NON-LINEAR REGRESSION BASED FEATURE EXTRACTION FOR CONNECTED-WORD RECOGNITION IN NOISE

F. Seide and A. Mertins

Telecommunications Group, Technical University of Hamburg
Eissendorfer Str. 40, 21071 Hamburg, Germany, e-mail: Mertins@tu-harburg.d400.de

ABSTRACT

This paper shows the application of non-linear regression to robust feature extraction for noisy speech recognition. In this approach, a non-linear estimator is used to compute noise invariant features from non-linear combinations of noise contaminated observations. The observations may be short-term subband-energies obtained from a filter bank analysis, cepstral coefficients or linear prediction coefficients. Instead of training the hidden Markov models (HMMs) under various noise conditions, they can be trained with clean data. The results show that this method leads to error rates comparable to those achieved by training in the presence of noise.

1 INTRODUCTION

It is well known that the performance of speech recognizers trained under low noise conditions dramatically decreases in the presence of noise. Therefore, robustness of speech recognizers is decisive when they are used in an unknown (noise contaminated) environment. One possibility to obtain robustness is to train under various noise conditions [4]. Generally, this method leads to good results but a large computation effort is required.

However, algorithms allowing training using only one data set and leading to robust recognizers for any (unknown) noise level, are more desirable. Several methods, such as spectral subtraction, noise masking, or first-order equalization [1], have been presented in the literature. Using neural networks to map noise-contaminated into noise-independent observations has also been considered [2], and in [3], a cumulant-based linear predictor is used for the same task. While knowledge of the SNR was assumed in [2], our paper addresses the problem of robust detection, which means that the same non-linear procedure is applied to obtain good results in any unknown noise environment.

The new method presented here uses a non-linear estimator to compute noise-invariant features from non-linear combinations of noise-contaminated observations. The basic idea is to take advantage of the non-Gaussian distribution of speech observations and the (more or less)[1] Gaussian distribution of background noise. The ability of higher order statistics to separate non-Gaussian from Gaussian signals is well known [5].

In addition to freedom in the choice of the input features and the kind of non-linear combinations, there are several possibilities to define the desired noise-invariant observations. The following methods have been tested:

- estimation of noise-free observations
- estimation of the mean vectors of observations corresponding to the same state. Here, in addition to noise reduction, the estimator is used to minimize the intra-class distances.

[1] Additive Gaussian noise on the speech signal does not exactly result in Gaussian noise on short-term subband-energies.

0-7803-1775-0/94 $3.00

The first estimator is trained using general speech signals at certain noise levels. Here, the computation effort is negligible compared to training the HMMs.

For the second method, the alignment of training observation vectors to their corresponding HMM states has to be computed.

2 REGRESSION ALGORITHM

Let $\boldsymbol{r} = [r_1, \ldots, r_N]^T$ be a noise-contaminated observation and $\boldsymbol{y}$ the corresponding desired feature vector. An estimator for $\boldsymbol{y}$ based on non-linear regression can be written as

$$\hat{\boldsymbol{y}} = \boldsymbol{M}\boldsymbol{x}, \tag{1}$$

where $\boldsymbol{x}$ contains non-linear combinations of the input data $r_1, \ldots, r_N$. For a polynomial estimator, this is

$$\boldsymbol{x} = \left[1\,,\, r_1\,,\, \ldots\,,\, r_N\,,\, r_1^2\,,\, r_1 r_2\,,\, \ldots\,,\, r_N^2\,,\, \ldots\right]^T, \tag{2}$$

but any other non-linear combination is also possible.

Minimizing the estimation error $E\left\{\|\hat{\boldsymbol{y}} - \boldsymbol{y}\|^2\right\}$ leads to

$$\boldsymbol{M} = \boldsymbol{R}_{xy}\boldsymbol{R}_{xx}^{+}, \tag{3}$$

where $\boldsymbol{R}_{xx}$ and $\boldsymbol{R}_{xy}$ are given by

$$\boldsymbol{R}_{xx} = E\left\{\boldsymbol{x}\boldsymbol{x}^T\right\}, \qquad \boldsymbol{R}_{xy} = E\left\{\boldsymbol{y}\boldsymbol{x}^T\right\}, \tag{4}$$

and $^+$ denotes the pseudoinverse. Obviously, this estimator can be trained easily. From (2) - (4), it can be seen, that for a polynomial estimator of order p, the matrix $\boldsymbol{M}$ contains moments up to the order $2p$.

A refined estimator can be obtained by using the observations of the current frame as well as those of adjacent observations as estimator inputs. In this case, $\boldsymbol{x}$ has to be extended accordingly.

3 ILLUSTRATION

To illustrate the effect of non-linear regression, the mean square errors $E\left\{\|\boldsymbol{r} - \boldsymbol{y}\|^2\right\}$ (without regression) and $E\left\{\|\hat{\boldsymbol{y}} - \boldsymbol{y}\|^2\right\}$ (with regression) are considered.

Clean feature vectors were computed from real speech signals by composing twelve short-term log subband-energies obtained every 10 ms from a Mel-scale filter bank. As a 13th feature, the short-term log frame energy was used. To obtain contaminated vectors, white noise at SNRs of ∞, 30, 20, and 10 dB was added in time-domain.

A third order polynomial estimator without cross terms ($r_i r_j r_k$, $i \neq j \neq k$) was used to estimate the noise-free feature vectors from the contaminated observations.

By applying the non-linear regressor to each frame independently, the mean square error is reduced from the original value of $E\left\{\|\boldsymbol{r} - \boldsymbol{y}\|^2\right\} = 6.33$ to $E\left\{\|\hat{\boldsymbol{y}} - \boldsymbol{y}\|^2\right\} = 2.16$. If two adjacent frames are taken into account, the error further decreases to $E\left\{\|\hat{\boldsymbol{y}} - \boldsymbol{y}\|^2\right\} = 1.66$. With four adjacent frames, the error is reduced to 1.31.

Two examples are shown in figure 1.

4 RESULTS

In the experiments described below, a speaker dependent, connected-word HMM recognizer with single mixture, multivariate Gaussian emission densities was used. The vocabulary consisted of the ten German digits and an additional silence model. Speech data was recorded through a hand microphone at a sampling rate of 8 kHz.

The word models were trained independently with 110 clean observations per digit using the BAUM-WELCH method. For the test data, another 88 clean observations per digit were recorded. As in section 3, each observation was contaminated in time-domain with white noise at SNRs of ∞, 30,

20, and 10 dB, giving a total of 352 observations per digit.

In training, the short-term log subband-energies obtained from the filter bank and the log frame energy were used directly as features. First order temporal derivatives were appended, resulting in a 26 dimensional feature space.

The following regression-based features were tested:

- third order polynomial estimates of the *clean* feature vectors, based on log subband-energies and the log frame energy of the current frame,
- estimates based also on the previous and subsequent frames' subband and frame energies,
- estimates based on the current and four adjacent frames, and
- third order polynomial estimates of the HMM states' *mean* vectors using no, two, and four adjacent frames.

The first order temporal derivatives were not used as regressor inputs, but appended after the regression had been applied.

Figure 2 shows the error rates in terms of the noise level in the case of estimating clean feature vectors. At 10 dB, the error rate vastly decreases from 13.8% to 2.7% (using four adjacent frames), the overall error rate decreases from 4.2% to 1.1%.

Figure 3 shows the results, when HMM state means are estimated. Here, the error rate at 10 dB decreases to 3.1%, the total error rate to 1.4%.

5 CONCLUSION

In this paper, a new robust feature extraction method based on non-linear regression has been presented. It has been shown that robust recognizers can be obtained by training with clean data and using our preprocessor (estimator) in testing.

Further work will comprise the combination of non-linear estimators with highpass filtering of features. We expect that this combination will lead to noise- and channel-invariant recognizers.

REFERENCES

[1] Juang, B.H.; Paliwal, K.P.: *Hidden Markov Models with First-Order Equalization for Noisy Speech Recognition,* IEEE Trans. on Signal Processing, vol. 40, No. 9, pp. 2136-2143, 1992.

[2] Trompf, M.: *Building Blocks for a Neural Noise Reduction Network for Robust Speech Recognition,* Proc. EUSIPCO, pp. 431-434, 1992.

[3] Paliwal, K.K.; Sondhi, M.M.: *Recognition of Noisy Speech Unsing Cumulant-Based Linear Prediction Analysis,* Proc. ICASSP, pp. 429-432, 1991.

[4] Dautrich, B.A.; Rabiner, L.R.; Martin, T.B.: *On the effect of varying filter bank parameters on isolated word recognition,* IEEE Trans. ASSP, vol. 31, pp. 793-806, Aug. 1983.

[5] Rosenblatt, M.: *Stationary Sequences and Random Fields,* Boston, Birkhuser, 1985.

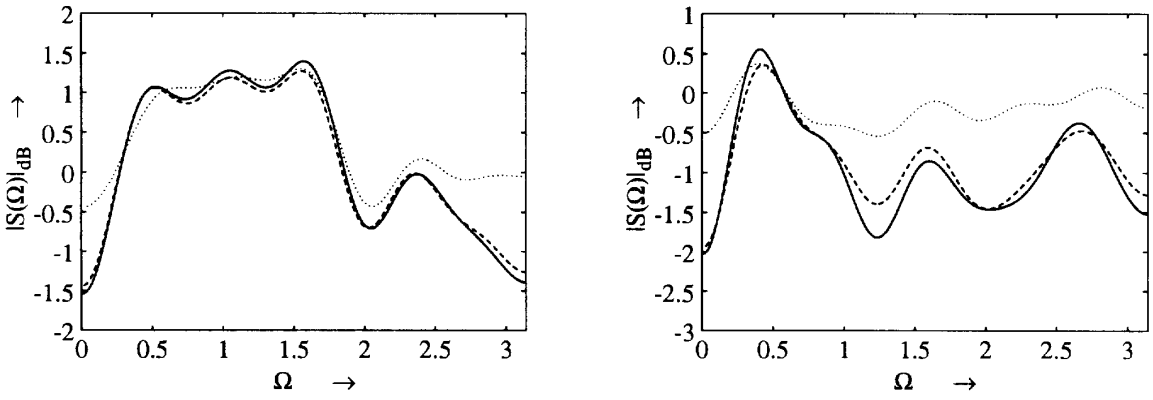

Figure 1: Examples of clean (——), noisy (. . .) and estimated (– – –) short-term log subband-energies.

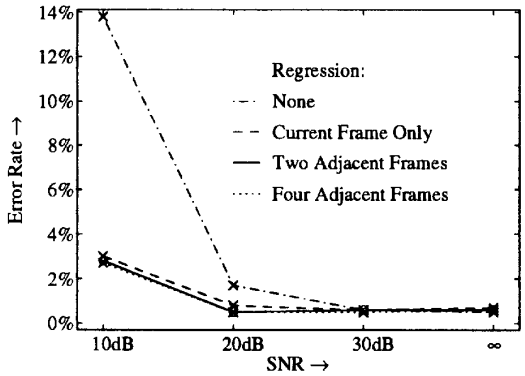

Figure 2: Error rates (estimating clean vectors)

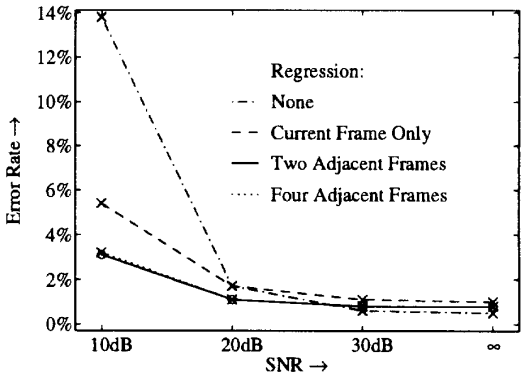

Figure 3: Error rates (estimating mean vectors)