A Comparison of Soft and Hard Spectral Subtraction for Speaker Verification*

Michael Padilla†and Thomas F. Quatieri‡

Stanford University†
MIT Lincoln Laboratory‡
mtpadilla@ieee.org†, quatieri@ll.mit.edu‡

Abstract

An important concern in speaker recognition is the performance degradation that occurs when speaker models trained with speech from one type of channel are subsequently used to score speech from another type of channel, known as channel mismatch. This paper investigates the relative performance of two different spectral subtraction methods for additive noise compensation in the context of speaker verification. The first method, termed "soft" spectral subtraction, is performed in the spectral domain on the $|DFT|^2$ values of the speech frames while the second method, termed "hard" spectral subtraction, is performed on the Mel-filter energy features. It is shown through both an analytical argument as well as a simulation that soft spectral subtraction results in a higher signal-to-noise ratio in the resulting Mel-filter energy features. In the context of Gaussian mixture model-based speaker verification with additive noise in testing utterances, this is shown to result in an equal error rate improvement over a system without spectral subtraction of approximately 7% in absolute terms, 21% in relative terms, over an additive white Gaussian noise range of 5-25 dB.

1. Introduction

This paper compares the relative performance of two variations of spectral subtraction for additive noise compensation within the context of Gaussian mixture model (GMM)-based speaker verification. One method of spectral subtraction (SS), termed "soft" SS, attempts to mitigate additive noise effects in the spectral domain [1]. The other method, termed "hard" SS, instead performs the noise subtraction in the domain of the Mel-filter energy features themselves [2]. This paper shows both analytically as well as empirically that soft SS produces a higher average signal-to-noise ratio (SNR) in the resulting Mel-filter energy features in typical operating scenarios with additive white Gaussian noise (AWGN), resulting in improved speaker verification performance [3].

The paper is organized as follows. First in Section 2, the general problem of speaker verification is defined as is the nature of the GMM speaker model employed. In addition, the standard Mel-filter energy speech features used in this study are briefly reviewed. In Section 3, a summary of soft and hard SS is presented as are analytical and empirical results showing the superiority of the soft variety. Section 4 then presents experimental results comparing the performance of two GMM-based speaker verification systems using soft and hard SS in an AWGN environment. The paper is concluded in Section 5.

*This work was sponsored by the United States Air Force under Air Force Contract F19628-00-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

2. Speaker Verification Essentials

This section briefly reviews the GMM approach to speaker verification and the Mel-filter energy speech features, both of which were employed in this work.

2.1. GMM-based Speaker Verification

In the statistical speaker model, each speaker is regarded as a random source producing the observed speech feature vectors $\mathbf{X}$ as a function of the vocal tract's state. If an M-state statistical speaker is assumed, the probability density function for the GMM speaker model [4] is given by

$$p(\mathbf{X}|\lambda) = \sum_{i=1}^{M} p_i b_i(\mathbf{X}) \tag{1}$$

where

$$b_i(\mathbf{X}) = \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}} \exp\{-\frac{1}{2}(\mathbf{X}-\mu_i)^T\Sigma_i^{-1}(\mathbf{X}-\mu_i)\}. \tag{2}$$

Here μ_i and $\mathbf{\Sigma}_i$ are the mean vector and covariance matrix for speaker state i, respectively, D is the dimensionality of the feature vectors, and p_i is the probability of being in state i. The set of quantities

$$\lambda = (p_i, \mu_i, \mathbf{\Sigma}_i), \quad \text{for i = 1,...,M} \tag{3}$$

constitutes a complete speaker model.

In the speaker verification task, specifically, there are two competing classifications: (1) "is claimant" and (2)

"is not claimant". A GMM model for the former, referred to as the target model, is derived from training utterances for the known speaker, while a GMM model for the latter, referred to as the background model , is constructed by using a diverse collection of speech from representative imposter speakers. For an unknown speech file $\mathbf{X}_T = [\mathbf{X}_1\ \mathbf{X}_2\ ...\ \mathbf{X}_T]$ (with speech vectors $\mathbf{X}_t$), T representing the number of frames, a claimant with model λ_c and background model λ_{bkgd} are used to obtain a log-likelihood ratio, given by

$$\Lambda(\mathbf{X}_T) = \log[p(\mathbf{X}_T|\lambda_c)] - \log[p(\mathbf{X}_T|\lambda_{bkgd})], \quad (4)$$

where the speaker and background log-likelihood scores of the utterance are calculated as

$$\log[p(\mathbf{X}_T|\lambda_{c/bkgd})] = \frac{1}{T}\sum_{t=1}^{T}\log[p(\mathbf{X}_t|\lambda_{c/bkgd})]. \quad (5)$$

The log-likelihood ratio is compared with a threshold to determine the relative number of misses and false acceptances over a threshold range. Typically, the equal error rate (ERR) point, i.e., the point at which the misses equal the false acceptances, is referenced as a good representative measure of performance.

2.2. Mel-filter Energy Speech Features

This study assumes Mel-filter energy speech features [5]. For each windowed speech segment $x[m,n]$, m denoting the frame number, the discrete short time Fourier transform (DSTFT) $X[m,k]$ is computed via the discrete Fourier transform (DFT), where k denotes the DFT frequency sample.

The values $|X[m,k]|^2$ are passed through a bank of triangular filters known as a *Mel-scale filterbank*. Assuming that there are N Mel-filters and that $M_l[k]$ represents the values of the l^{th} filter, the *Mel-filter "energy"* at the output of the l^{th} filter is given by

$$\mathcal{M}[m,l] = \sum_{k=L_l}^{U_l} |M_l[k]X[m,k]|^2, \quad (6)$$

where L_l and U_l are the lower and upper frequencies of the l^{th} filter, respectively. Note that typically the M_l values are normalized so that each has unit energy ($\sum |M_l[k]|^2 = 1\ , \forall\ l$). The resulting N-element vector constitutes the speech features for frame m.

3. Noise Compensation by Spectral Subtraction

This section reviews hard and soft forms of SS for noise compensation in speaker recognition, showing both analytically and empirically that soft SS provides superior performance.

3.1. Mel-Filter Energy Domain (Hard) Spectral Subtraction

Consider a speech signal $x[n]$ that has been corrupted by stationary additive noise $a[n]$, to produce a noisy speech signal $y[n]$:

$$y[n] = x[n] + a[n]. \quad (7)$$

With the assumption that $x[n]$ and $a[n]$ are independent zero mean wide-sense stationary processes, it is easily shown that the expected value of $|Y[m,k]|^2$ is

$$E[|Y[m,k]|^2] = E[|X[m,k]|^2] + E[|A[m,k]|^2]. \quad (8)$$

The resulting Mel-filter energy features $\mathcal{M}[m,l]$ will on average be composed of a signal component and an additive noise component. By analyzing silence frames, an estimate of the average Mel-filter energies for the noise, $\hat{\mathcal{N}}[m,l]$, may be found. The true value of the signal component of the corrupted Mel-filter energy features, $\mathcal{M}_{true}[m,l]$, is then estimated as

$$\hat{\mathcal{M}}_{true}[m,l] = \mathcal{M}[m,l] - \hat{\mathcal{N}}[m,l]. \quad (9)$$

Due to issues involving artificially created musical tones in re-synthesized speech when SS is used in the $|DFT|^2$ domain, a nonlinear flooring operation is typically employed [2]. Although the motivation for this nonlinear modification is not directly relevant to the Mel-filter energies in the context of speaker identification, a similar formulation has been adopted in this domain as well. The relationship for this perceptually motivated form of SS, referred to as *generalized SS* [2], is given by

$$\mathcal{D}[m,l] = \mathcal{M}[m,l] - \alpha\hat{\mathcal{N}}[m,l] \quad (10)$$

and

$$\hat{\mathcal{M}}_{true}[m,l] = \begin{cases} \mathcal{D}[m,l], & \text{if } \mathcal{D}[m,l] > \beta\hat{\mathcal{N}}[m,l] \\ \beta\hat{\mathcal{N}}[m,l], & \text{otherwise,} \end{cases} \quad (11)$$

where $\alpha \geq 1$ and $0 < \beta \leq 1$. The parameter α is used to overestimate the Mel-filter energy of the noise, and β determines the level of the spectral flooring.

3.2. $|DFT|^2$ Domain (Soft) Spectral Subtraction

In hard SS, because the noise component of the feature is being estimated from its average, necessarily different from its instantaneous value, the noise level will frequently be overestimated. In this case, the nonlinearity will clip the resulting speech feature estimate to 0. This will often times result in a very large deviation from the true underlying clean speech feature.

It is this consideration that makes the application of SS-based feature enhancement in the $|DFT|^2$ domain a

potentially more attractive option. Because the nonlinear flooring operation is now applied to the $|DFT|^2$ values prior to the Mel-filterbank, the error component of each spectrally subtracted value may be decreased through averaging via the Mel-filters $M_l[k]$. Consider the following two expressions that show the resulting Mel-filter energy $\mathcal{M}[m,l]$ when SS, the nonlinear flooring operation of which is symbolized by the operator $\mathcal{S}$, is applied in the $|DFT|^2$ domain and in the Mel-filter energy domain, respectively:

$$\mathcal{M}_{dft}[m,l] = \sum_k M_l^2[k]\mathcal{S}(|Y[m,k]|^2 - \hat{N}_{dft}[m,k]) \tag{12}$$

and

$$\mathcal{M}_{mel}[m,l] = \mathcal{S}(\{\sum_k |M_l[m,k]Y[m,k]|^2\} - \hat{N}_{mel}[m,k]). \tag{13}$$

Here $\hat{N}_{dft}[m,k]$ is the noise spectral estimate in the $|DFT|^2$ domain, and $\hat{N}_{mel}[m,k]$ is the noise spectral estimate in the Mel-filter energy domain. It is seen that while $\mathcal{M}_{mel}[m,l]$ is exposed to a single nonlinear operation $\mathcal{S}$, $\mathcal{M}_{dft}[m,l]$ is computed by a weighted summation of the outputs of a large number of such nonlinear operations. This has the effect of reducing the noise introduced into the enhanced speech feature through averaging of the nonlinear operator outputs.

More specifically, in the expression for $\mathcal{M}_{mel}[m,l]$ above, the Mel-filter domain spectral noise estimate, $\hat{N}_{mel}[m,k]$, may be re-written as

$$\hat{N}_{mel}[m,k] = \sum_k M_l^2[k]\hat{N}_{dft}[m,k], \tag{14}$$

producing

$$\mathcal{M}_{mel}[m,l] = \mathcal{S}(\sum_k M_l^2[k]\{|Y[m,k]|^2 - \hat{N}_{dft}[m,k]\}). \tag{15}$$

Let $\hat{X}_{ms}[m,k] = |Y[m,k]|^2 - \hat{N}_{dft}[m,k]$ be the estimate for $|X[m,k]|^2$ (ms denoting "magnitude squared"). The expressions for $\mathcal{M}_{dft}[m,l]$ and $\mathcal{M}_{mel}[m,l]$ then simplify to

$$\mathcal{M}_{dft}[m,l] = \sum_k M_l^2[k]\mathcal{S}(\hat{X}_{ms}[m,k]) \tag{16}$$

and

$$\mathcal{M}_{mel}[m,l] = \mathcal{S}(\sum_k M_l^2[k]\hat{X}_{ms}[m,k]). \tag{17}$$

A simplified analysis may be performed on these expressions to develop intuition for what they mean in terms of the SNR of the resulting features. Here SNR refers to the variance of the true component of the feature over the mean squared error of the difference between this true component and the actual corrupted component. This analysis assumes that the spectral components of the DFT are independent.

Let $\epsilon[m,k]$ denote the contribution to the difference between $\hat{X}_{ms}[m,k]$ and $|X[m,k]|^2$ due to the nonlinear flooring operation in $\mathcal{S}$ and model it as a random variable with an unknown distribution with mean μ_ϵ and variance $\sigma^2_{\epsilon_{nl}}$. As the non-zero values for each of the 24 Mel-filters are normalized to have unit energy (i.e. $\sum_k M_l[k]^2 = 1$) [5], we approximate the nonlinear noises in $\mathcal{M}_{dft}[m,l]$ and $\mathcal{M}_{mel}[m,k]$, $\epsilon[m,k]$ and $\epsilon[m,l]$, as having roughly the same first and second moments. Then

$$\begin{aligned}\mathcal{M}_{dft}[m,l] &\approx \sum_k M_l^2[k](\hat{X}_{ms}[m,k] + \epsilon[m,k]) \quad (18)\\ &= \sum_k M_l^2[k]\hat{X}_{ms}[m,k] \\ &\quad + \sum_k M_l^2[k]\epsilon[m,k] \quad (19)\end{aligned}$$

and

$$\mathcal{M}_{mel}[m,l] \approx \sum_k M_l^2[k]\hat{X}_{ms}[m,k] + \epsilon[m,l]. \tag{20}$$

Let

$$\zeta_{dft}[m,l] = \sum_k M_l^2[k]\epsilon[m,k] \tag{21}$$

and

$$\zeta_{mel}[m,l] = \epsilon[m,l] \tag{22}$$

represent the two terms that are different between these two feature types, each representing the respective error introduced by $\mathcal{S}$ for speech frame m and feature l. It is then seen that given the normalization of the $M_l[k]$ terms, the means of both terms are approximately μ_ϵ while the variances have the relationship

$$\sigma^2_{\zeta_{dft}} < \sigma^2_{\zeta_{mel}}. \tag{23}$$

The benefit of soft SS is that the error term in the resulting Mel-feature resulting from the nonlinearity, while having roughly the same mean as in the hard case, has a variance that is smaller.

To empirically demonstrate the SNR improvement that soft SS provides in comparison to hard SS, a simulation was performed in which speech DFT values were corrupted with AWGN and the resulting SNR in the output $\mathcal{M}_{dft}[m,l]$ and $\mathcal{M}_{mel}[m,l]$ feature classes compared. The results are in figure 1. It is seen that the resulting speech feature under soft SS outperforms that of hard SS by approximately 2 dB across the entire SNR range of the AWGN degrading $X[m,k]$.

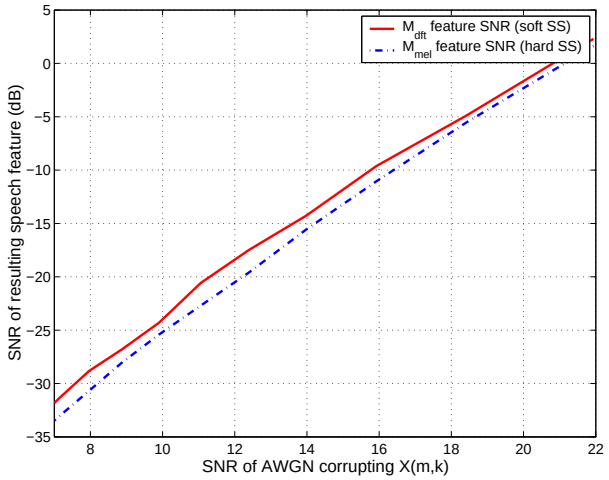

Figure 1: *Comparison of SNR in speech features resulting from hard and soft SS in an AWGN environment.*

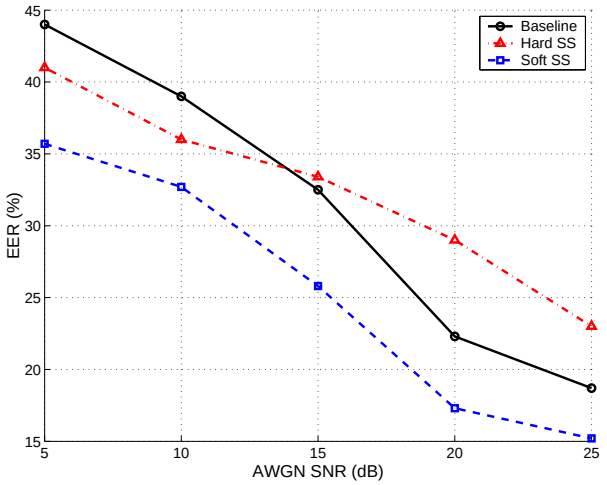

Figure 2: *EER vs. SNR (AWGN) for both the baseline case and hard and soft SS cases.*

4. Results with Speaker Verification

Speaker verification experiments were performed to evaluate both Mel-filter energy and $|DFT|^2$ domain SS. Both focused on the train clean/test dirty case where the corruption was AWGN. In all of the simulations the speech features used in scoring were Mel-filter energies and the sampling rate was 8 khz, resulting in 24 features. 1024 mixtures were used in each GMM speaker model. The speech files were taken from the TSID corpora, a collection of 35 male/female speakers reading digits, sentences, and directions. In this study, controlled "dirty" speech was simulated by adding AWGN to the speech of the "clean" channel. A baseline case, without any SS, was included for comparison. Approximately 20 and 100 minutes of speech was used for training speaker and background models, respectively. Speaker and background models were each trained with approximately 5 and 40 minutes each, respectively. Test utterances were approximately 2 minutes each.

The performance when hard SS is employed, shown in figure 2, is seen to depend on the level of noise corrupting the speech. For SNR values below approximately 14 dB, application of hard SS in this domain was found to improve performance over the baseline by about 4%, while for higher SNRs performance was degraded by about 5%.

In contrast, it is seen in figure 2 that the performance when soft SS is applied in the $|DFT|^2$ domain is improved substantially *at all SNR levels* over the baseline. At high SNRs above approximately 15 dB the EER is improved by about 4% and this improvement increases to approximately 9% at low SNR levels. At all SNRs, results with soft SS outperform those using hard SS.

5. Conclusion

This study has investigated the techniques of hard and soft SS for GMM-based speaker verification in a train clean/test dirty scenario with AWGN corruption. It has been shown that over a range of 5-25 dB SNR for the additive noise that the resulting soft SS speech features have an improved SNR of roughly 2 dB compared to those derived from hard SS. By use of soft SS, the resulting ERR improvement over baseline was shown to be between 4% and 9%, depending on the SNR of the additive noise.

Acknowledgement: The authors thank Dr. Doug Reynolds (MIT Lincoln Laboratory) and Dr. Jack McLaughlin (University of WA) for help with the GMM code and the TSID corpus.

6. References

[1] J. Ortega-Garcia and J. Gonzalez-Rodriguez, *Overview of Speech Enhancement Techniques for Automatic Speaker Recognition*, Proc. ICSLP, Oct. 1996

[2] A. Drygajlo and M. El-Maliki, *Use of Generalized Spectral Subtraction and Missing Feature Compensation for Robust Speaker Verification*, RLA2C, April 1998.

[3] M. Padilla, *Applications of Missing Feature Theory to Speaker Recognition*, S.M. Thesis, MIT, 2000

[4] D.A. Reynolds, *Automatic Speaker Recognition Using Gaussian Mixture Speaker Models*, The Lincoln Laboratory Journal, Vol.8-2, Pages 173-192, 1995.

[5] T.F. Quatieri, *Principles of Discrete-Time Speech Processing*, Prentice Hall, 2000.