

An Evaluation of VTS and IMM for Speaker Verification in Noise

Suhadi*, Sorel Stan**, Tim Fingscheidt, and Christophe Beaugeant

Siemens AG, ICM Mobile Phones, 81675 Munich, Germany

{firstname}.{lastname}@siemens.com

Abstract

The performance of speaker verification (SV) systems degrades rapidly in noise rendering them unsuitable for security-critical applications in mobile phones, where false acceptance rates (FAR) of $\sim 10^{-4}$ are required. However, less demanding applications for which equal error rates (EER) comparable to word error rates (WER) of speech recognizers are acceptable could benefit from the SV technology.

In this paper we evaluate two feature-based noise compensation algorithms in the context of SV: vector Taylor series (VTS) combined with statistical linear approximation (SLA), and Kalman filter-based interacting multiple models (IMM). Tests with the YOHO database and the NTT-AT ambient noises show that EERs as low as 5%–10% in medium to high noise conditions can be achieved for a text-independent SV system.

1. Introduction

Recent studies [1] show that state-of-the-art SV technologies achieve EERs ranging from 0.1% for text-dependent systems trained and tested using single-microphone clean data to 25% or higher for text-independent systems and military radio data. Somewhere in the middle lies the performance of systems for telephony with EERs anywhere between 1% for text-dependent, digit strings combinations, to 10% for text-independent, conversational data.

Although SV biometric does not seem to fully qualify for security-critical applications (e.g. e-banking) yet, one could think of an entire spectrum of applications with relaxed security requirements such as password-protected user profiles or screen savers for mobile phones, given an acceptable performance in the typically noisy operating environments of these devices.

In this paper we evaluate in the context of SV two recent algorithms developed for noise robust speech recognition (SR) using Mel-frequency cepstral coefficients (MFCC) [2] as features: Kalman filter IMM [3] and an extension of VTS [4, 5, 6, 7]. We use a text-independent SV system, based on Gaussian mixture models (GMM) and an MFCC front-end [8]. The EERs are computed on speech data contaminated with car and street noise.

Our paper is organized as follows: in Section 2 we describe briefly the SV system employed, then we present in Section 3 the noise compensation algorithms, followed by the experimental results in Section 4, and the conclusions in Section 5.

2. Speaker Verification Approach

2.1. Feature Extraction

We employ a standard front-end processing based on MFCC features [2, 8]. The speech signal sampled at 8 kHz is divided

into overlapping frames of 32 ms length with a frame shift of 15 ms. Energy-based voice activity detection (VAD) is performed on each frame to discard silence segments.

Each frame is multiplied by a Hamming window prior to transformation by a 256-point FFT. The squared spectral magnitudes are pre-emphasized by a first order FIR filter and then transformed to the Mel-frequency domain using $P = 15$ triangle-shaped filters. After taking the logarithm a vector of 15 *log-spectral* coefficients \mathbf{x}_t is obtained for frame t . Subsequent application of the discrete cosine transformation (DCT) yields a vector of 12 *cepstral* coefficients $\mathbf{x}_t^{\text{cep}}$.

The convolutional effects of a linear transmission channel appear in the cepstral domain as additive biases, which can be eliminated by estimating the mean and subtracting it from each cepstral feature vector. Cepstral mean subtraction (CMS) is part of the standard front-end processing in SV systems [8], and we also apply it to obtain channel-compensated feature vectors.

2.2. Models and Training

The world and the speaker are represented in the *cepstral* domain using GMMs with $J = 32$ Gaussian probability density functions (PDF) each. A GMM is completely characterized by the parameter set $\lambda = \{w_j, \mu_j, \Sigma_j\}_{j=1, \overline{J}}$ and the density

$$p(\mathbf{x}_t^{\text{cep}} | \lambda) = \sum_{j=1}^J w_j \mathcal{N}(\mathbf{x}_t^{\text{cep}}; \mu_j, \Sigma_j), \quad (1)$$

where $\mathcal{N}(\cdot)$ is a Gaussian PDF and w_j is the mixing weight satisfying $\sum_{j=1}^J w_j = 1$.

The world model λ_w is obtained via EM training [9] from data pooled together from 20 male and 20 female speakers, with each speaker contributing 1–2 minutes of speech. A speaker model λ_s is derived from the world model via Bayesian adaptation of means and variances [8] using about 1 minute of speech. We trained 18 gender-balanced speaker models, i.e. 9 female and 9 male. All training data is taken from the YOHO database [10].

2.3. Acceptance/Rejection Criterion

Given a sequence of cepstral vectors $\mathbf{X}^{\text{cep}} = \{\mathbf{x}_t^{\text{cep}}\}_{t=1, \overline{T}}$ the decision rule for identifying the authorized user or the impostor is based on comparing the log-likelihood ratio with a threshold

$$LLR = \log \frac{p(\mathbf{X}^{\text{cep}} | \lambda_s)}{p(\mathbf{X}^{\text{cep}} | \lambda_w)} \geq \theta, \quad (2)$$

where “>” means acceptance and “<” means rejection. It is common practice to make the approximation

$$\log p(\mathbf{X}^{\text{cep}} | \lambda) \approx \sum_{t=1}^T \log p(\mathbf{x}_t^{\text{cep}} | \lambda), \quad (3)$$

which holds exactly if and only if the feature vectors are statistically independent.

*Suhadi was on leave from the Technical University of Hamburg-Harburg.

**Corresponding author.

3. Noise Compensation Approaches

3.1. Environment Model

The effect of the environment on the clean speech can be modeled in the time domain by a linear transmission channel and additive noise as shown in Fig. 1. Having incorporated CMS in

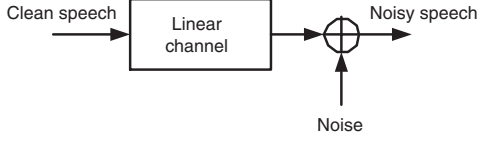


Figure 1: *Environment model.*

our system for channel normalization (see Section 2.1), we will consider only the effect of the additive noise.

Although linear in the time domain, the noise contamination rule in the log-spectral domain changes to the non-linear function

$$\mathbf{z}_t = \mathbf{f}(\mathbf{x}_t, \mathbf{n}_t) = \mathbf{x}_t + \log[\mathbf{1} + \exp(\mathbf{n}_t - \mathbf{x}_t)] \quad (4)$$

where \mathbf{x}_t , \mathbf{n}_t , and \mathbf{z}_t denote the clean speech, noise, and noisy speech *log-spectral* vectors, respectively [4].

3.2. Linear Approximations

Assuming that the PDF of clean speech *log-spectra* can be well represented by a GMM with $K = 32$ components $p(\mathbf{x}_t) = \sum_{k=1}^K w_{\mathbf{x},k} \mathcal{N}(\mathbf{x}_t; \mu_{\mathbf{x},k}, \Sigma_{\mathbf{x},k})$ and the PDF of noise *log-spectra* is a single Gaussian $p(\mathbf{n}_t) = \mathcal{N}(\mathbf{n}_t; \mu_{\mathbf{n}}, \Sigma_{\mathbf{n}})$, what is the distribution of the noisy speech log-spectral vectors \mathbf{z}_t ? Looking at the non-linear model from Eq. 4, we find no easy answer. However, if we approximate the noise contamination rule in each GMM component by a linear model

$$\mathbf{z}_t \approx \mathbf{A}_k \mathbf{x}_t + \mathbf{B}_k \mathbf{n}_t + \mathbf{c}_k, \quad k = 1, \dots, K, \quad (5)$$

then each Gaussian of clean speech log-spectra will transform into a Gaussian of noisy speech log-spectra. Note that \mathbf{A}_k and \mathbf{B}_k are $P \times P$ matrices and \mathbf{c}_k is a $P \times 1$ vector. If the matrices $\{\mathbf{A}_k, \mathbf{B}_k\}_{k=1, \dots, K}$ are diagonal, the transformation models the shifting of means and scaling of variances.

Moreno [4, 5] follows this approach and proposes the Taylor series as a way to linearize Eq. 4 around the mean of each Gaussian component. Furthermore a batch EM algorithm is employed to learn the noise statistics and compute the PDF of noisy speech. Once the distribution of noisy speech is known, the unobserved clean speech log-spectra are obtained from the noisy speech log-spectra via minimum mean squared error (MMSE) estimation.

Raj *et al.* [6] argue that it is more important for a linear approximation of the noise contamination function to optimally represent the parameters of the Gaussians describing the noisy speech log-spectra rather than the function itself. A third order polynomial approximation of Eq. 4 is used as the basis for deriving the optimal slope and intercept of a linear model.

Kim [7] introduces SLA, where the expected error between a higher order Taylor series expansion of the non-linear noise contamination rule and the linear model is minimized, giving similarly to [6] a statistical meaning to the optimal linear approximation. It is found that a third order Taylor series suffices.

3.3. VTS-SLA

We present a batch EM algorithm to estimate the statistics of noise log-spectra. Given are the GMM model parameters of the clean speech log-spectra $\lambda_{\mathbf{x}} = \{w_{\mathbf{x},k}, \mu_{\mathbf{x},k}, \Sigma_{\mathbf{x},k}\}_{k=1, \dots, K}$ and the observed sequence of noisy log-spectral vectors $\mathbf{Z} = \{\mathbf{z}_t\}_{t=1, \dots, T}$. The goal of the EM algorithm is to obtain an estimate $\hat{\lambda}_{\mathbf{n}} = \{\hat{\mu}_{\mathbf{n}}, \hat{\Sigma}_{\mathbf{n}}\}$ of the noise PDF parameters. It starts with an initial guess for $\hat{\lambda}_{\mathbf{n}} = \hat{\lambda}_{\mathbf{n}}^{(0)}$ computed from the first few frames, and improves it iteratively to become $\hat{\lambda}_{\mathbf{n}}^{(i)}$ in the i -th step.

In the i -th step, we first apply the third order SLA to linearize $\mathbf{f}(\mathbf{x}_t, \mathbf{n}_t)$ around the mean log-spectral vectors $\mu_{\mathbf{x},k}$ and $\hat{\mu}_{\mathbf{n}}^{(i-1)}$ for each k [7]. This computation yields $\{\hat{\mathbf{A}}_k, \hat{\mathbf{B}}_k, \hat{\mathbf{c}}_k\}_{k=1, \dots, K}$ in the i -th step¹. The linear approximation maps the k -th clean speech Gaussian component to the corresponding noisy speech Gaussian with mean and covariance

$$\begin{aligned} \hat{\mu}_{\mathbf{z},k}^{(i-1)} &= \hat{\mathbf{A}}_k \mu_{\mathbf{x},k} + \hat{\mathbf{B}}_k \hat{\mu}_{\mathbf{n}}^{(i-1)} + \hat{\mathbf{c}}_k \\ \hat{\Sigma}_{\mathbf{z},k}^{(i-1)} &= \hat{\mathbf{A}}_k \Sigma_{\mathbf{x},k} \hat{\mathbf{A}}_k' + \hat{\mathbf{B}}_k \hat{\Sigma}_{\mathbf{n}}^{(i-1)} \hat{\mathbf{B}}_k', \end{aligned} \quad (6)$$

which determines the distribution of \mathbf{z} for a given $\hat{\lambda}_{\mathbf{n}}^{(i-1)}$ as the GMM model $\hat{\lambda}_{\mathbf{z}}^{(i-1)} = \{w_{\mathbf{z},k}, \hat{\mu}_{\mathbf{z},k}^{(i-1)}, \hat{\Sigma}_{\mathbf{z},k}^{(i-1)}\}_{k=1, \dots, K}$, where $w_{\mathbf{z},k} = w_{\mathbf{x},k}$ are fixed. The prime denotes transposition.

Let us define $\log \mathcal{L}(\hat{\lambda}_{\mathbf{n}}^{(i)} | \mathbf{Z}) = \log p(\mathbf{Z} | \hat{\lambda}_{\mathbf{n}}^{(i)})$ to be the log-likelihood of $\hat{\lambda}_{\mathbf{n}}^{(i)}$. In order to make the optimization of the log-likelihood mathematically tractable, we assume apart from the *observed* data \mathbf{Z} the existence of some additional *unobserved* data \mathcal{K} . The unobserved data $\mathcal{K} = \{k_t\}_{t=1, \dots, T}$ with $k_t \in \{1, \dots, K\}$ specifies the sequence of dominant mixture components of $\hat{\lambda}_{\mathbf{z}}^{(i-1)}$ generating $\mathbf{Z} = \{\mathbf{z}_t\}_{t=1, \dots, T}$.

Extending the log-likelihood by taking into account the observed data as well as the unobserved data $(\mathbf{Z}, \mathcal{K})$, the expectation value w.r.t. the unobserved data \mathcal{K} given the observation data \mathbf{Z} and the previous step's estimation of the noise statistics $\hat{\lambda}_{\mathbf{n}}^{(i-1)}$ can be written as (called E-step)

$$Q(\lambda_{\mathbf{n}}, \hat{\lambda}_{\mathbf{n}}^{(i-1)}) = E \left\{ \log p(\mathbf{Z}, \mathcal{K} | \lambda_{\mathbf{n}}) | \mathbf{Z}, \hat{\lambda}_{\mathbf{n}}^{(i-1)} \right\}. \quad (7)$$

This is to be maximized with respect to $\lambda_{\mathbf{n}}$ yielding the next estimate (called M-step)

$$\hat{\lambda}_{\mathbf{n}}^{(i)} = \arg \max_{\lambda_{\mathbf{n}}} Q(\lambda_{\mathbf{n}}, \hat{\lambda}_{\mathbf{n}}^{(i-1)}). \quad (8)$$

In a bit more detail, computation of the E-step leads to

$$Q(\lambda_{\mathbf{n}}, \hat{\lambda}_{\mathbf{n}}^{(i-1)}) = \sum_{k=1}^K \sum_{t=1}^T \log p(\mathbf{z}_t, k | \lambda_{\mathbf{n}}) \cdot p(k | \mathbf{z}_t, \hat{\lambda}_{\mathbf{n}}^{(i-1)}), \quad (9)$$

with the contribution of the k -th Gaussian to the PDF $p(\mathbf{z}_t | \lambda_{\mathbf{n}})$

$$p(\mathbf{z}_t, k | \lambda_{\mathbf{n}}) = w_{\mathbf{z},k} \mathcal{N}(\mathbf{z}_t; \mu_{\mathbf{z},k}, \Sigma_{\mathbf{z},k}), \quad (10)$$

and the probability of the k -th Gaussian mixture component

$$p(k | \mathbf{z}_t, \hat{\lambda}_{\mathbf{n}}^{(i-1)}) = \frac{p(\mathbf{z}_t, k | \hat{\lambda}_{\mathbf{n}}^{(i-1)})}{\sum_{\mathcal{K}=1}^K p(\mathbf{z}_t, \mathcal{K} | \hat{\lambda}_{\mathbf{n}}^{(i-1)})}, \quad (11)$$

simply computed by filling the known terms from Eq. 6 into

$$p(\mathbf{z}_t, k | \hat{\lambda}_{\mathbf{n}}^{(i-1)}) = w_{\mathbf{z},k} \mathcal{N}(\mathbf{z}_t; \hat{\mu}_{\mathbf{z},k}^{(i-1)}, \hat{\Sigma}_{\mathbf{z},k}^{(i-1)}). \quad (12)$$

¹For simplicity, the superscript i is omitted here.

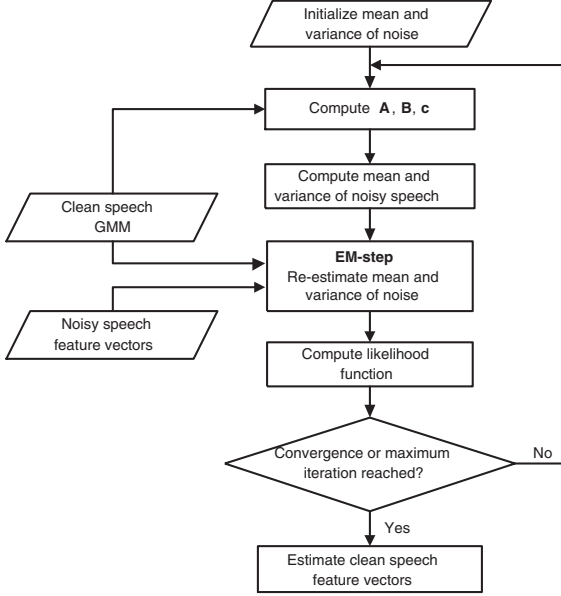


Figure 2: Noise statistics estimation using EM.

The unknown model parameters in Eq. 10 are replaced by

$$\begin{aligned} \mu_{\mathbf{z},k} &= \hat{\mathbf{A}}_k \mu_{\mathbf{x},k} + \hat{\mathbf{B}}_k \mu_{\mathbf{n}} + \hat{\mathbf{c}}_k \\ \Sigma_{\mathbf{z},k} &= \hat{\mathbf{A}}_k \Sigma_{\mathbf{x},k} \hat{\mathbf{A}}_k' + \hat{\mathbf{B}}_k \Sigma_{\mathbf{n}} \hat{\mathbf{B}}_k' \end{aligned} \quad (13)$$

where the only remaining unknowns $(\mu_{\mathbf{n}}, \Sigma_{\mathbf{n}})$ constitute the model $\lambda_{\mathbf{n}}$ which is subject to optimization in Eq. 8.

The E-step and the M-step are repeated either until the log-likelihood has converged, or up to a maximum iteration i_{\max} . A flowchart description of the algorithm is given in Fig. 2.

Once the noise statistics $\hat{\lambda}_{\mathbf{n}} = \hat{\lambda}_{\mathbf{n}}^{(i_{\max})}$ has been estimated, the log-spectral vectors of the clean speech are MMSE estimated by

$$\hat{\mathbf{x}}_t = E \left[\mathbf{x}_t \mid \mathbf{z}_t, \hat{\lambda}_{\mathbf{n}} \right] = \int_{\mathbf{x}} \int_{\mathbf{n}} \mathbf{x}_t p(\mathbf{x}_t, \mathbf{n}_t \mid \mathbf{z}_t, \hat{\lambda}_{\mathbf{n}}) d\mathbf{x}_t d\mathbf{n}_t, \quad (14)$$

which reduces to

$$\hat{\mathbf{x}}_t = \mathbf{z}_t - \sum_{k=1}^K p(k \mid \mathbf{z}_t, \hat{\lambda}_{\mathbf{n}}) \{ (\hat{\mathbf{A}}_k - I) \mu_{\mathbf{x},k} + \hat{\mathbf{B}}_k \hat{\mu}_{\mathbf{n}} + \hat{\mathbf{c}}_k \} \quad (15)$$

by using Eq. 5. Note that in Eq. 15 above also the parameters $\{\hat{\mathbf{A}}_k, \hat{\mathbf{B}}_k, \hat{\mathbf{c}}_k\}$ depend on the noise statistics estimate $\hat{\lambda}_{\mathbf{n}}$.

3.4. IMM

In contrast to VTS-SLA, IMM provides an estimate $\hat{\lambda}_{\mathbf{n},t}$ for each log-spectral vector \mathbf{z}_t [3]. The algorithm employs a bank of K Kalman filters, which share the state transition equation but have different observation models

$$\begin{aligned} \mathbf{n}_t &= \mathbf{n}_{t-1} + \mathbf{u}_{t-1} \\ \mathbf{z}_t &= \hat{\mathbf{A}}_k \mathbf{x}_t + \hat{\mathbf{B}}_k \mathbf{n}_t + \hat{\mathbf{c}}_k. \end{aligned} \quad (16)$$

Note that the noise log-spectral vector is treated as the state of interest, and \mathbf{u}_t is a zero-mean Gaussian process with covariance matrix $\Sigma_{\mathbf{u}}$.

The sequential IMM (S-IMM) algorithm uses an initial guess for $\hat{\lambda}_{\mathbf{n},t=1}$ and applies the Kalman prediction/update scheme to get an estimate of noise statistics for each mixture component. The K estimates are combined in the mixing step to obtain a single estimate

$$\begin{aligned} \hat{\mu}_{\mathbf{n},t} &= \sum_{k=1}^K p(k \mid \mathbf{z}_t) \hat{\mu}_{\mathbf{n},t,k} \\ \hat{\Sigma}_{\mathbf{n},t} &= \sum_{k=1}^K p(k \mid \mathbf{z}_t) \left(\hat{\Sigma}_{\mathbf{n},t,k} + \Delta \hat{\mu}_{\mathbf{n},t,k} \Delta \hat{\mu}_{\mathbf{n},t,k}' \right) \end{aligned} \quad (17)$$

where $\mathbf{Z}_t = \{\mathbf{z}_1, \dots, \mathbf{z}_t\}$ and $\Delta \hat{\mu}_{\mathbf{n},t,k} = (\hat{\mu}_{\mathbf{n},t,k} - \hat{\mu}_{\mathbf{n},t})$. This estimate is used to recompute $\hat{\mathbf{A}}_k, \hat{\mathbf{B}}_k, \hat{\mathbf{c}}_k$ for the next frame and as the initial value for $\hat{\lambda}_{\mathbf{n},t+1}$.

If batch processing is allowed, then S-IMM is applied both forward and backward on the buffered sequence \mathbf{Z} to obtain two estimates of the noise parameters for each frame t , which are subsequently combined in a single estimate

$$\begin{aligned} \hat{\Sigma}_{\mathbf{n},t}^{-1} &= \left(\hat{\Sigma}_{\mathbf{n},t}^f \right)^{-1} + \left(\hat{\Sigma}_{\mathbf{n},t}^b \right)^{-1} \\ \hat{\mu}_{\mathbf{n},t} &= \hat{\Sigma}_{\mathbf{n},t} \left[\left(\hat{\Sigma}_{\mathbf{n},t}^f \right)^{-1} \hat{\mu}_{\mathbf{n},t}^f + \left(\hat{\Sigma}_{\mathbf{n},t}^b \right)^{-1} \hat{\mu}_{\mathbf{n},t}^b \right]. \end{aligned} \quad (18)$$

The superscripts ‘‘f’’ and ‘‘b’’ denote the forward and the backward estimate, respectively. This method of estimation is called fixed interval smoothing IMM (FIS-IMM) [3]. It is important to note that FIS-IMM provides an estimate of the noise parameters for each frame t , similarly to S-IMM.

4. Experimental Results

Both training and testing data is taken from the YOHO database [10]. The testing data is contaminated with noises from the NTT-AT database [11] as follows. The clean speech signal is level normalized to -26 dB using the ITU tools [12]. The noise signal level is adjusted to yield SNR levels between 0 and 40 dB after addition to the clean speech signal. The experiments are conducted on clean speech contaminated with car or street noise.

For testing each of the 18 speakers acts once as the authorized user and 17 times as the impostor for the other speakers. The testing data is different from the training data and consists of 2 to 3 minutes utterances per speaker.

Based on the pooled male and female training data that was used for the GMM world model λ_w in the cepstral domain, log-spectral vectors \mathbf{x} are taken to compute the clean speech GMM model $\lambda_{\mathbf{x}}$ via EM iterations. Three algorithms using $\lambda_{\mathbf{x}}$ for noise compensation, namely VTS-SLA, S-IMM, and FIS-IMM, were applied to the testing data. Initialization of noise statistical parameters of all algorithms was taken from the first five frames, and the maximum number of iterations in the VTS algorithm is set to three.

The SV performance was evaluated by its equal error rate (EER) based on the LLR scores from Eq. 2. The frame shift for LLR computation is a single feature vector and the segment length in Eq. 3 is $T = 200$. Taking the mean of the LLRs over all segments, one gets 18×18 scores. 18 of them belong to the true speakers (target scores), the rest to the impostors (non-target scores). The EER is determined as the intersection of the cumulative density functions of the two sets of scores.

Figs. 3 and 4 show that the plain SV system *without* any noise compensation approach achieves a good EER = 2.43% on clean

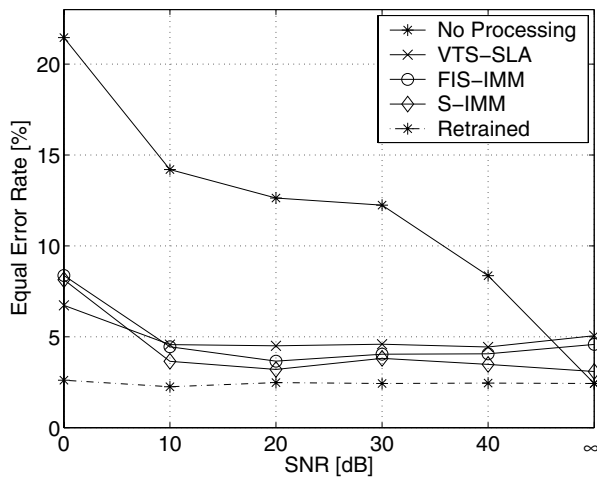


Figure 3: EER for speech in car noise.

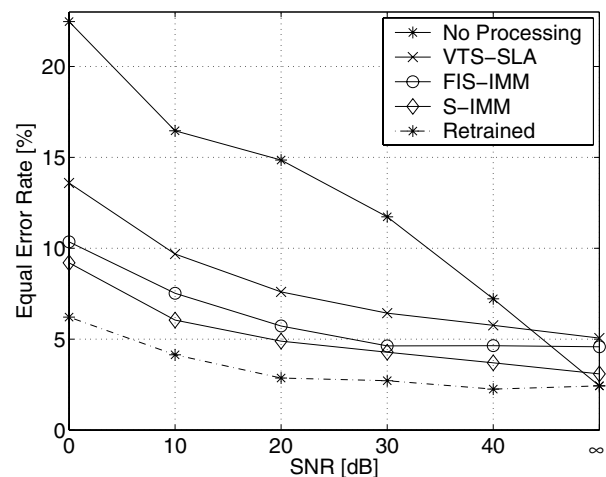


Figure 4: EER for speech in street noise.

speech. On noisy data, the SV performance degrades severely up to an EER around 22% for SNR = 0 dB. As a second reference, the EER system performance is given when it is simply retrained on noisy data.

Looking at the results for the SV system *with* a noise compensation technique, it can be clearly seen that a significant improvement is achievable. Particularly in car noise, the EER improves drastically to values of 6% to 8% for SNR = 0 dB. In general, S-IMM turns out to perform slightly better than any other algorithm tested. This is a surprising result, since S-IMM is a real-time approach, while the other two compensation techniques are processed in batch mode.

In Fig. 3, it is shown that all algorithms in car noise have more or less similar performance. This was expected, because the car noise exhibits stationary characteristics which allows for a single estimate of the noise statistics for the whole utterance as it is done in VTS-SLA. VTS-SLA reveals a good performance especially in very low SNR conditions.

Fig. 4 shows that in highly non-stationary street noise both versions of IMM perform better than the VTS-SLA algorithm, as they provide a new estimate of the noise for each log-spectral vector and are thus able to track the non-stationary characteristics. Performing noise compensation to a long utterance in highly non-stationary noise becomes a disadvantage of our specific VTS-SLA implementation, since a single estimate for the whole utterance is then no longer sufficient.

5. Conclusions

In this contribution we evaluated a batch and a sequential processing variant of the IMM algorithm as well as an extension of VTS, called VTS-SLA. All three algorithms investigated in our baseline text-independent SV system show a significant decrease of EER down to a range of 5% to 10% for medium to high background noise levels.

For street noise, which is highly non-stationary, both IMM variants perform better than VTS-SLA. This is mostly due to the fact that the IMM algorithms compute an estimate of the noise parameters for each frame. This allows for better tracking of the background noise statistics than VTS-SLA. In car noise however, which exhibits mostly stationary characteristics, all algorithms perform about equally well.

6. References

- [1] Reynolds, D. A. and Heck, L. P., "Speaker Verification Tutorial", Proc. of ICASSP, Salt Lake City, UT, May 2001.
- [2] Davis, S. B. and Mermelstein, P., "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences", IEEE Trans. on ASSP, 28(4):357-366, Aug. 1980.
- [3] Kim, N. S., "Feature Domain Compensation of Nonstationary Noise for Robust Speech Recognition", Speech Communication, 37:231-248, 2002.
- [4] Moreno, P. J., *Speech Recognition in Noisy Environments*, Ph.D. Thesis, Carnegie Mellon University, 1996.
- [5] Moreno, P. J., Raj, B., and Stern, R. M., "A Vector Taylor Series Approach for Environment Independent Speech Recognition", Proc. of ICASSP, Atlanta, GA, May 1996.
- [6] Raj, B., Gouvea, E. J., Moreno, P. J., and Stern, R. M., "Cepstral Compensation by Polynomial Approximation for Environment-Independent Speech Recognition", Proc. ICSLP, pp. 2340-2343, Philadelphia, PA, Oct. 1996.
- [7] Kim, N. S., "Statistical Linear Approximation for Environment Compensation", IEEE Signal Processing Letters, 5(1):8-10, 1998.
- [8] Reynolds, D.A., Quatieri, T. F., and Dunn R. B., "Speaker Verification Using Adapted Gaussian Mixture Models", Digital Signal Processing, 10:19-41, Oct. 2000.
- [9] Dempster, A., Laird, N., and Rubin, D., "Maximum Likelihood from Incomplete Data via the EM Algorithm", J. Royal Stat. Soc., vol. 39, 1977.
- [10] Campbell, J. P. and Reynolds D. A., "Corpora for the Evaluation of Speaker Recognition Systems", Proc. of ICASSP, Phoenix, AZ, March 1999.
- [11] NTT-AT, "Ambient Noise Database for Telephony", <http://www.ntt-at.com/products_e/noise-DB/>, Tokio, Japan, 1996.
- [12] ITU-T, "SVP56: The Speech Voltmeter", in *Software Tool Library 2000 User's Manual*, pp. 151-161, Geneva, Switzerland, December 2000.