ENHANCING ACOUSTIC MODELS FOR ROBUST SPEAKER VERIFICATION

Juan A. Nolazco-Flores, L. Paola García-Perera

Computer Science Department
Tecnológico de Monterrey, Campus Monterrey,
Monterrey, NL, MX
{jnolazco,paola.garcia}@itesm.mx

ABSTRACT

Acoustic model enhancement (AME) refers to adapting the acoustic models to compensate for the distortion induced by a speech enhancement technique. This work extends the AME technique for speaker verification recently presented by incorporating the corresponding adaptation of the model variances, and by exploring the trade off between noise over-estimation and flooring distortion in the verification error.

By using spectral subtraction (SS) as the speech enhancement technique, the extended AME highly outperformed SS alone particularly at moderately low SNRs (0dB-15dB), where the adaptation of the variance was found to considerably improve the equal error rate (EER).

Index Terms— Speaker verification, robustness, minimum verification error, additive noise, spectral subtraction, acoustic model enhancement.

1. INTRODUCTION

Speaker Verification (SV) is the task of accepting legitimate registered users and rejecting impostors from a voice segment and a claimed identity. Depending on the underlying assumptions in an application, SV systems can be designed to adopt a desired functionality. For example, text-independent SV can be used for automatic undisclosed monitoring of conversations. The inputs of this SV system are a speech sequence and a claimed identity, producing an output that either accepts or rejects the claim along with a confidence score. If the input speech sequence length is assumed infinite (or very long), a decision from SV could be drawn when a desired confidence level has been reached or when a decision is requested (decide-now button). In this work, we study such conventional text-independent SV system.

There are two sessions in SV: *enrollment* and *verification*. In enrollment, the user being registered provides several segments of speech, also called *positive tokens*. The verification session consists of a series of verification *trials* which can be identified either as *target* or *impostor* depending on the actual legitimacy of the claim.

Although there has been significant progress in SV, in part driven by the evaluation organized by NIST [1], SV remains an interesting topic of research. The acoustic mismatch between the enrollment and verification sessions rapidly degrades the performance. Robustness methods are indeed needed to enable SV systems be widely adopted in real applications. Among the main sources of acoustic mismatch the most widely used are convolutive noise (channel distortion) widely studied in [1], coding distortion [2], and additive ambient noise[3]. This work addresses this last case, not only found in outdoor environment but also in indoor places where air-condition and computer fans are operated.

Mitigating the acoustic mismatch between training and testing is the fundamental problem of robustness. In the presence of additive noise, one can either clean the noisy speech (speech enhancement) [3] [4], or modify the acoustic models originally trained with clean data to match the noise condition (model adaptation) [5]. Other methods make the features match the noise conditions (feature-based adaptation) such as SPLICE (State Based Piecewise Linear Compensation for Enviroments) [6], MEMLIN (Phoneme-Dependent Multi-Enviroment Enhanced Models Based LInear Normalization) [7].

Acoustic model enhancement (AME) is the adaptation of the acoustic models to compensate for the distortion induced by a speech enhancement technique. In this work the AME technique for speaker verification recently presented in [8] is extended by exploring the adaptation of the model variances. Moreover, several experiments were performed to anlyse the trade-off between noise under-subtraction and spectral flooring.

First the Section 2 introduces the principles of SV and explains the methodology for AME, then Section 3 explores a set of experiments that test the efficacy of AME. Finally, conclusions are presented in Section 4.

2. METHODOLOGY

2.1. SV framework

The decision of acceptance/rejection for a given trial is drawn from a hypothesis test, where the null hypothesis H_0 is to accept the speaker as legitimate and the alternative hypothesis H_1 is to reject it. This hard decision is based on the log-likelihood ratio test:

$$\theta = \ln \frac{P(H_0)}{P(H_1)}; \qquad \theta \begin{array}{c} accept \\ \geq \\ < \\ reject \end{array} \tau. \tag{1}$$

The acoustic modeling for the i-th speaker consists of two parts: a *target model* λ_{i-tgt} that captures the intrinsic characteristics of the speaker, and a corresponding *anti-model* λ_{i-anti} that provides a contrast of this speaker against impostors. The anti-model can be defined under two schemes: as a cohort (a set of GMMs: $\mathcal{C}$) or as a background model (a single-GMM), depending on the chosen scope of contrast.

The author acknowledges the support received from Tecnológico de Monterrey through grant number CAT009 to carry out the research reported in this paper.

During the speaker enrollment session, a set of positive tokens are collected from the speaker being registered. Another set of *negative tokens* can be collected beforehand from a pool of speakers, which ideally represents the population of impostor speakers.

2.1.1. Maximum likelihood

Acoustic modeling under the maximum likelihood estimation (MLE) criterion fits the observed data O to a probability density function (PDF), a GMM in this case, by using the expectation maximization (EM) algorithm that iteratively finds a model λ such that (n denotes the iteration number):

$$P(O|\lambda^{(n+1)}) \geq P(O|\lambda^{(n)}). \tag{2}$$

The positive tokens are used to fit the target model and the negative ones to fit the anti-model. Traditionally, the anti-model is shared by all speakers as a universal background model (UBM).

Since the availability of positive tokens is limited, the target models is built by first using negative tokens from a large pool of speakers to create a speaker-independent flat-start model, then performing maximum a posteriori (MAP) speaker adaptation [9]:

$$\theta_{MAP} = \underset{\lambda,\vartheta}{\operatorname{argmax}}\, P(O|\lambda;\vartheta)P(\lambda;\vartheta) \tag{3}$$

where $\theta = (\lambda, \vartheta)$ and ϑ is the meta-parameter of the distribution of λ. Equation 3 is also solved using the EM algorithm.

2.2. Acoustic model enhancement

When the testing and trainning occur in different acoustic conditions, the SV system accuracy degrades. Mitigating this mismatch is the main problem of robustness. In the presence of additive noise, one can either utilize speech enhancement [3] or model adaptation [5]. Since the noise or distortion can never be totally removed by enhancement algorithms, the strategy of AME is to obtain the best of both techniques.

The goal of AME is to mimic in the clean acoustic models (for H_0 and H_1) the distortion that results from the speech enhancement, therefore inducing a matched condition with the enhanced test data.

For models trained under the MLE criterion, AME attempts to find a transformation of the clean models (PDFs) to a new one where the enhanced speech fits, therefore preserving the maximum likelihood.

In this work spectral subtraction (SS) [3], as defined in Equation 4 and depicted in Figure 4, is used as the speech enhancement technique:

$$\mathbf{Y}_D = max(\mathbf{Y} - \alpha\hat{\mathbf{N}}, \beta\mathbf{Y}) \tag{4}$$

where $\mathbf{Y}$ is the observed spectrum of noisy signal, ($\hat{\mathbf{N}}$) is an estimate of the noise spectrum and $\mathbf{Y}_D$ is the corresponding spectrum of the enhanced (spectral subtracted) signal.

In practice, the phase of $\mathbf{Y}_D$ is set to the phase of $\mathbf{Y}$, and only the magnitude spectra is considered in Equation 4. The parameters α and β adjust the gain of the subtraction and the flooring level, respectively. These values depend on the noise estimate and are to be determined empirically from a development set. In general, $\beta \ll \alpha$.

In order for AME to adapt the clean MLE speaker models, the statistics of the enhanced speech ($\mathbf{Y}_D$) need to be found as a function of the SS used. Since the predominant model for text-independent speaker verification is the mixture of Gaussians, and the acoustic features are based on MFCCs (Mel-frequency cepstral coefficients),

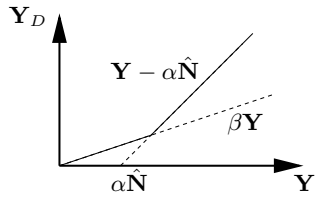

Fig. 1. The solid line shows the non-linear characteristic for SS. The flooring level of Y_D is proportional to the observed noisy signal $\mathbf{Y}$.

such a case will be illustrated. Nevertheless, similar procedures can be devised for other types of spectral-based features or HMM-based models.

First, let us consider $\lambda = (\mathbf{c}, \mu, \mathbf{\Sigma})$, a GMM trained with clean speech and PDF:

$$f_\chi(\mathbf{x}) = \sum_k c_k \mathcal{N}(\mathbf{x}; \mu_k, \mathbf{\Sigma}_k), \qquad \sum_k c_k = 1, \tag{5}$$

where χ is the feature vector in cepstral domain. The covariance matrix $\mathbf{\Sigma}_k$ is often simplified as a diagonal matrix. Then, this model can be easily transformed to log-spectral domain by taking the inverse discrete cosine transform, obtaining $\mathbf{X}^\ell$ with the PDF:

$$f_{\mathbf{X}^\ell}(\mathbf{x}) = \sum_k c_k \mathcal{N}(\mathbf{x}; \mu_k^\ell, \mathbf{\Sigma}_k^\ell), \tag{6}$$

$$\mu_k^\ell = \mathbf{C}^{-1}\mu_k, \qquad \mathbf{\Sigma}_k^\ell = \mathbf{C}^{-1}\mathbf{\Sigma}_k(\mathbf{C}^{-1})^T. \tag{7}$$

Our notation can be reduced if we consider only the k-th Gaussian mixture component. Additionally, by approximating $\mathbf{\Sigma}_k^\ell$ to be a diagonal matrix, a single dimension in $\mathbf{X}^\ell$ can be considered: X^ℓ (a Mel-frequency channel).

The log-spectrum (X^ℓ) can be transformed to magnitude spectrum (X), which has a log-normal distribution with:

$$E(X) = e^{\mu^\ell + (\sigma^\ell)^2/2}, \qquad E(X^2) = e^{2\mu^\ell + 2(\sigma^\ell)^2}. \tag{8}$$

From Equation 4, it can be found that the first and second moments of $\mathbf{Y_D}$, for a single dimension (Y_D), are:

$$\begin{aligned} E(Y_D) = E(Y) &- \alpha\hat{N}\left[1 - F_Y^{(0)}(a)\right] \\ &+ (\beta - 1)\, F_Y^{(1)}(a), \end{aligned} \tag{9}$$

and

$$\begin{aligned} E(Y_D^2) = E(Y^2) &+ \alpha^2\hat{N}^2\left[1 - F_Y^{(0)}(a)\right] \\ &+ 2\alpha\hat{N}\left[F_Y^{(1)}(a) - E(Y)\right] \\ &+ \left(\beta^2 - 1\right) F_Y^{(2)}(a), \end{aligned} \tag{10}$$

where $a = \alpha\hat{N}/(1-\beta)$, and

$$F_Y^{(0)}(a) = \int_{-\infty}^{a} f_Y(y)dy, \tag{11}$$

$$F_Y^{(1)}(a) = \int_{-\infty}^{a} y f_Y(y)dy, \tag{12}$$

$$F_Y^{(2)}(a) = \int_{-\infty}^{a} y^2 f_Y(y)dy. \tag{13}$$

We can use the approximations:

$$E(Y) \approx E(X) + \hat{N}, \tag{14}$$

$$E(Y^2) \approx E(X^2) + 2\hat{N}E(X) + \hat{N}^2, \tag{15}$$

$$f_Y(y) \sim \log \text{ normal}, \tag{16}$$

to finally obtain $E(Y_D)$ and $E(Y_D^2)$ from Equations 9 and 10. Equation 16 allows the transformation of the adapted distribution back to log-spectrum (Y_D^ℓ) with mean and variance:

$$\mu_D^\ell = 2\ln E(Y_D) - 0.5\ln E(Y_D^2), \tag{17}$$

$$(\sigma_D^\ell)^2 = \ln E(Y_D^2) - 2\ln E(Y_D). \tag{18}$$

The resulting adaptation in cepstrum domain for the means and variance is:

$$\mu \rightarrow \mu_D = \mathbf{C}\mu_D^\ell, \tag{19}$$

$$\mathbf{\Sigma} \rightarrow \mathbf{\Sigma}_D = \mathbf{C}\mathbf{\Sigma}_D^\ell\mathbf{C}^T. \tag{20}$$

This adaptation is done for every mixture component, and every speaker's target and anti model, and is named AME^2 since it is an extension of AME^1 (only mean adaptation).

In the case when $\alpha = 0$ (no SS), it is clear from Equations 9 and 10 that the distribution of Y_D becomes the estimated distribution of Y (conventional model adaptation). If $\alpha = \infty$, the distribution of Y_D becomes the distribution of βY.

3. EXPERIMENTS

3.1. Data set

The experiments were conducted using a modified version of TIMIT: SV-TIMIT. By itself, TIMIT is not recommended for exploring new techniques in speaker recognition because of its unrealistic acoustic conditions, and the lack of intra-speaker variability that results in a nearly perfect performance. Nevertheless, this data set can still be useful for isolating the effect of a particular phenomenon [10], as it is the case of additive noise. Moreover, the public availability of TIMIT allows the results to be reproducible.

SV-TIMIT was assembled as follows. The SV task was split into two independent ones: *male* and *female*, resulting in 326 male and 136 female enrolled speakers obtained from the entire training set of TIMIT. This gender split does not entail a simplification because gender attributes can be assumed to be part of the enrollment data and cross-gender trials are easy to detect as impostors. Discarding the 'sa' utterances (sentences intended to show dialectial variants) to avoid acoustic bias, each registered speaker has 8 enrollment utterances from which two are randomly selected and sent to the verification set to induce their target trials. Since the speakers set that conforms the testing part of TIMIT are not present in the training set, they were used as impostors. In SV-TIMIT, a ratio of 4:1 impostor-to-target trials was used. On average, each utterance is about 2.5 seconds of active speech, therefore the enrollment consists of around 15 seconds, and the decision for each verification trial is drawn from only 2.5 seconds of speech. Earlier experiments with TIMIT [11] explored a different configuration of the data, using only the test set (168 speakers) and including 'sa' utterances, for enrollment and verification.

For the first set of experiments the white noise was synthetically produced. The spectral subtraction and AME were performed using the noise estimate $\hat{\mathbf{N}}$, which was fixed to the mean value across the utterance of the actual noise added. Although a frame by frame estimate is more accurate than a fixed one, this value of $\hat{\mathbf{N}}$ sets a reference that is easy to reproduce.

For the second set of experiments, each of the noise added was extrated from NOISEX database [12], and the noise estimate $\hat{\mathbf{N}}$ was computed using the average of 2 seconds of noise. The types of noise employed were: white noise, pink noise, tank noise (Leopard) and fighter jets (F16). Each type of noise was added for SNRs from 0 dB to 25 dB to the speech waveforms. Spectral subtraction and AME^2 were performed using the noise estimate $\hat{\mathbf{N}}$.

3.2. Results

Several approximations were used in the formulation of AME (Equation 14-16). This section experimentally tests the performance of the proposed technique.

Female				
SNR	no-Enh	SS	AME^1	AME^2
clean	.76	-	-	-
25dB	8.1	4.0	4.0	4.7
20dB	14.7	8.4	7.3	8.8
15dB	26.1	16.2	15.8	12.8
10dB	37.5	31.6	29.4	24.3
5dB	47.9	41.6	40.1	36.4
0dB	48.5	45.2	44.5	43.0

Male				
SNR	no-Enh	SS	AME^1	AME^2
clean	.61	-	-	-
25dB	8.0	2.3	2.1	4.2
20dB	16.0	5.1	4.4	6.4
15dB	27.7	14.0	12.6	11.6
10dB	38.5	30.4	29.3	22.7
5dB	46.5	42.2	40.8	35.0
0dB	49.2	47.4	47.0	44.6

Table 1. Equal error rates (EER) in % for different SNRs in dB and speaker models trained under maximum likelihood (MLE). No-Enh shows the results when neither the verification speech nor the models were enhanced, in SS only the speech was enhanced, and in AME both the speech and the models were enhanced, where AME^1 adapts the means only and AME^2 adapts the means and the variance.

The results for white noise synthetically produced are shown in Table 1. In this experiment, β was set to 0.1 and $\alpha = 1$. First, we can observe that Speech enhancement (SS) successfully removes part of the noise at the cost of a non-linear distortion, achieving an improvement in EER, specially for medium-high SNRs.

The early version of AME (presented as AME^1), where only the means are adapted (for each target and anti models), shows a consistent moderate improvement w.r.t SS alone for all SNRs. The extension of AME, proposed in Section 2 (shown as AME^2), considerably outperforms SS alone for medium-low SNRs, but not for higher SNRs where AME^1 does. Moreover, AME^2 outperforms AME^1 for low SNR.

Based on this results, the next set of experiments explore the performance of AME^2 against noisy speech signal, SS and AME^1. The NOISEX database was employed for this task. A comparison table for male and female speakers and for different types of noises is depicted in Table 2.

In general, results for male and female share a consistent trend. The no-Enh column demonstrates how EER is degraded even for a high SNR. The SS column shows the results for enhanced speech.

Male																
Noise	white				pink				tank (Leopard)				jets(F16)			
SNR	No Enh	SS	AME^1	AME^2	No Enh	SS	AME^1	AME^2	No Enh	SS	AME^1	AME^2	No Enh	SS	AME^1	AME^2
clean	0.6															
25dB	28.5	16.0	16.2	9.0	1.7	13.7	12.9	9.0	9.2	2.5	3.83	5.1	14.4	12.4	11.96	9.7
20dB	37.3	25.1	27.4	11.2	4.1	22.7	22.5	11.2	9.7	5.9	7.1	6.9	20.3	22.6	20.3	11.8
15dB	44.2	29.5	40.3	15.4	12.9	30.7	33.2	13.8	11.0	11.4	11.0	8.9	27.9	32.1	28.9	14.1
10dB	47.4	34.4	47.8	21.3	25.2	34.9	43.6	17.8	12.8	20.1	17.0	11.3	38.0	38.5	36.9	17.3
5dB	49.2	42.5	49.5	29.1	39.1	38.8	48.3	24.1	16.3	28.5	22.2	12.9	45.1	40.8	44.3	21.2
0dB	49.7	46.3	49.5	37.2	46.6	42.9	48.7	28.5	21.8	36.0	26.8	14.1	48.0	43.2	48.4	24.1

Female																
Noise	white				pink				tank (Leopard)				jets(F16)			
SNR	No Enh	SS	AME^1	AME^2	No Enh	SS	AME^1	AME^2	No Enh	SS	AME^1	AME^2	No Enh	SS	AME^1	AME^2
clean	0.7															
25dB	27.5	11.0	14.3	11.4	2.5	10.3	10.6	8.9	14.9	4.3	4.7	7.0	18.3	9.5	10.3	10.3
20dB	34.9	18.0	23.8	13.6	5.2	13.9	16.1	11.0	17.2	6.2	6.6	8.7	21.3	15.8	13.6	10.6
15dB	43.0	23.9	37.4	16.9	13.2	20.2	24.9	13.6	19.2	10.6	8.8	11.0	29.3	22.4	21.3	12.8
10dB	47.4	27.6	44.8	20.9	28.3	27.6	37.7	18.7	22.7	17.3	12.5	12.5	41.2	29.3	32.2	15.3
5dB	48.1	34.9	48.8	26.1	41.5	36.3	47.7	23.8	26.8	23.4	16.1	14.3	45.8	36.8	41.9	18.7
0dB	49.6	43.3	50.3	33.0	48.2	42.7	50	28.3	31.9	31.6	20.9	14.3	48.4	42.2	47.1	26.4

Table 2. Equal error rates (EER) in % for different SNRs in dB and speaker models trained under maximum likelihood (MLE). No-Enh shows the results when neither the verification speech nor the models were enhanced, in SS only the speech was enhanced, and in AME^2 both the speech and the models were enhanced. The best (lowest) EER values were selected employing $\alpha = 1$ to $\alpha = 2.5$ and $\beta = 0.1$.

The AME^1 and AME^2 columns depict the best (lowest) EER percentage value obtained for different values of α (1 to 2.5). The choice of α allows to balance the distortion induced from noise under subtraction and a non linear distortion that results from the spectral flooring. Table 2 shows how the performance of AME^2 can dramatically improve when an appropriate value of α is selected. For example, for jet noise at 0 dB, the EER drops from 48.47% to 26.47% for female speakers. However, AME^1 does not show a consistent improvement, when compared to both, No-Enh and SS columns.

4. CONCLUSIONS

In this work, we extended the AME technique for speaker verification by incorporating the corresponding adaptation of the model variances. Using spectral subtraction (SS) as the speech enhancement technique, the extended AME highly outperformed SS alone particularly at moderately low and low SNRs (5dB-15dB) for synthetic white noise, where the adaptation of the variance was found to considerably improve the equal error rate (EER). However, it also degrades at high SNRs (> 20 dB) compared to means adaptation only. For real noise condition, AME ^{2} outperforms SS and AME^1 for SNRs less than 20dB. These results encourge us to work on a better variance approximation. Since the goal of this work was to show the efectiveness of the extended-AME technique to compensate the noise, then Lombard effect was not taken into account. As a future work, we are planning to use this technique with a large vocabulary automatic speech recognition task.

5. REFERENCES

[1] M. Przybocki and A. Martin, "Nist speaker recognition evaluations," in *Proc. of the International Conference on Language Resources and Evaluation (LREC).*, Grenada, Spain, 1998, pp. 331–335.

[2] R. B. Dunn, T. F. Quatieri, D. A. Reynolds, and J. P. Campbell, "Speaker recognition from coded speech and the effects of score normalization," *Proceedings of the Twenty-Sixth Asilomar Conference on Signals, Systems, and Computers*, vol. 2, pp. 1562–1567, 2001.

[3] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 27, pp. 113–120, 1979.

[4] P. Lockwood and J.Boudy, "Experiments with a nonlinear spectral subtraction (NSS), hidden markov models and the projection, for robust speech recognition in cars," *speech Communications*, vol. 11, pp. 215–228, 1992.

[5] M. Gales and S. Young, "Robust continuous speech recognition using parallel model combination," *IEEE Trans. on Speech and Audio Processing*, 1996.

[6] J. Dropo, L. Deng, and A. Acero, "Evaluation of the SPLICE algorithm on the Aurora2 database (web update)," Aalborg, Denmark, 2001, vol. 2, pp. 217–220.

[7] L. Buera, E. Lleida., A. Miguel, and A. Ortega, "Multi-environment models based linear normalization for speech recognition in car conditions," in *Proc. ICASSP*, Montreal, CAN, May 2004.

[8] A. Moreno-Daniel, J. A. Nolazco-Flores, T. Wada, and B.-H. Juang, "Acoustic model enhancement: An adaptation technique for speaker verification under noisy environments," in *Proc. ICASSP*, Honolulu, USA, April 2007.

[9] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of markov chains," *IEEE Trans. on Speech and Audio Processing*, vol. 2, pp. 291–299, Apr. 1994.

[10] J.P. Campbell and D.A. Reynolds, "Corpora for the evaluation of speaker recognition systems," *Proc. ICASSP*, vol. 2, pp. 829–832, 1999.

[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] M. Tomlinson A. Varga, H. Steeneken and D. Jones, "The noisex-92 study on the affect of additive noise on automatic speech recognition," *DRA Speech Research Unit*, 1992.