ACOUSTIC MODEL ENHANCEMENT: AN ADAPTATION TECHNIQUE FOR SPEAKER VERIFICATION UNDER NOISY ENVIRONMENTS

A. Moreno-Daniel [†‡], *J. A. Nolazco-Flores* [‡*], *T. Wada*[†] *B.-H. Juang* [†]

[†]Center for Signal and Image Processing
Georgia Institute of Technology, Atlanta GA, USA

[‡]Computer Science Department
Tecnológico de Monterrey, Campus Monterrey, Monterrey NL, MX

ABSTRACT

This work presents an acoustic model adaptation method for speaker verification (SV) in environments with additive noise. In contrast to traditional acoustic model adaptation techniques that adapt the models parameters based on a model of the noise, acoustic model enhancement (*AME*) belongs to a new scheme in which the models are adapted to the speech enhancement strategy. The theoretical framework is presented for spectral subtraction (*SS*) as the enhancement technique and GMM as the acoustic models. In order to study the effect of additive noise only, a modified TIMIT dataset was used. The experimental setup uses two types of noise: one with fixed spectrum that helps as a proof of concept, and another with time-varying spectrum as a more realistic performance reference for *AME*. The results for this latter type show that at 20 dB SNR, the equal error rate (EER) dropped from 17% to around 8.9% when the noisy speech was enhanced with *SS*, whereas it further dropped to 8.1% with *AME*.

Index Terms— Speaker recognition, robustness, speech enhancement, model adaptation, acoustic model enhancement

1. INTRODUCTION

Secure applications such as electronic banking, e-commerce and access control to restricted areas, where privacy and security of information storage or transference are vital, have become a very important topic in today's daily life. The most common way of access control is the use of passwords to restrict the access to unauthorized users and to secure critical information; however, it is often not secure enough because the passwords can be stolen, intercepted, or even guessed by a clever program if not carefully chosen. On the other hand, the biometrics can be used for user identification, verification, and lately for generating keys for cryptosystems [1], with the advantage that the user does not have to memorize the key nor store it in a secret place.

Among all the biometrics, we will concentrate on speech, particularly in the speaker verification (SV) problem. Since speech in real life applications is generally corrupted by noise, we are specially interested in SV in adverse conditions. Specifically, we will examine the additive noise case, a condition often found when the speaker is located in a public area or in a moving vehicle (neglecting the Lombard effect).

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

Approaches for tackling this problem are diverse; some of these approaches directly process the signal in time, spectral or cepstral domain, recovering the clean speech given the noise estimate [2, 3], while others adapt the acoustic models to the noisy environment [4, 5, 6, 7]. We refer to acoustic model enhancement (*AME*) as the approach that automatically adapts the clean acoustic models to the effects of the enhanced speech, in spectrum [8] or cepstrum [9] (the cepstral subtraction method).

The algorithm proposed in this paper extends the idea shown in [8], a technique for robust ASR using *SS*, to GMM-based (Gaussian mixture model) text independent SV. *SS* was chosen as enhancement technique and GMMs as speaker models because they have been shown to be effective techniques [2, 10].

The organization of this document is as follows. Section 2 explains the speech enhancement algorithm as well as the corresponding theoretical framework for *AME*. Section 3 complements the study with an experimental analysis. Finally, in section 4, conclusions and future work are discussed.

2. ENHANCEMENT

2.1. Enhancing speech with spectral subtraction

The idea behind *SS* is to obtain back the clean speech ($\mathbf{X}$) by subtracting an estimate of the noise spectrum ($\hat{\mathbf{N}} \approx \mathbf{N}$) from the noisy observation ($\mathbf{Y}$):

$$\mathbf{Y} = \mathbf{X} + \mathbf{N}. \tag{1}$$

Although there are a number of sophisticated *SS* techniques derived from [2], we study the Generalized Spectral Subtraction [11]:

$$\begin{aligned} D(\mathbf{Y}) &= \mathbf{Y} - \alpha\hat{\mathbf{N}}, \\ \mathbf{Y_D} &= \max\left(D(\mathbf{Y}), \beta\mathbf{Y}\right). \end{aligned} \tag{2}$$

This *SS* scheme is, in essence, a floored subtraction of $\hat{\mathbf{N}}$ from $\mathbf{Y}$, where an overestimation factor $\alpha > 0$ adjusts the amount subtracted while $0 < \beta \ll 1$ sets the floor level as a fraction of $\mathbf{Y}$. Flooring will occur more often in signals with low SNR (signal to noise ratio).

State of the art *SS* systems update $\hat{\mathbf{N}}$ in the frame by frame basis with the help of a silence detector. In such a case, both the speech and the speaker models would be enhanced every frame with the corresponding latest noise estimate.

For spectral-based features, *SS* can be easily performed in an intermediate step of the front-end stage of the noisy waveform ($\mathbf{y}$). In the particular case of MFCC (as shown in Fig. 1), a smoothed

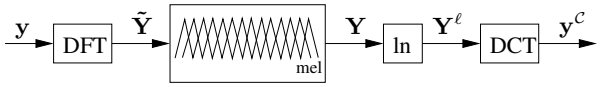

Fig. 1. MFCC (Mel-frequency cepstral coefficients) front-end stage. Log spectral and cepstral domain are denoted with the superscripts ℓ and $\mathcal{C}$, respectively.

version ($\mathbf{Y}$) from the short-time spectrum ($\tilde{\mathbf{Y}}$) is first obtained from a bank of triangular filters uniformly spaced in mel-scale; then the MFCC feature is computed as a truncated cepstrum with a simple linear transformation of the log-spectrum ($\mathbf{Y}^\ell$):

$$\mathbf{Y}^\ell = \ln \mathbf{Y}, \tag{3}$$

$$\mathbf{y}^\mathcal{C} = \mathbf{C} \cdot \mathbf{Y}^\ell, \tag{4}$$

where log-spectral and cepstral domains are denoted with the superscripts ℓ and $\mathcal{C}$, respectively, and $\mathbf{C}$ is a DCT (discrete cosine transform) matrix.

Similarly, the noise estimate ($\hat{\mathbf{N}}$), calculated after the triangular filters, is subtracted from ($\mathbf{Y}$) as in Eqn. 2 to give $\mathbf{Y}_D$, which is later transformed to log-spectrum ($\mathbf{Y}_D^\ell$) and then to the *SS*-cepstral feature ($\mathbf{y}_D^\mathcal{C}$).

2.2. Enhanced acoustic models

In contrast to the traditional acoustic model adaptation techniques that adapt the models parameters based on a model of the noise, *AME* belongs to a new scheme in which the models are adapted to the speech enhancement strategy.

Let $\lambda_i(\mathbf{w}_i, \mu_i, \mathbf{\Sigma}_i)$ be the corresponding GMM for the i-th speaker. The goal of *AME* is to find a transformation $\hat{\mu}_i = T_\mu\{\mu_i\}$ for which the resulting enhanced model $\hat{\lambda}_i(\mathbf{w}_i, \hat{\mu}_i, \mathbf{\Sigma}_i)$ resembles λ_i^*, where λ_i^* is the model that matches the *SS* condition (a model trained with $\mathbf{y}_D^\mathcal{C}$). This section presents a way for finding such a transformation.

Consider a single speaker model λ (without confusion, the subscript i has been dropped) trained with clean speech cepstral features ($\mathbf{x}^\mathcal{C}$). The probability density function (PDF) of the GMM is:

$$f_{\mathbf{x}^\mathcal{C}}(\mathbf{x}) = \sum_m w_m \mathcal{N}(\mathbf{x}; \mu_m, \mathbf{\Sigma}_m); \quad \sum_m w_m = 1. \tag{5}$$

First, the spectral domain is suitable for adapting the models to the effect of *SS* since the noise studied here is additive in linear-spectrum; therefore, the speaker models should be projected to this domain.

The cepstral features ($\mathbf{x}^\mathcal{C}$) can be projected back to log-spectral domain by simply inverting the DCT:

$$\mathbf{X}^\ell = \mathbf{C}^{-1}\mathbf{x}^\mathcal{C}. \tag{6}$$

Consequently, the log-spectral features ($\mathbf{X}^\ell$) have the PDF:

$$f_{\mathbf{X}^\ell}(\mathbf{x}) = \sum_m w_m \mathcal{N}(\mathbf{x}; \mu_m^\ell, \mathbf{\Sigma}_m^\ell); \quad \sum_m w_m = 1; \tag{7}$$

$$\mu_m^\ell = \mathbf{C}^{-1}\mu_m; \qquad \mathbf{\Sigma}_m^\ell = \mathbf{C}^{-1}\mathbf{\Sigma}_m(\mathbf{C}^{-1})'.$$

At this point, we introduce a first simplification to Eqn. 7 by approximating $\mathbf{\Sigma}_m^\ell$ with a diagonal matrix, thus allowing us to treat each dimension (i.e. Mel-frequency bin) independently. Furthermore, without loss of generality, let us consider a single dimensional X^ℓ, with μ_m^ℓ as mean and σ_m^2 as variance of its m-th mixture component.

Then, while the log-spectrum (X^ℓ) distributes as a Gaussian mixture, our simplified linear-spectrum (X) distributes as a log-normal mixture:

$$\begin{aligned} f_X(x) &= \frac{1}{x} f_{X^\ell}(\ln x), \qquad 0 < x < \infty, \\ &= \sum_m \frac{w_m}{x\sigma_m\sqrt{2\pi}} e^{-(\ln x - \mu_m^\ell)^2/(2\sigma_m^2)}, \\ &= \sum_m w_m f_{X,m}(x). \end{aligned} \tag{8}$$

Next, in order to characterize the effects of *SS*, the first two moments of the enhanced speech Y_D are obtained as follows.

2.2.1. *First moment of* Y_D

Given $f_X(x)$, the expression for the PDF of the linear-spectrum of clean features (Eqn. 8), let us find the effect in this distribution caused by the addition of noise followed by *SS* (Eqn. 2). Let us further reduce our notation and consider only the m-th mixture component:

$$h_X(x) \equiv f_{X,m}(x), \tag{9}$$

$$h_Y(y) = h_X(y - \hat{N}), \tag{10}$$

$$\begin{aligned} E(Y_D) = &\int_a^\infty (y - \alpha\hat{N}) h_Y(y) dy \\ &+ \int_{-\infty}^a \beta\, y\, h_Y(y) dy, \end{aligned} \tag{11}$$

where $a \equiv \frac{\alpha\hat{N}}{1-\beta}$.

Equation 11 can be expressed in terms of $E(X)$ as follows:

$$E(Y_D) = E(X) + (1 - \alpha + \alpha A)\hat{N} + (\beta - 1)B, \tag{12}$$

$$A \equiv H_Y(a); \qquad B \equiv \int_{-\infty}^a y h_Y(y) dy,$$

where $H_Y(y)$ is the cumulative distribution function (CDF) of Y. After grouping terms, Eqn. 12 can be rewritten as:

$$E(Y_D) = E(X) + \delta_m, \tag{13}$$

$$\delta_m \equiv (1 - \alpha + \alpha A)\hat{N} + (\beta - 1)B, \tag{14}$$

which is a simple mean shift.

Let us analyze Eqn. 2 and Eqn. 12 for the extreme cases: when $\alpha = 0$, we have a naive model adaptation and no *SS* at all; on the other hand, when $\alpha = 1$ and $\beta = 0$, we have a naive *SS* and no model adaptation at all. Therefore, the terms A and B indeed account for the non-linear distortion caused by the flooring in *SS* as it was defined in Eqn. 2.

2.2.2. *Second moment of* Y_D

A second simplification is introduced by approximating the variance of Y_D by the variance of X, therefore obtaining an expression for the second moment for Y_D:

$$E(Y_D^2) \approx E(X^2) - E(X)^2 + (E(X) + \delta_m)^2. \tag{15}$$

Although it is clear from Eqn. 9 that neither Y nor Y_D has a log-normal distribution, a third simplification is used to make the

problem tractable by fitting Y_D into a log-normal distribution with mean $E(Y_D)$ and variance $E(Y_D^2) - E(Y_D)^2$. This way, the corresponding statistics of the log-spectrum $Y_D^\ell = \ln(Y_D)$ are:

$$\hat{\sigma}^2 = \ln\left[Var(X) + E(Y_D)^2\right] - 2\ln\left[E(Y_D)\right], \quad (16)$$

$$\hat{\mu}^\ell = \ln\left[E(Y_D)\right] - \hat{\sigma}^2/2, \quad (17)$$

where $\hat{\mu}^\ell$ is the new mean and $\hat{\sigma}^2$ is the new variance in log-spectrum domain for a fixed dimension in the m-th mixture component of a given speaker GMM.

Finally, the resulting transformation for mean in cepstral domain is:

$$T_\mu\{\mu\} = \hat{\mu} = \mathbf{C}\hat{\mu}^\ell. \quad (18)$$

3. EXPERIMENTS

These experiments were performed to explore the effect of speech enhancement (*SS*) and *AME* in the text-independent SV task. The SV paradigm followed is explained in subsection 3.1; the modified TIMIT is explained in section 3.2; the experimental conditions explored are presented in subsection 3.3; finally, the results are presented and discussed in subsection 3.4.

3.1. Speaker verification framework

The SV framework is essentially a statistical hypothesis test. The null hypothesis ($\mathcal{H}_0$) denotes the hypothesis to accept the observation as being produced by the target speaker, while the alternative hypothesis ($\mathcal{H}_1$) rejects it. Each trial consists of a speech segment and a claimed identity. The log likelihood ratio (LLR) is used to determine a score (θ). The greater the score, the more likely the trial is indeed a legitimate target trial.

$$\theta = \ln\frac{f_{0,Y}(y)}{f_{1,Y}(y)}; \qquad \theta \begin{array}{c} accept \\ \geq \\ < \\ reject \end{array} \tau. \quad (19)$$

The value of the threshold (τ) is selected to minimize the expected cost for each type of error (I and II), which is application dependent. In our analysis, we consider the EER (equal error rate) as the performance measure.

For $\mathcal{H}_0$, the acoustic model for each target speaker is a Gaussian mixture with 64 components. On the other hand, the corresponding impostor model (for $\mathcal{H}_1$) is meant to provide a contrast capable of distinguishing those speakers that are easily confused with the target. Although properly trained speaker-dependent impostor models are suitable, for the purpose of testing our model adaptation technique under a generic setup, a speaker-independent impostor model is used with a single Gaussian mixture with 64 components UBM (universal background model), trained with the entire enrollment set (no actual impostor data was used in this training).

The feature set used consists of 18 MFCC plus the 0-th cepstral coefficient and their corresponding Δ appended; thus forming a 38 dimensional vector.

3.2. Data set

We used a text-independent SV system with TIMIT as our data set. Since in NIST evaluations data sets [12] it is not possible to isolate the effect of *SS*, we use a modified text-independent SV TIMIT as data set. As it is traditionally done in SV NIST evaluations, our SV system is also gender dependent (a reasonable assumption because any claimed identity can easily have gender attribute); for a total of 326 male and 136 female target speakers.

All "sa" utterances from TIMIT were ignored to avoid acoustic bias; and for every speaker, two randomly selected utterances were moved from the training set (enrollment) to the testing set (verification), inducing this way target trials; while the entire testing set was used for impostor trials since it was recorded from a different speaker set. In total, each putative speaker has 2 target trials and 20 impostor trials. Each trial is on average 2.5 s long, and each speaker model was trained with an average of 7.8 s of active speech.

3.3. Acoustic conditions

For the purpose of studying the effectiveness of the algorithm presented, two types of acoustic conditions are considered: type *A* and *B*. Type *A* artificially adds synthetic noise directly to the spectrum, serving as a proof of concept and a measure of *AME*'s potential (these conditions are enumerated with Roman numbers: ACI-V). Type *B* artificially adds a white noise sequence (from NOISEX-92) to the speech waveform (these conditions are enumerated with Arabic numbers: AC1-5).

On the one hand, type *A* conditions (ACI-V) synthetically add a noise with known and fixed spectrum: $\mathbf{N}_A$, nevertheless, if we knew what the noise spectrum was, one would always achieve perfect enhancement by a naive subtraction. Yet in these experiments, the known noise $\mathbf{N}_A$ was blindly treated as a simple estimate ($\hat{\mathbf{N}}$), therefore the non-linear effects (spectral flooring) due to *SS* are indeed present. For this case, the value of $\hat{\mathbf{N}}$ was obtained from the average spectrum of a Gaussian white noise sequence.

On the other hand, the conditions type *B* (AC1-5) add a white noise signal with an unknown time-varying spectrum $\mathbf{N}_B$. The $\hat{\mathbf{N}}$ was conservatively estimated as the average spectrum from one second of silence before and after the utterance.

A summary of the notation used for all the conditions explored is shown in Table 1. Notice that the conditions ACV and AC5 correspond to the proposed *AME* algorithm, where the clean speaker models have been adapted to work with the speech enhancement technique (*SS*).

The parameter α, as shown in Eqn. 2, adjusts the performance trade-off between distortion due to spectral flooring (controlled by β) and noise under-subtraction. The values for α and β are a function of the noise estimate and the data, thus they are to be determined empirically from a development set. A reasonable value for β is 0.1, and a conservative value for α is 1.0.

3.4. Results

The experimental results for the conditions type *A* (ACI-V) are shown in Table 2. These experiments help as a proof of concept, and represent a performance upper-bound for *AME* because the noise estimate is perfect. The results for conditions type *B* (AC1-5) are shown in Table 3. These experiments represent a performance lower-bound for *AME* because the *SS* and the noise estimate were done in a conservative way.

First, the results from AC0, ACI and AC1 show that the clean models are very sensitive to the presence of even a small amount of noise in the verification utterance. For example, the EER increased from 1.8% in AC0 (clean) to 11% for ACI and to 10% for AC1 at 25 dB SNR for female subjects. Although the noise type *B* is time-variant and the type *A* is fixed, both conditions showed comparable degradation.

Condition	Enroll	Verify
AC0	cln	cln
ACI	cln	cln+$\mathbf{N}_A$
ACII	cln	ss(cln+$\mathbf{N}_A$)
ACIII	cln+$\mathbf{N}_A$	cln+$\mathbf{N}_A$
ACIV	ss(cln+$\mathbf{N}_A$)	ss(cln+$\mathbf{N}_A$)
ACV	*cln**	ss(cln+$\mathbf{N}_A$)
AC1	cln	cln+$\mathbf{N}_B$
AC2	cln	ss(cln+$\mathbf{N}_B$)
AC3	cln+$\mathbf{N}_B$	cln+$\mathbf{N}_B$
AC4	ss(cln+$\mathbf{N}_B$)	ss(cln+$\mathbf{N}_B$)
AC5	*cln**	ss(cln+$\mathbf{N}_B$)

Table 1. Acoustic conditions, where "cln" denotes clean speech, "$\mathbf{N}_A$" denotes white noise type *A* with known and fixed spectrum and "$\mathbf{N}_B$" denotes white noise with unknown time-varying spectrum. The star in "*cln**" means that the clean models were adapted with the proposed *AME* algorithm.

Gender	Condition	5dB	10dB	15dB	20dB	25dB	clean
Female	ACI	47.0	42.0	31.0	18.0	11.0	1.8
	ACII	10.0	4.4	2.2	1.8	1.8	-
	ACIII	5.1	4.8	3.7	3.3	2.2	-
	ACIV	2.2	2.6	2.1	2.2	1.9	-
	ACV	6.2	3.3	1.9	1.8	1.8	-
Male	ACI	47.0	41.0	30.0	19.0	10.0	1.2
	ACII	9.2	4.5	1.8	1.4	1.2	-
	ACIII	5.8	5.1	4.1	3.4	2.5	-
	ACIV	3.6	2.6	1.9	1.8	1.4	-
	ACV	4.6	2.6	1.7	1.5	1.4	-

Table 2. EER for the acoustic conditions type *A*. AC0 (clean condition) is shown on the right-most column of ACI.

Second, the comparison of ACI with ACII and AC1 with AC2 shows how effective the *SS* was in making the verification noisy speech match the clean models. The distortion induced by the spectral flooring (a hard lower clipping) was softened by the cepstral truncation (18 coefficients in this case). For both types of noise (A and B), the *SS* improved the EER. The type A case outperformed type B because of the time-variance of $\mathbf{N}_B$ and its fixed estimate ($\hat{\mathbf{N}}$).

Third, although ACIII, ACIV, AC3 and AC4 are unfeasible scenarios (it would require having a model for every noise condition), they provide a reference to what conventional noise adaptation and the proposed *AME* could each achieve. For most noise levels, ACIV and AC4 (*SS*-matched) outperformed ACIII and AC3 (noise-matched), which is a sign of *AME*'s potential.

Finally, the results for the proposed algorithm (*AME*) are shown in ACV and AC5. With respect to no adaptation at all (ACII and AC2), *AME* improves the EER for every noise level type *A*, while it moderately does for the noise type *B* case, where as the SNR drops, also does the effectiveness of *AME* because $\hat{\mathbf{N}}$ becomes a poor estimate.

4. CONCLUSIONS AND FUTURE WORK

The *AME* method was presented and shown to be effective in reducing the EER, particularly at high SNRs for noise type *A* (easily estimated) and at low SNRs for noise type *B* (poorly estimated). This new method adapts the means of the GMM-based speaker models to match the effect of *SS*, which was chosen as the speech enhancement strategy. A text-independent SV system with relatively short verification utterances (2.5 s long) was used for testing purpose.

For noise type *A*, at 5 dB SNR, the EER dropped from 47% to around 10% when the noisy speech was enhanced with *SS*, whereas it further dropped to around 5% when the *AME* was added. For noise type *B*, at 20 dB SNR, the EER dropped from 17% to 8.9% after *SS*, and it was further reduced to 8.1% with *AME*.

Gender	Condition	5dB	10dB	15dB	20dB	25dB	clean
Female	AC1	47.0	41.0	28.0	17.0	10.0	1.8
	AC2	43.0	35.0	21.0	8.9	4.4	-
	AC3	14.0	9.9	7.0	5.2	3.7	-
	AC4	15.0	9.7	6.2	5.2	3.3	-
	AC5	43.0	35.0	21.0	8.1	4.1	-
Male	AC1	46.0	39.0	29.0	18.0	9.6	1.2
	AC2	40.0	30.0	18.0	8.6	4.1	-
	AC3	15.0	10.0	6.4	4.7	3.2	-
	AC4	15.0	9.1	6.0	3.7	2.5	-
	AC5	40.0	30.0	17.0	7.8	3.6	-

Table 3. EER for the acoustic conditions type *B*. AC0 (clean condition) is shown on the right-most column of AC1.

The results also suggest that the three simplifications taken in the *AME* theoretical framework kept the problem tractable and preserved an acceptable level of performance. Future work includes the incorporation of an improved frame-by-frame noise estimator, the adaptation of the variances, and the use of a larger data set.

5. REFERENCES

[1] U. Uludag, S. Pankanti, S. Prabhakar, and A. K. Jain, "Biometric cryptosystems: Issues and challenges," *Proceedings of the IEEE*, vol. 92, no. 6, pp. 948–960, June 2004.

[2] 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.

[3] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *Journal of the Acoustical Society of America*, vol. 55, pp. 1304–1312, Jun. 1974.

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

[5] A. P. Varga and R. K. Moore, "Hidden markov model decomposition of speech and noise," in *Proc. ICASSP*, Albuquerque, NM, 1990, pp. 845–848.

[6] L. Deng, A. Acero, J. L. Droppo, and J. X. Huang, "High-performance robust speech recognition using stereo training data," in *Proc. ICASSP*, Salt Lake City, UT, May 2001, pp. 301–304.

[7] B. J. Frey, T. T. Kristjansson, L. Deng, and A. Acero, "ALGONQUIN learning dynamic noise models from noisy speech for robust speech recognition," *NIPS*, pp. 1165–1171, 2001.

[8] J.A. Nolazco-Flores and S. Young, "Continuous speech recognition in noise using spectral subtraction and hmm adaptation," in *Proc. ICASSP*, Adelaide, Australia, 1994, pp. 409–412.

[9] H. K. Kim and R. C. Rose, "Cepstrum-domain acoustic feature compensation based on decomposition of speech and noise for asr in noisy environments," *IEEE Trans. on Speech and Audio Processing*, vol. 11, no. 5, pp. 435–446, Sept. 2003.

[10] D. A. Reynolds, "Speaker identification and verification using gaussian mixture speaker models," *Speech Communication*, vol. 17, no. 1-2, pp. 91–108, 1995.

[11] D.V. Compernolle, "DSP techniques for speech enhancement," *ESCA Workshop on Speech Processing in Adverse Conditions*, pp. 21–30, 1992.

[12] 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.