Missing Feature Theory with Soft Spectral Subtraction for Speaker Verification*

Michael T. Padilla†, Thomas F. Quatieri‡, and Douglas A. Reynolds§

Stanford University†
MIT Lincoln Laboratory‡§
mtp@stanford.edu†, quatieri@ll.mit.edu‡, dar@sst.ll.mit.edu§

Abstract

This paper considers the problem of training/testing mismatch in the context of speaker verification and, in particular, explores the application of missing feature theory in the case of additive white Gaussian noise corruption in testing. Missing feature theory allows for corrupted features to be removed from scoring, the initial step of which is the detection of these features. One method of detection, employing spectral subtraction, is studied in a controlled manner and it is shown that with missing feature compensation the resulting verification performance is improved as long as a minimum number of features remain. Finally, a blending of "soft" spectral subtraction for noise mitigation and missing feature compensation is presented. The resulting performance improves on the constituent techniques alone, reducing the equal error rate by about 15% over an SNR range of 5 - 25 dB.

Index Terms: speaker verification, GMM, spectral subtraction, missing features.

1. Introduction

An important concern in speaker verification is the degradation that occurs when speaker models trained with speech from one type of channel are subsequently used to score speech from another, known as channel mismatch. This paper presents a Gaussian mixture model (GMM)-based speaker verification system [1] that uses a merging of a "soft" variation of spectral subtraction (SS) [2] with missing feature theory that helps mitigate this problem. Training speech is clean, while testing speech is corrupted with additive white Gaussian noise (AWGN). Missing feature theory recognizes that at times features may be too corrupted to be usable and attempts to detect and gracefully remove such features from the scoring process via a method known as *missing feature compensation* (MFC) [3] [4]. In this case SS may be employed not only as a noise compensation technique, but also as a *missing feature detector* (MFD). This paper explores this detection method and presents a controlled study of performance limits for MFD, providing evidence that with AWGN corruption, missing features can be detected with reasonable accuracy relative to a perfect detector. The resulting process of MFD and MFC is shown to significantly enhance performance in certain circumstances so long as there is a sufficient number of speech features retained for scoring. Finally, the paper considers the linear combination of SS, as a noise mitigation technique, with MFC in an AWGN environment and demonstrates that this combination improves performance with respect to equal error rate (EER, where % misses = % false alarms) by $\approx 15\%$ across a wide range of noise levels relative to the case where the constituent methods are applied alone and in comparison to the baseline case of no attempted noise mitigation.

*This work was sponsored by the Department of Defense under Air Force Contract FA8721-05-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

2. Comparison of Hard and Soft SS

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

$$y(n) = x(n) + a(n). \tag{1}$$

With the assumption that $x(n)$ and $a(n)$ are independent zero mean wide-sense stationary processes,

$$E[|Y(m,k)|^2] = E[|X(m,k)|^2] + E[|A(m,k)|^2], \tag{2}$$

where $X(m,k)$, $A(m,k)$, and $Y(m,k)$ represent discrete short time Fourier transforms [5] for frame m, k denoting the frequency variable. The resulting Mel-filter energy features [5] $\mathcal{M}(m,l)$, l denoting the feature number, are composed of signal and additive noise components. Using non-speech 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 in the basic form of SS as

$$\hat{\mathcal{M}}_{true}(m,l) = \mathcal{M}(m,l) - \hat{\mathcal{N}}(m,l). \tag{3}$$

Generalized SS, described in [3], is given by

$$\mathcal{D}(m,l) = \mathcal{M}(m,l) - \alpha\hat{\mathcal{N}}(m,l) \tag{4}$$

and

$$\hat{\mathcal{M}}_{true}(m,l) = \begin{cases} \mathcal{D}(m,l), & \text{if } \mathcal{D}(m,l) > \beta\hat{\mathcal{N}}(m,l) \\ \beta\hat{\mathcal{N}}(m,l), & \text{otherwise,} \end{cases} \quad (5)$$

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. This variant of SS is referred to here as *hard SS*.

By applying the SS to the $|DFT|^2$ values prior to the Mel-filterbank, the error component of each feature may be decreased through averaging [2]. Consider the following two expressions that give the resulting Mel-filter energy $\mathcal{M}(m,l)$ when SS, the nonlinear flooring operation (equation (5)) of which is symbolized by the operator $\mathcal{S}$, is applied in the $|DFT|^2$ domain and Mel-filter energy domains, respectively:

$$\mathcal{M}_{dft}(m,l) = \sum_k M_l^2(k)\mathcal{S}(|Y(m,k)|^2 - \hat{N}_{dft}(m,k)) \quad (6)$$

and

$$\mathcal{M}_{mel}(m,l) = \mathcal{S}(\{\sum_k |M_l^2(k)Y(m,k)|^2\} - \hat{N}_{mel}(m,k)), \quad (7)$$

where $M_l^2(k)$ denotes the values of the l^{th} Mel-filter. $\hat{N}_{dft}(m,k)$ 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 operations. In [2] it is shown that the resulting speech features under this operation, referred to as *soft SS*, outperform those of hard SS by approximately 2 dB across a wide SNR range of AWGN degrading $X(m,k)$.

3. Missing Feature Compensation

This section first introduces the general theory of MFC in a speaker recognition framework. The issue of detecting missing features is then discussed and results of baseline speaker verification experiments with missing feature compensation are described.

3.1. General Theory

Missing feature theory recognizes that including a highly corrupted feature in the scoring mechanism may worsen performance compared to if it were omitted. One manner discussed in [3] and [4] to deal with a missing feature is to adapt the GMMs to remove it from inclusion in scoring. This method, MFC, uses generalized SS to classify features as "missing" or "present", rather than as a speech enhancement pre-processor. Since the covariance matrix in GMM, denoted for each mixture as Σ_i, is often assumed diagonal, each of the multi-variate Gaussian PDFs Φ may be re-written as a product of single-variate Gaussian PDFs as

$$p(\mathbf{X}|\lambda) = \sum_{i=1}^{M} p_i \prod_{k=1}^{D} \Phi_i(x_k, \mu_{ki}, \sigma_{ki}^2),$$

where μ_{ki} and σ_{ki}^2 are the mean and variance of feature element x_k in state i and p_i is the prior probability of the speaker model being in state i. Assuming the ability to distinguish "present" and "missing" features, to be discussed in section 3.3, we may divide our overall speech feature vector $\mathbf{X} = \{x_1, x_2, ..., x_D\}$ into two sub-vectors, $\mathbf{X}_{present}$ and $\mathbf{X}_{missing}$ to give

$$p(\mathbf{X}|\lambda) = \sum_{i=1}^{M} p_i \prod_{j=1}^{D_{present}} \Phi_i(x_j, \mu_{ji}, \sigma_{ji}^2) \prod_{k=1}^{D_{missing}} \Phi_i(x_k, \mu_{ki}, \sigma_{ki}^2)$$

MFC then proceeds by removing the second product term representing the sub-vector $\mathbf{X}_{missing}$ and leaving only the data associated with $\mathbf{X}_{present}$:

$$p_{mfc}(\mathbf{X}|\lambda) = \sum_{i=1}^{M} p_i \prod_{j=1}^{D_{present}} \Phi_i(x_j, \mu_{ji}, \sigma_{ji}^2),$$

which is then used in place of the full GMM $p(\mathbf{X}|\lambda)$.

3.2. Potential Performance Benefits

To study the potential degradation caused by a single missing feature, a simulation was run in which the 10^{th} linear Mel-energy feature for every frame for an otherwise clean/clean case was nulled[1]. The resulting EER (32%) is given in table 1. For comparison, results are also given for cases of uncorrupted clean/clean (3.26%), clean/clean with AWGN (23%), as well as MFC applied to the 10^{th} feature in every frame (3.39%). Throughout this paper the sampling rate was 8 khz, resulting in 24 features/frame. GMMs had 1024 mixtures. All simulations used speech data from the TSID corpus[2]. Controlled dirty speech was simulated by adding AWGN to the speech of the clean channel. *Speaker and background* models were each trained with approximately 5 and 40 minutes each, respectively. Test utterances were approximately 2 minutes each in duration. The results demonstrate that missing features can significantly reduce performance if included and that performance may be significantly restored through MFC.

Scenario	EER (%)
Clean/Clean	3.26
Clean/(Clean w/ MFC applied to every 10th feature)	3.39
Clean/(Clean + AWGN) (20dB)	23
Clean/(Clean + Every 10th Feature Zeroed)	32

Table 1: Potential benefit of applying MFC.

[1] The 10^{th} feature (out of 24) was selected as it has been experimentally shown to represent information in a relatively important spectral band [6].

[2] Tactical Speaker ID Speech Corpus, Linguistic Data Consortium, http://www.ldc.upenn.edu.

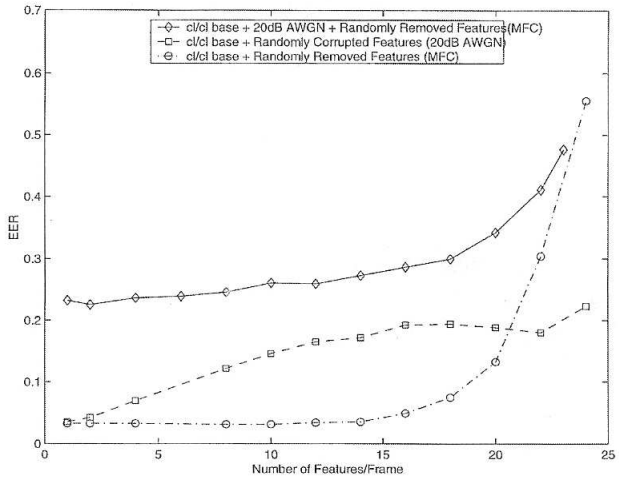

Figure 1: Trade-off between corrupted features and removing speech information in MFC.

To further investigate the speech distortion/noise compensation trade-off, simulations involving the corruption of randomly chosen features every frame for the clean test data were run. The number of corrupted Mel-filter energy features was fixed, although the particular features were determined randomly frame-by-frame. The corruption was AWGN at 20 dB. One simulation looked at verification performance using this randomly corrupted data. For comparison, a second study had a certain predetermined number of features randomly chosen and removed by MFC each frame. Finally, a third study added 20 dB AWGN to all features and a fixed number of features each frame were randomly chosen to be removed by MFC. The resulting EER points are in figure 1. The application of MFC maintains EER performance until the removal of $\simeq 15$ features, outperforming the case of no MFC up to the removal of about 20. In these cases MFC improves performance to a certain point, beyond which the noise removal cannot compensate for the loss of a minimum amount of necessary speech information.

3.3. Speaker Verification with Missing Feature Compensation

In order to remove missing features, the speaker verification system must first be able to distinguish which features are missing and which are present. As suggested by Drygajlo and El-Maliki [7], generalized SS may be used as a missing feature detector:

$$\mathcal{M}(m,l) - \alpha\hat{\mathcal{N}}(m,l) \underset{\mathcal{M}(m,l)\text{ "missing"}}{\overset{\mathcal{M}(m,l)\text{ "present"}}{\gtrless}} \beta\hat{\mathcal{N}}(m,l),$$

with typically $\beta = 0$. It was experimentally found that $\alpha = 3$ produced the best EER results and this was used throughout this work. This suggests that *overestimating* the amount of noise energy is preferred and relates to the noise vs. speech information removal trade-off.

Assuming the criterion for detecting missing features above is appropriate, a "*perfect*" missing feature detector is available for cases where AWGN is added since both the clean linear Mel-energy feature as well as its corrupted version are available; $\hat{\mathcal{N}}(m,l)$ is known perfectly. In contrast is the "*imperfect*" detector, where the expressions above still hold but the noise estimate $\hat{\mathcal{N}}(m,l)$ is derived from averaging features during non-speech frames.

Of interest is the similarity between decisions made by the imperfect detector, using only an estimate of $\overline{\mathcal{N}}(m,l)$, and those made by a perfect detector, which has access to the actual noise component $\mathcal{N}(m,l)$. The correlation between the perfect and imperfect missing feature detectors for each feature was studied in an AWGN environment at SNR levels from 5 dB to 20 dB. For most features the correlation, averaged over the range of AWGN studied, between the perfect and imperfect (with $\alpha = 3.0$) missing feature detectors was high, greater than 0.7 and closer for 0.9 for the majority. Hence at various SNR levels the imperfect detector's decisions will closely resemble those of the perfect missing feature detector for most features under AWGN. Also of interest is the similarity between perfect and imperfect detectors with respect to other aspects of MFD, such as the average number of times a particular feature is declared missing per frame and the total number of missing features declared missing per frame. These were also measured at various SNR values in an AWGN environment. For the former study, at 20 dB, it was again found that for $\alpha = 3.0$ the match was close with the typical deviation being 0.02 features/frame, the maximum deviation occuring for the 10^{th} feature at roughly 0.05 features/frame. The results for the latter study, with 20 dB AWGN, may be seen in figure 2, where very similar behaviour is observed. The speaker recognition performance of MFC with both the perfect and imperfect missing feature detectors relative to the baseline case was studied. As before, all tests were train clean/test dirty (AWGN) and the performance metric was EER. The results are in table 2. The cases with

SNR (dB)	Baseline	Perf. MFD/MFC	Non-Perf. MFD/MFC
5	44.3	44.3	45.2
10	39.5	39.7	40.7
15	32.5	34.8	34.2
20	22.7	19.9	19.9
25	18.7	14.9	15

Table 2: EER values for various scenarios of perf/imperf. MFD + MFC.

MFC are very similar, with the maximum deviation being less than 1% at lower SNR values. While the application of MFC at lower SNR values degrades performance, at higher SNRs of roughly 17 dB or more the MFC begins to outperform the baseline case. At noise levels starting

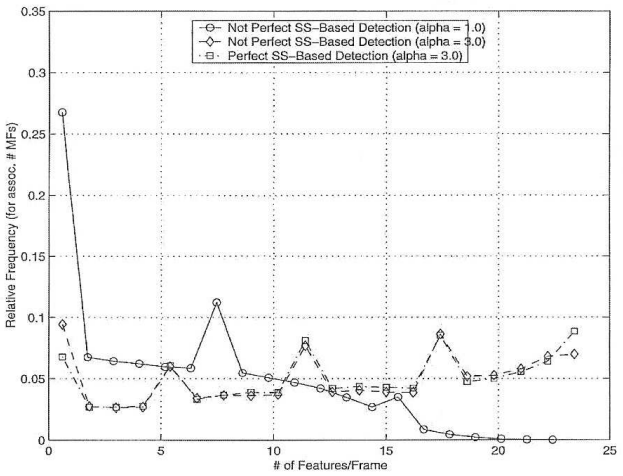

Figure 2: Frequency per frame that a given total number of missing features are detected by the perfect ($\alpha = 3.0$) and non-perfect missing feature detectors ($\alpha = 1.0$ and 3.0).

at 20 dB the improvement over the baseline rises to approximately $3 \sim 4\%$.

The lack of uniform improvement at different SNRs using MFC is likely due to the noise vs. speech information trade-off, seen in figure 1. At a critical point of noise compensation via MFC, the loss of speech information accompanying the removal of missing features causes a catastrophic loss of necessary speaker information.

4. Cascade Noise Handling Systems

SS (hard and soft varieties) and MFC may be applied jointly for further performance enhancement. To study this, systems incorporating SS with MFC were used in clean/dirty (AWGN) tasks with noise levels varying from 5 to 20 dB. Here logarithmic Mel-filter energies were employed, the rest of the experimental set-up as before. MFD was performed by imperfect detectors.

The results are in table 3. The combination of soft SS and MFC outperforms the pure soft SS system, the combination of hard SS and MFC, as well as the baseline. This trend holds at all SNR levels and achieves a reduction in EER by an average of as much as $\approx 15\%$ at the various SNR levels tested. The next best system was the pure soft SS system, which was outperformed by the combination system at all SNR levels by $\approx 10\%$. *Thus, for the particular channel and training/testing conditions given, the system combining soft SS and MFC has been shown to perform better than each system individually*. The resulting performance gain is equivalent to a 13 dB gain as measured by equivalent noise power.

Case	5 dB	10 dB	15 dB	20 dB	25 dB
Soft SS/MFC	32.3	22.4	15.9	8.2	6.7
Soft SS	35.2	31.4	25.5	18.1	15.3
Hard SS/MFC	36.2	33.0	25.5	18.6	17.4
Hard SS	40.9	35.8	33.4	28.7	23.5
Baseline	44.1	39.3	32.5	22.8	18.6

Table 3: EER values (%) for various combinations in clean/clean + AWGN scenario.

5. Conclusion

This paper has studied missing feature theory and MFC, combined with SS [2]. Unlike other methods that enhance noisy features, here highly corrupted features are estimated and removed from inclusion in the GMM based scoring mechanism.

The problem of MFD was also discussed. Through correlation studies it was found that a working imperfect missing feature detector, using an estimate of the background noise spectrum, produced decisions very close to those of the perfect missing feature detector with the true noise spectrum under the noise scenarios considered.

The MFC system's results, as with the soft SS system, depend on the SNR. At low SNR values the system performed below the baseline by about 1% and at high SNRs it improved the EER by close to 4%. Finally, the joint combination of soft SS and MFC was considered, again in a clean/dirty task with AWGN. This combination improved performance over the baseline and the performance seen by each technique applied individually. For an SNR range of 5 - 25 dB it was found to outperform the baseline case with respect to EER by $\approx 15\%$. In future work we will study the application of these methods to actual corrupted speech, moving beyond synthetic noise.

6. References

[1] D.A. Reynolds and R.C. Ross, *Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models*, IEEE Transactions on Speech and Audio Processing, Vol. 3, No. 1, Pages 72-83, Jan. 1995.

[2] M. Padilla and T.F. Quatieri, *A Comparison of Soft and Hard Spectral Subtraction for Speaker Verification*, ICSLP, 2004.

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

[4] R.P. Lippman and B.A. Carlson, *Using Missing Feature Theory to Actively Select Features for Robust Speech Recognition with Interruptions, Filtering, and Noise*, Proc. Fifth European Conf. on Speech Communication and Technology, Vol. 1, Pages KN37-40, Rhodes, Greece, Sept. 1997.

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

[6] S. VanVuuren and H. Hermansky, *On the Importance of Components of the Modulation Spectrum for Speaker Verification*, ICSLP, 1998.

[7] M. El-Maliki and A. Drygajlo, *Missing Features Detection and Handling for Robust Speaker Verification*, Proc. Sixth European Conf. on Speech Communication and Technology, Vol. 2, Pages 975-978, Budapest, Hungary, Sept. 1999.