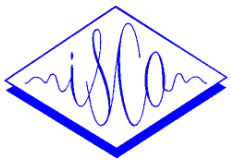

ESCA Tutorial and Research Workshop on
Speech Processing in Adverse Conditions
Cannes-Mandelieu, France
November 10-13, 1992

Robust Signal Preprocessing for Word Recognition in Noisy Environment

H. Eckhardt, M. Trompf, G. Angleys, H. Hackbarth

ALCATEL SEL, Dept. ZFZ/SC3, Lorenzstr. 10, 7000 Stuttgart 40, Germany
email: eckhardt@rcs.sel.de, trompf@rcs.sel.de

Reliable word recognition from degraded speech signals requires the incorporation of robust feature extraction as well as noise reduction into the system. This paper is focused on the comparison of different feature extraction methods with respect to different types of distortions. The experiments were made to tackle the problems of additive and convolutional noise as well as with the speaker-stress related Lombard effect, and are part of our ongoing work in the area of robust speech recognition.

1. Introduction

In most real-world applications, we have to deal with different types of possible distortions. Among them are additive noise caused by the superposition of time signal components on the microphone membrane as well as convolutional noise due to acoustic or electrical transfer functions, and speech variations (Lombard effect) because of speaker stress, e.g. while driving at high speed.

Different sources of distortion can be treated individually. For development of a robust system, the combination of different approaches is already known from publications by Kroschel (1988), Blanchet et al. (1992), and others. Most of them deal with later removal of these distortions. In contrast, our approach is based on the combination of robust feature extraction with regard to the different types of distortion and subsequent reduction of remaining noise signal portions.

In section 2, the system architecture and the database are described. Section 3 summarizes the experiments designed to minimize the contributions of the different sources of distortion to the overall decrease of accuracy. In section 4, a combination of robust feature extraction and different noise reduction methods is proposed. Finally, a summary of the results is given.

2. Robust Word Recognition System

Our multi-speaker database for speaker-dependent isolated word recognition contains two different data sets with 30 German words each. One of them is an offic-related vocabulary (OFFICE) and the other one is from a cellular phone application in a car environment (CAR). Both of them contain the ten digits. Five recordings of each data set were taken from ten speakers (5 male and 5 female; OFFICE) and six speakers (3 male and 3 female; CAR), respectively. In addition, OFFICE contains five noise-free "Lombard repetitions" per speaker produced under stress due to noise played back over earphones.

The OFFICE database was recorded in a noise-free environment in order to separate the problem of additive noise from speaker stress. In order to obtain a Lombard-free noisy speech signal, the noise recordings contained in a separate NOISE database were digitized and added to the speech signal samples in the time domain at different SNRs. This additional database contains recordings from a line printer and a computer room noise as well as computer generated white noise.

The CAR database was recorded in a real traffic environment in order to test the combination of different distortion sources. The utterances were spoken in driving situations at different speed.

All signals were lowpass filtered with a cutoff frequency of 3.4 kHz, and the feature vector coefficients were extracted every 20 ms from overlapping time segments. After feature extraction, the feature vector sequence is passed to the noise reduction module (optional), time normalized to 40 feature vectors for each word, and classified by a "scaly" neural network (Krause and Hackbarth, 1988) or a DTW- based speaker-dependent word recognizer.

3. Robust Feature Extraction

The different aspects of robust feature extraction are evaluated independently from each other (sections 3.1 to 3.3). Section 3.4 describes experimental results obtained from real-world recordings and hence a mixture of different types of distortions.

3.1 Additive Noise

Two different sets of feature vector coefficients were tested on their robustness in degraded speech: lpc-cepstral and plp (perceptually based lpc, Hermansky 1990) coefficients. Previous work of Hanson and Applebaum (1990) indicates that the derivatives of subsequent feature vectors are also necessary for robust recognition in degraded environment. The following simulations were designed to compare both types of feature vectors extracted from the OFFICE database with additive computer room noise. Each frame was represented by a vector of ten coefficients, and their first and second derivatives were calculated from three subsequent input frames. The word accuracy was evaluated with a DTW-based word recognizer.

From initial experiments, we found that a normalization scheme similar to the one used by Atal (1974) works best: each coefficient of a feature vector is normalized by first subtracting its mean value computed within the word boundaries of the current speech template and second by scaling these values with the standard deviation of long term speaker- independent statistics of speech.

The results for noise-free training and different test data SNRs are shown in figure 1. A comparison of lpc-cepstrum with plp results shows that lpc-cepstral coefficients work better for clean speech and high SNR (not depicted in figure 1), whereas plp coefficients are better for low SNR and additive noise. Furthermore, a combination of the original coefficients with their first derivatives leads to increased word accuracy mainly in highly distorted speech. No significant increase of performance could be obtained from the second derivative (not shown).

These simulations were repeated with the neural network based recognizer. The results were qualitively similar, although the recognition rates were somewhat worse.

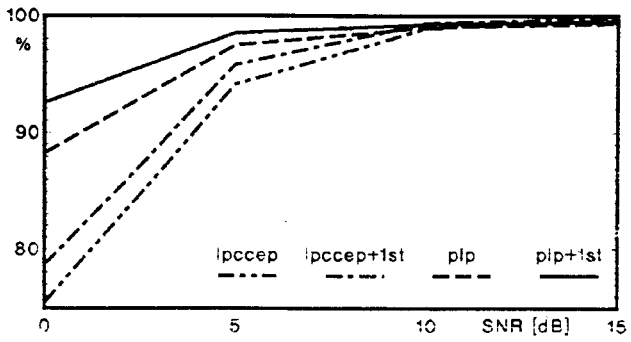

Figure 1: Comparison of word recognition results obtained from lpc-cepstrum and plp coefficients and their first derivatives on the OFFICE database with additive computer room noise using the DTW word recognizer.

3.2 Convolutional Noise

Previous work by Hermansky et al. (1991) suggests that RASTA-plp coefficients be used to eliminate the influence of transfer functions such as microphone or telephone channel characteristics. To verify this and to test the influence of additive noise on RASTA-plps, tests were made with noise-free speech as well as with additive noise using the neural network based word recognizer. A preemphasis of the digitized noisy speech signal (highpass characteristic) with

$$H(z) = 1 - 0.95\, z^{-1}$$

was used to simulate convolutional noise and test the behaviour of the RASTA filtering method. Training was always done with noise-free data and without preemphasis. The test results obtained from the original plp and RASTA-plp coefficients without derivatives are shown in figure 2.

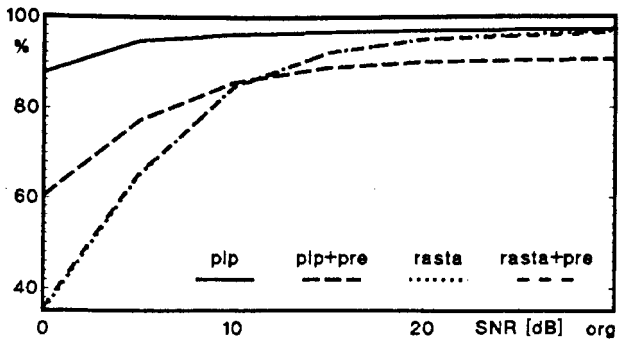

Figure 2: Comparison of plp and RASTA-plp coefficients for noisy and noise-free speech with and without speech signal preemphasis with noise added at different SNR levels. Word accuracy in %, neural network based recognizer.

Two conclusions can be drawn from the results: First, there is no difference between the original and the preemphasized RASTA-results, hence the RASTA-method is highly efficient for removal of convolutional noise. Second, the performance decrease for original plps due to signal preemphasis can only be prevented by RASTA-filtering for small additive noise components (>10 dB). However, if we compare the results for decreasing SNR, we see that the RASTA method is prohibitive in the presence of strong additive noise.

3.3 Lombard Speech

In order to separate the problem of additive noise from the problem of speaker stress, similar experiments were repeated with the noise-free Lombard data set for test. Training was done with the noise- and Lombard-free data set. The results shown in table 1 indicate that plp coefficients are highly sensitive against the Lombard effect, whereas lpccep coefficients are more robust in this case. The difference of the results for the original and the first derivatives (0.7% error for lpccep versus 3.7% for plp) shows that the choice of the coefficient set is crucial and that a significant contribution to the error rate is due to the variability of the speech signal. Again, the first derivative is necessary to improve the results, whereas the second is not useful. Obviously, different coefficient sets seem to be able to cope with different parts of the problem: plp are best for robustness against additive noise, lpc-cepstral coefficients give more reliable results in the presence of Lombard speech.

derivatives	-	1st	2nd
lpccep	98.9	99.3	99.2
plp	95.3	96.3	95.2

Table 1: Comparison of lpc-cepstrum and plp coefficients for noise-free Lombard speech. Recordings with noise over headphones. Word accuracy for the 30-word vocabulary in %.

3.4 Real World Environment

In order to verify the results obtained so far from feature extraction based experiments, a real-world task with a combination of both, additive noise and Lombard speech was investigated. This was done by using the CAR database and the DTW word recognizer. The results are shown in figure 3.

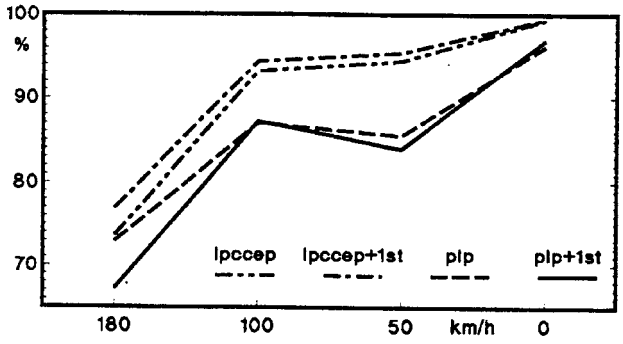

Figure 3: Word recognition results from real-world recordings in a car environment. Comparison of lpc-cepstral with plp coefficients in Lombard speech and additive noise using the DTW word recognizer.

The comparison of the two different feature extraction methods shows better performance for the lpc-cepstral coefficients. One reason for this might be due to the dominance of the Lombard related distortions over the additive noise. The word accuracy for plp coefficients combined with their derivatives is surprisingly low, with the dip at 50 km/h due to the driving situation.

4. Noise Reduction

4.1 Spectral Subtraction

We have inferred from section 3.4 that for high SNR levels the Lombard related distortions predominate the impact of additive noise. In an additional experiment, a combination of Lombard-insensitive lpc-cepstral coefficients and spectral subtraction for the removal of slowly varying parts of the noise spectrum was tested. In our implementation, we used speech pause information for the noise spectrum update as reported in Dvorak and Hörmann (1991).

Different driving situations at four different speeds between 0 and 180 km/h were tested. The recognition accuracy at low SNR levels remained almost the same as without spectral subtraction. Only for the 180 km/h recording an improvement of 4.5 % word accuracy compared to the result in figure 3 was achieved. From these experiments, we conclude that spectral subtraction is useful mainly for stationary driving situations at high speed and noise levels.

4.2 Neural Noise Reduction

Another adaptive method for additive noise reduction is neural noise reduction mapping. Our results reported previously in Trompf (1992) indicate that this method is efficient even for signals which vary in a moderate parameter range. By now, they have only been tested with additive noise, even though they can handle nonlinear mappings in general. Improvements on noisy OFFICE data were up to 40% word accuracy or around 96% recognition rate at a 6 dB SNR level with the neural network based word recognizer.

5. Summary

We have shown experimental results obtained from different feature vector sets to measure the sensitivity against additive and convolutional noise, Lombard-related distortions and combinations thereof. The most promising feature set is a combination of lpc-cepstral coefficients and their first derivatives together with an template-based normalization method based on the coefficients' statistics. Plp features proved to be quite robust in presence of additive noise. However, they show poor performance in Lombard speech. RASTA-plp coefficients turned out to be robust only against transfer functions. Spectral subtraction should only be applied to stationary noise signals at low SNR-levels. Further methods such as neural noise reduction mapping are currently tested in more detail with an application related data set.

6. Acknowledgements

The authors whish to thank Hynek Hermansky from US West for fruitful discussions and support with RASTA plp extraction software. This work was partially sponsored by the German Ministry of Development and Research under contract number 01 IV 101 I/2.

7. References

Atal B S (1974) Effectiveness of Linear Prediction Characteristics of the Speech Wave for Automatic Speaker Identification and Verification. JAcoustSocAm 55(6), June 1974, pp 1304-1312

Blanchet M, Boudy J, Lockwood Ph (1992) Environment Adaptation for Speech Recognition in Noise. EUSIPCO 1992, Vol 1, pp 391-394

Dvorak S, Hörmann T (1991) Noise-Robust Speech Recognition by Template Adaptation. DAGA 1991

Hanson B, Applebaum T (1990) Robust Speaker-Independent Word Recognition Using Static, Dynamic and Acceleration Features: Experiments with Lombard and Noisy Speech. IEEE ICASSP 1990, pp 857-860

Hermansky H (1990) Perceptual Linear Predictive (PLP) Analysis of Speech. JAcoustSocAm 87(4), April 1990, pp 1738-1752

Hermansky H, Morgan N, Bayya A, Kohn P (1991) Compensation for the Effect of the Communication Channel in Auditory-Like Analysis of Speech (RASTA-plp). Eurospeech 1991, pp 1367-1370

Krause A, Hackbarth H (1989) Scaly Artificial Networks for Speaker-Independent Recognition of Isolated Words. IEEE ICASSP 1989

Kroschel K (1988) Umgebungs-geräuschreduktion bei Sprachkommunikationssystemen. Frequenz, 42(1988), pp 79-84

Trompf M (1992) Building Blocks for a Neural Noise Reduction Network for Robust Speech Recognition. EUSIPCO 1992, Vol 1, pp 431-434