are highly time-varying in nature and consists of engine noises from various vehicles, blaring horns and blowing winds etc. The spectrogram of a typical street noise sample is given in Figure 1. Five samples of the street noise were selected and each test utterance was corrupted with one of the five samples chosen randomly. The noise signals were added at SNRs of 10dB and 5dB. The SNR was computed by taking the ratio between the average speech energy and the average of noise energy. Silent frames were identified by setting a threshold on frame energy and discarded.

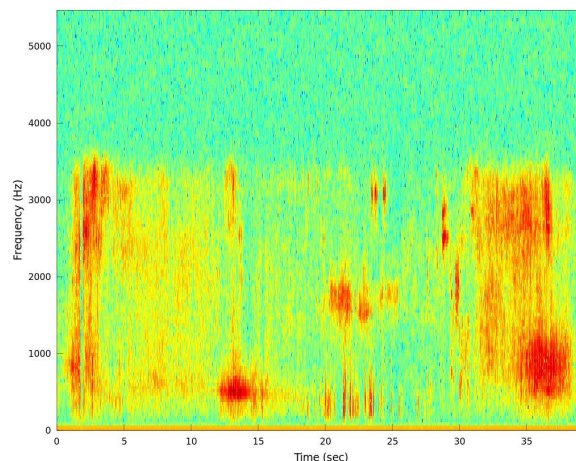


Figure 1: Spectrogram of a Typical Street Noise Sample

UBM, and consequently, the speaker models were modeled by GMMs with 256 Gaussian components. The UBM was trained by pooling all the training speech utterances. It should be noted that this method of training the UBM is a variation of the cohort approach of training background models described in [11] and it does not mean that the verification was closed set. We have followed the NIST-2000 evaluation scheme of target speakers and claimants. This cohort approach of UBM training was necessary, as we did not have access to any developmental data. Speech utterances were divided into 20 ms frames with 10 ms overlap. 20 MFCCs were computed from each of the frame using 27 mel-filters. 0th cepstral coefficients as well as the last six cepstral coefficients were not considered for training or scoring, but they were kept in the model so that the parameters can be transformed back to mel-spectral domain without any approximations. For this work, we have used only static feature coefficients (i.e., the MFCC coefficients) and not the dynamic coefficients (i.e., the delta and double-delta coefficients). The compensation of dynamic coefficients is being investigated currently and will be addressed in the near future.

We have implemented four systems. The first employs the feature space CMVN. The second system employs MCMVN proposed in Section 3. The third system implements the VTS algorithm described in [12]. The fourth system implements the Psy-Comp algorithm along with the MCMVN. The results, in terms of Equal Error Rate (EER) and minimum Decision Cost Function (min-DCF) are summarized in Table 1 and 2. The min-DCF was computed using the cost parameters from NIST-2000 evaluation plan (i.e. $C_{Miss} = 10$, $C_{FalseAlarm} = 1$, $P_{Target} = 0.01$). The Detection Error Trade-off (DET) plot for street noise at 5dB SNR and factory-1 noise at 5dB SNR are provided in Figure 2 and Figure 3, respectively.

From the results, it can be observed that the proposed method significantly outperforms the other robustness algorithms in terms of the EER as well as the min-DCF. The most surprising result for us was the performance of the VTS algorithm. It performed even worse than the CMVN and the MCMVN. We can think of two possible explanations for this. First, we have not considered the dynamic coefficients in this work. As reported in [12], VTS provides significant improvements for compensation of dynamic coefficients. Therefore, we expect VTS to

	Street Noise 5dB		Street Noise 10dB	
	EER	Min-DCF	EER	Min-DCF
CMVN	19.8	0.0924	16.51	0.0783
MCMVN	20.04	0.0893	17.00	0.0769
VTS	25.24	0.0995	24.91	0.1000
Proposed	17.58	0.0742	16.06	0.0675

Table 1: EER and Min-DCF for various methods in street noise condition

	Factory Noise 5dB		Factory Noise 10dB	
	EER	Min-DCF	EER	Min-DCF
CMVN	29.67	0.1000	21.39	0.1000
MCMVN	30.86	0.0999	21.72	0.0999
VTS	30.49	0.1000	24.50	0.0936
Proposed	26.6	0.0998	20.45	0.0823

Table 2: EER and Min-DCF for various methods in factory noise condition

perform better once we take dynamic coefficients into consideration. Second, in our observations, cepstral variance normalization plays a significant role in robust speaker recognition. The VTS compensates for the channel mismatch by adjusting the model mean vectors only, which is akin to cepstral mean subtraction. The variance normalization is not performed by the VTS and consequently, the positive effect of cepstral variance normalization is not observed. It can also be seen from the results that the performance of the proposed MCMVN is equal to the performance of the feature space CMVN. The benefit of the MCMVN is that it can be applied on the model parameters after compensation has been performed for additive noise through the Psy-Comp. The complimentary nature of Psy-Comp and MCMVN has resulted in the significant gains in performance reported here.

6. Conclusions and Future Work

We have proposed a compensation scheme for the compensation of the static model parameters for robust speaker recognition in additive noise. We have shown that the proposed scheme performs considerably better than the CMVN and the VTS algorithms. The performance of the proposed scheme is also remarkable since it is computationally more efficient than the VTS algorithm. Our current research goal is the compensation of the dynamic coefficients (i.e. the deltas and double deltas). We are developing the dynamic coefficients compensation scheme along the lines of VTS and PMC. Once we address the compensation of dynamic coefficients, we will investigate the methods of using the proposed scheme with i-vector technique. Given that the VTS has been used with i-vector technique in [6] to obtain good results in noisy conditions, it will be interesting to see how Psy-Comp fares compared to the VTS.

7. References

- [1] N. V. Prasad and S. Umesh, "Improved cepstral mean and variance normalization using Bayesian framework," in *2013 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2013, pp. 156–161.
- [2] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verifica-

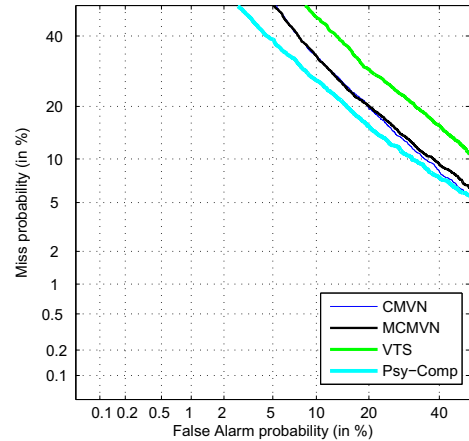


Figure 2: DET curves for various methods in street noise (5 dB) condition

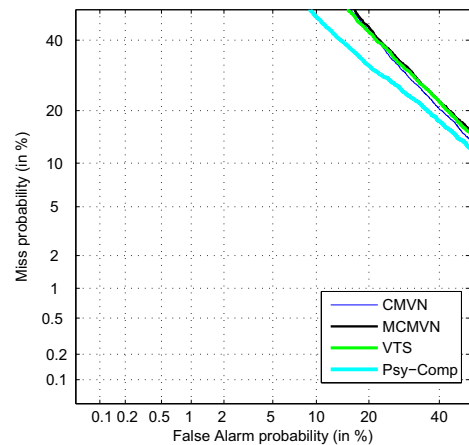


Figure 3: DET curves for various methods in factory noise (5 dB) condition

tion," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

- [3] N. B. Yoma and M. V. Fernandez, "Speaker verification in noise using a stochastic version of the weighted Viterbi algorithm," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 3, pp. 158–166, 2002.
- [4] L. P. Wong and M. Russell, "Text-dependent speaker verification under noisy conditions using parallel model combination," in *Proceedings of ICASSP'01*, vol. 1, 2001, pp. 457–460.
- [5] J. Ming, T. J. Hazen, J. R. Glass, and D. A. Reynolds, "Robust speaker recognition in noisy conditions," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 15, no. 5, pp. 1711–1723, 2007.
- [6] Y. Lei, L. Burget, and N. Scheffer, "A noise robust i-vector extractor using vector Taylor series for speaker recognition," in *Proceedings of ICASSP'13*, 2013, pp. 6788–6791.

- [7] Y. Lei, M. McLaren, L. Ferrer, and N. Scheffer, "Simplified VTS-based i-vector extraction in noise-robust speaker recognition," in *Proceedings of ICASSP'14*, 2014, pp. 3037–4041.
- [8] A. Panda and T. Srikanthan, "Psychoacoustic model compensation for robust speaker verification in environmental noise," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 3, pp. 945–953, 2012.
- [9] A. Panda, "Psychoacoustic model compensation with robust feature set for speaker verification in additive noise," in *Proceedings of ISCSLP'14*, 2014.
- [10] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19–41, 2000.
- [11] D. A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," *Speech Communication*, vol. 17, no. 1-2, pp. 91–108, 1995.
- [12] J. Li, L. Deng, D. Yu, Y. Gong, and A. Acero, "High-performance HMM adaptation with joint compensation of additive and convolutive distortions via vector Taylor series," in *Proceeding of the 2007 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2007.