# SPECTRAL SUBTRACTION AND RASTA-FILTERING IN TEXT-DEPENDENT HMM-BASED SPEAKER VERIFICATION

*Detlef Hardt* [1] *and Klaus Fellbaum* [2]

[1]Institute of Telecommunication and Theoretical Electrical Engineering, Technical University of Berlin, Germany
[2]Communications Engineering, Brandenburg Technical University of Cottbus, Germany

## ABSTRACT

*In real text-dependent telephone-based speaker verification systems, both, additive and convolutional noise influence the error rate considerably. In this paper, different procedures which make a speaker verification system more robust against noise are compared. We either use the spectral subtraction in addition to the MFCC-feature extraction or only the PLP and RASTA-PLP (without spectral subtraction). Considering spectral subtraction two modifications were examined: one version which was pre-connected to the system and a second one being integrated into the MFCC computation. The first version has the advantage that the window length can be chosen independently on those of the MFCC procedure. This led to better results. However, the most effective procedure for telephone speech data is the J-RASTA-PLP, but the estimation of the optimal J factor is difficult. At first we used a fixed J factor based on the off-line measurement of the noise power. Finally, we performed some experiments to optimize the system with the adaptive estimation of the J factor during the utterance. This procedure is based on the method of spectral mapping which has been shown to be very effective in automatic speech recognition.*

## 1. INTRODUCTION

Speaker identity is important for many applications such as access control, automatic money transfer, telephone shopping, etc.. The quality of telephone-based speaker verification (SV) systems depends on the noise power of the speech data and the telephone channel. There are several well known procedures for automatic speech recognition, which make those systems more robust against noise [3, 4, 8]. In this paper, we compare some of these procedures for speaker verification. In our experiments, the speech data were disturbed by additive and convolutive noise.

## 2. DESCRIPTION OF THE DATABASES

We used two databases (IFT and TUBTEL), both consist of the same German sentences which include 18 phonemes [1, 2].

**Table 1.** Summary of the main features of both speech databases

| speech data | TUBTEL | IFT |
|---|---|---|
| number of speakers | 50 | 10 |
| quality ($f_T$=8 kHz) | ISDN (G711) | 0,3 .. 3,4 kHz |
| number of repetitions (test / training) | 14 (9 / 5) | 24 (18 / 6) |

The TUBTEL corpus was collected in a real telephone environment and contained convolutional noise from telephone channel. White noise was added to the clean speech in order to test the speaker verification system at different SNR levels.

## 3. VERIFICATION SYSTEM

The speaker verification system shown in fig. 1 consists of the following components: *spectral subtraction, endpoint detection, feature extraction* and *classification*. Features can be extracted from MFC-, PLP- [3] and RASTA-PLP-coefficients [4]. For the classification we used a modified HMM recognizer according to the HMM Toolkit (HTK) [5].

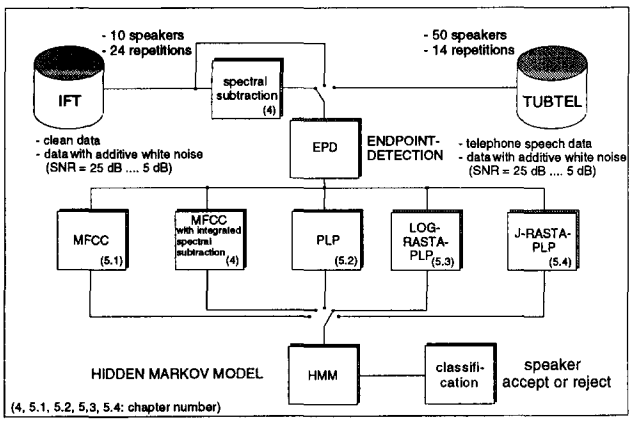


**Figure 1.** Block diagram of the speaker verification system

## 4. SPECTRAL SUBTRACTION

The spectral subtraction (SS) was either pre-connected (external) to the system or integrated into the MFCC feature extraction. If the additive noise $n(t)$ is stationary and uncorrelated to the clean speech signal $s(t)$, then the power spectrum of the noisy speech $u(t)$ is the sum of both power spectra. The clean speech spectrum can be estimated by a simple spectral subtraction of the noise spectrum weighted by a parameter $a$:

$$\hat{P}_s(\Omega) = P_u(\Omega) - a \cdot P_n(\Omega) \qquad (1)$$

The optimal analysis window length of the spectral subtraction, which was integrated into a MFCC feature extraction, was determined by the feature extraction. Bad results were achieved with a window length of 32 ms per frame. These results are due to the correlation between speech and noise signal. In this case the equal error rates (EERs) were above that of the system with-

out spectral subtraction (fig. 2). By using an external spectral subtraction pre-connected to the system (fixed noise and a window length of 128 ms for the SS) better results were achieved (fig.2). In case of non stationary (car) noise the spectral subtraction was less successful [6].

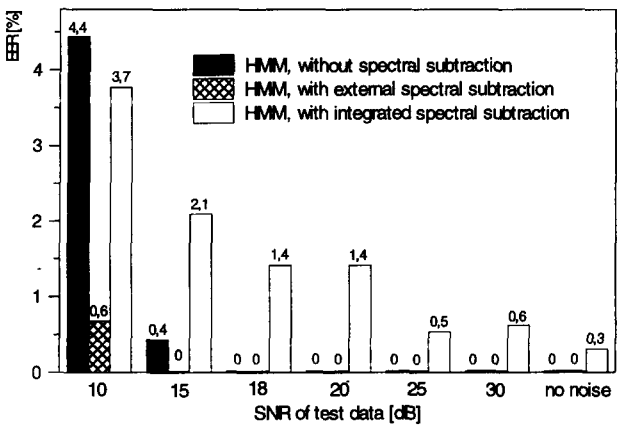


**Figure 2.** EER versus different SNR levels of the test data (IFT corpus, MFCC: N=16, white noise, HMM: 55 states, 2 mix.)

## 5. FEATURE EXTRACTION & RASTA-FILTERING

The following experiments were carried out without spectral subtraction in order to directly compare the influence of the MFCC-, PLP- and RASTA-PLP analysis against noise. All results are based on the comparisons between undistorted training data and test data with different SNR levels.

### 5.1 MFCC-Analysis

Figure 3 compares the EER from the number of states of the HMM for different SNR levels of the test data. For 18 phonemes per sentence and Q=2 to 3 states per phoneme, there are either Q=36 or Q=54 states per sentence possible. Additionaly the HTK required 2 non emitting states. For a MFCC model order of N=16, the best EER=0.06 % was achieved with Q=56 states (TUBTEL corpus). A disadvantage of the MFCC feature extraction is the increase of the EERs during additive noise [6].

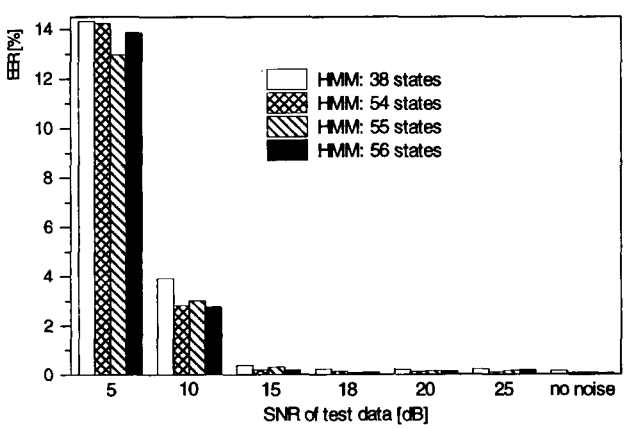


**Figure 3.** EER versus different SNR levels of the test data (TUBTEL corpus, MFCC: N=16, white noise, HMM: 1 mix.)

### 5.2 PLP-Analysis

For an HMM system with 38 states and a PLP model of order N=20 we got the best results (IFT corpus).

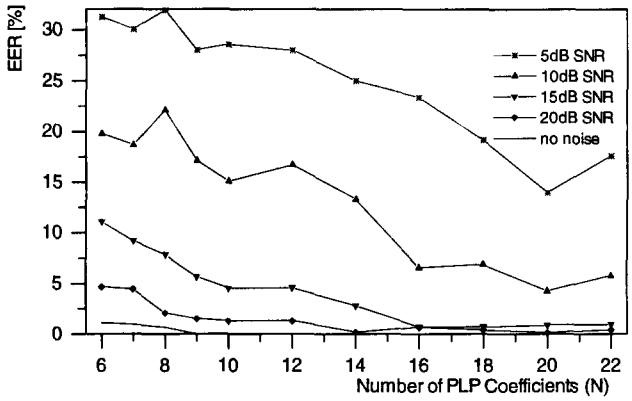


**Figure 4.** EER versus PLP model order N in the case of different SNR levels of the test data and undistorted training data (IFT corpus, HMM: 38 states, 2 mixtures)

This becomes obvious when considering the error rates (fig. 4) and their decreased factors for high PLP orders. For an additional use of delta cepstral coefficients we observed improvements for all SNR levels. Figure 5 shows the best equal error rates for a delta window length between 128 ms and 256 ms (see also table 2).

**Table 2.** EER for static and delta cepstral PLP coefficients (IFT corpus, PLP: N=20, HMM: 38 states, 2 mixtures)

| PLP analysis (SNR) | EER [%] (15 dB) | EER [%] (10 dB) | EER [%] (5 dB) |
|---|---|---|---|
| static coefficients only | 0.86 | 4.25 | 13.95 |
| static & delta coeff. (delta window size) | 0.31 (256 ms) | 2.14 (128 ms) | 10.99 (128 ms) |

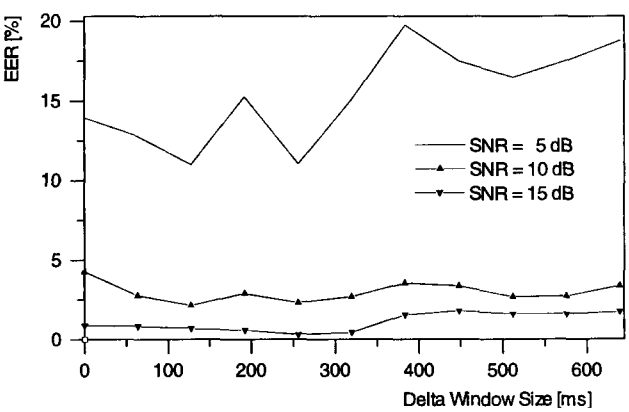


**Figure 5.** EER versus delta window length of delta cepstral coefficients for SNR = 5, 10 and 15 dB (IFT corpus, PLP: N=20, HMM: 38 states, 2 mixtures)

### 5.3 LOG-RASTA-Analysis

The LOG-RASTA-PLP can be used for the elimination of convolutional noise. We achieved good results for speech data which are only distorted by telephone channel noise. If the telephone speech data is additionally distorted by white noise, the LOG-RASTA-PLP analysis again leads to better results for all SNR levels compared to PLP (fig. 6).

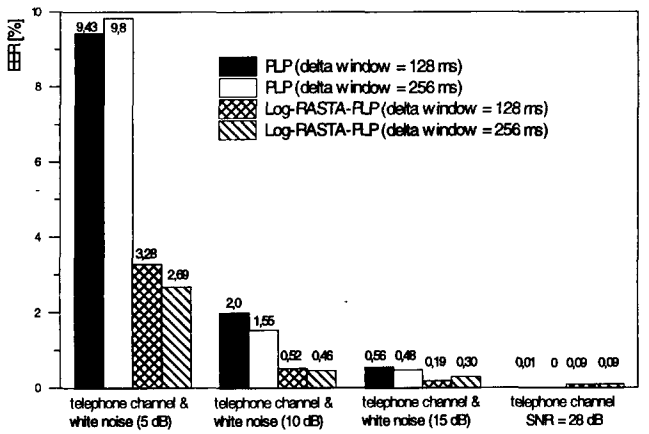


**Figure 6.** EER versus feature extraction, interval length of delta cepstral coefficients and type of the noise. (TUBTEL corpus, PLP: N=20, HMM: 56 states, 1 mixture)

### 5.4 J-RASTA-Analysis

The J-RASTA-PLP is useful for the elimination of both, additive and convolutional noise. For applying the J-RASTA-PLP, the estimation of the optimal J factor is necessary. The J factor depends on the noise power of the speech data. For its estimation, three different approaches are possible:

- without noise estimation (trial-and-error)
- off-line noise estimation (fixed J)
- adaptive noise estimation during the utterance

The J factor is defined by [7]:

$$J = \frac{1}{C \cdot E_{noise}} \qquad (2)$$

At first our system uses a fixed J factor which was estimated by an off-line noise measurement of the entire speech database. Figure 7 shows, that the results for the IFT corpus are nearly independent from the J factor of the test speech data but they strongly depend on the J factor of the training speech data. Figure 8 shows similiar dependence for the TUBTEL database. Extensive research is needed to determine the optimal J factor above all, because the estimation of the J factor is necessary to determine the SNR level. Hermansky et. al. found a J factor of C=3 in equation (2) to be optimal [7]. For our databases, the best results were reached with C=0.03, giving a J factor of $10^{-08}$. With this J factor, the equal error rate can be reduced by J-RASTA (table 4, SNR=5 dB) down to EER=1.41 % for the IFT corpus and down to EER=3.08 % for the TUBTEL corpus. Figure 7 and 8 show in principle the same results for both speech corpora, but depending on the SNR of the utterance, different J factors for training and test data are necesssary. In order to enable an efficient use of this J-RASTA-PLP, a specific procedure, which reduces the estimation for finding the right J factor, is needed.

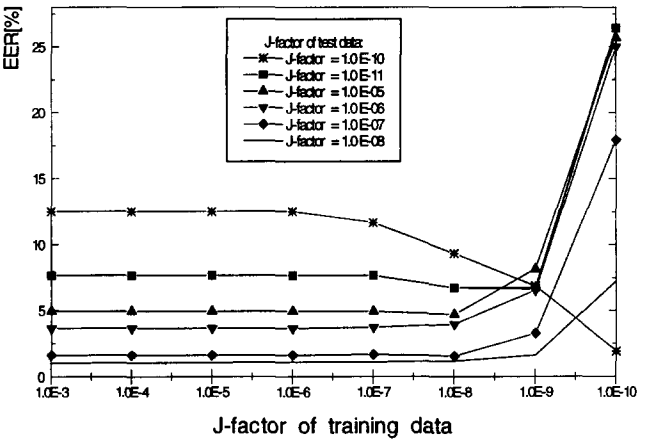


**Figure 7.** EER versus J factor of training and test data (IFT corpus, PLP: N=20, SNR=10 dB, HMM: 38 states, 2 mixtures)

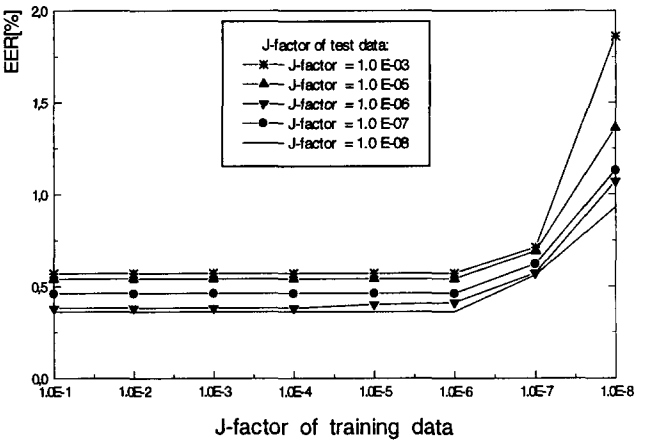


**Figure 8.** EER versus J factor of training and test data (TUBTEL corpus, PLP: N=20, SNR=10 dB, HMM: 56 states, 1 mix.)

### 5.5 Adaptive J-RASTA-Analysis (spectral mapping)

Further improvements of the results can be achieved by an adaptive determination of the J factor. Due to the dependence between the J factor and the SNR of the speech, the estimation of the noise level for adaptive settings of the J factor during the sentence is possible. The use of a time-varying J factor brings in a new problem that must be considered in training and verification, because changing a J factor over a time series introduces a new source of variability in the analysis. The J factor, as required by varying noise conditions, generates different dynamic of the spectra. The training system must be contend with a new source of variability. One approach to handle this variability, which has been successfully used in automatic speech recognition, is spectral mapping [7]. We therefore adapted

the method for our (SV) application. Finding a set of mapping coefficients, we applied the following J factors which gave the best results (chapter 5.4) for the J-RASTA-PLP (1.0E-02, 1.0E-03, 1.0E-04, 1.0E-05, 1.0E-06, 3.2E-02, 3.2E-03, 3.2E-04, 3.2E-05, 3.2E-06, 6.2E-02, 6.2E-03, 6.2E-04, 6.2E-05 and 6.2E-06). Figure 9 (see also table 3) shows the results for the adaptive RASTA-PLP with different J factor of training data and varying constant factor C (equation (2)). First experiments have shown significant improvements of result. For example, the EER for SNR=5 dB could be reduced from 9.8 % to 1.38 % compared to PLP and adaptive J-RASTA-PLP (table 4).

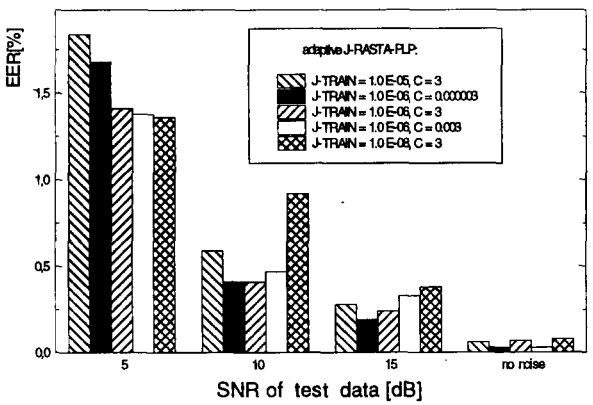


**Figure 9.** EER versus SNR of test data data for adaptive J-RASTA (TUBTEL corpus, adapt. RASTA-PLP: N=20, HMM: 56 states)

**Table 3.** EER for different constant factors C (TUBTEL corpus, adaptive RASTA-PLP: N=20, HMM: 56 states, 1 mixture)

| train J factor (SNR) | constant factor C (equation (2)) | EER [%] additive noise (5 dB) | EER [%] additive noise (10 dB) | EER [%] additive noise (15 dB) | EER [%] only telephone noise |
|---|---|---|---|---|---|
| 1.0 E-06 | 3 | 1.41 | 0.40 | 0.24 | 0.07 |
| 1.0 E-06 | 0.003 | 1.68 | 0.41 | 0.19 | 0.03 |
| 1.0 E-06 | 0.000003 | 1.38 | 0.47 | 0.18 | 0.03 |

**Table 4.** Best of equal error rates for different feature extraction (IFT corpus, PLP: N=20, HMM: 38 states, 2 mixtures) (TUBTEL corpus, MFCC/PLP: N=16/20, HMM: 56 states, 1 mix.)

| feature extraction (SNR) | speech database | EER [%] additive noise (5 dB) | EER [%] additive noise (10 dB) | EER [%] additive noise (15 dB) | EER [%] only telephone noise |
|---|---|---|---|---|---|
| MFCC | TUBTEL | 13.86 | 2.78 | 0.20 | 0.06 |
| PLP | TUBTEL | 9.80 | 1.55 | 0.48 | 0 |
| LOG-RASTA | TUBTEL | 2.69 | 1.55 | 0.30 | 0.09 |
| J-RASTA | TUBTEL | 1.41 | 0.39 | 0.19 | 0.04 |
| adaptive J-RASTA | TUBTEL | 1.38 | 0.40 | 0.18 | 0.03 |
| PLP | IFT | 10.80 | 2.10 | 0.62 | - |
| J-RASTA | IFT | 3.08 | 1.08 | 0.19 | - |

## 6. CONCLUSION

In this paper, we compared several procedures which make a speaker verification system more robust against noise. Our first experiments using spectral subtraction have shown that it can be used successfully if it is pre-connected to the speaker verification processing procedures. Otherwise, if spectral subtraction is integrated into the MFCC computation, the window length is too short and results in disturbing correlation effects. Therefore the advantage of low computation of the integrated spectral subtraction cannot be applied. Moreover, spectral subtraction (even if it is pre-connected) is restricted to stationary additive noise. But non stationary noise, which normally occurs in telephone channels, is much more difficult to handle.

If the telephone speech data are additionally distorted by white noise, the LOG-RASTA-PLP yields to better results for all SNR levels compared to PLP and MFCC analysis. Both, additive and convolutional noise can be better eliminated by the J-RASTA-PLP, but the off-line estimation of the right J factor is difficult.

Adaptive estimation of the J factor during the utterance improves the equal error rate in our speaker verification system. The procedure of spectral mapping is a well known effective methode to employ such systems. The adaptive RASTA-PLP is an useful procedure which makes the speaker verification system more robust against noise too. Additionally, we observed further small improvements by changing the constant factor C (equation (2)). In the future, we will concentrate on different versions of spectral mapping and some experiments with non stationary noise which occurs in a real telephone environment.

## 7. REFERENCES


[1] Fliegner, L.: „Textabhängige Sprecherverifizierung unter Berücksichtigung der Endpunktdetektion.", Dissertation, TU Berlin, 1995.

[2] Schürer, T.; Fellbaum, K., Hardt, D., et.al. : "TUBTEL - eine deutsche Telefonsprachdatenbank", Elektronische Sprachsignalverarbeitung, Wolfenbüttel, 1995.

[3] Hermansky, H: „Perceptual Linear Predictive (PLP) Analysis of Speech." JASSA, 4.1990

[4] Hermansky, H.; Morgan, N.; Bayya, A.; Kohn, P.: „RASTA-PLP Speech Analysis." TR-91-069, International Computer Science Institute, Berkley, CA 94704, 1991.

[5] Young, S.J., Woodland, P.: „The HTK Tied-State Continuous Speech Recognize", EUROSPEECH, pp. 2207-2210, Berlin, 1993

[6] Hardt, D.,: "Untersuchungen zum Einsatz der Störreduktion in der Sprecherverifizierung", Elektronische Sprachsignalverarbeitung, Berlin, 1996.

[7] Hermansky, H.; Morgan, N.; Bayya, A.; Kohn, P.: „Compensation for the effect of the communication channel in auditory-like analysis of speech (RASTA-PLP)", EUROSPEECH, Genova, Italy, pp. 1367, 1991

[8] Koehler, J.; Morgan, N.; Hermansky, H.; Hirsch, G.; Tong, G.: „Integrating RASTA-PLP into speech recognition.", ICASSP, Adelaide, pp. I- 421, 1994