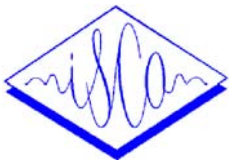

EUROSPEECH '97
5th European Conference on Speech Communication and Technology
Rhodes, Greece, September 22-25, 1997

# Noise Robust Recognition Using Feature Selective Modeling

Michael K. Brendborg and Børge Lindberg
Center for PersonKommunikation,
Aalborg University,
Fredrik Bajers Vej 7A2, DK-9220 Aalborg, Denmark
E-mail: mkb@cpk.auc.dk, bli@cpk.auc.dk



## ABSTRACT

In automatic speech recognition (ASR) systems immunity to additive noise may either be applied at the preprocessing stage or at the pattern matching stage.

The Feature Selective Modeling (FSM) approach suggested in this paper is applied in the pattern matching stage, but in contrast to most existing methods, it is optimized on a model basis such that noise robust and phonetically descriptive parameters of a particular model can be set in focus.

For sonorant sounds this is done by marking the lowest *n* mean values of each HMM density function as being sensitive to noise in a log filterbank representation. The noise robustness is obtained by de-emphasizing the marked feature dimensions. Two different methods for de-emphasizing - mean value masking and dimensional reduction - are presented and experimentally compared to the PMC-algorithm [2].

## 1. INTRODUCTION

Of the algorithms for noise robust recognition, some of the more promising approaches have been applied at the pattern matching stage by using models of the noise present in the environments of ASR systems. Examples are Noise Masking [3], Model Decomposition [6] and Parallel Model Combination (PMC) [2]. The main drawback of these algorithms is, however, the need of noise models, being problematic to pretrain or adapt to the noisy environments. Furthermore, it may be argued that there is no reason for modeling the noise as it contains no discriminative speech information.

To obtain further progress in noise robust ASR it is believed that noise immune ASR algorithms need to handle phonemes individually to fully utilize the natural noise robustness built into human speech communication.

The suggested approach of marking the lowest *n* mean values may obtain noise immunity to most speech sounds, but it is targeted towards sonorant sounds, i.e. sounds exhibiting a formant structure. However, FSM in general enables the possibility to apply different strategies according to the natural noise robustness of the sounds to be modeled.

To illustrate how sonorant sounds are distorted by additive noise a one state and one mixture HMM has been trained on the Danish unrounded, front, high vowel /i/ distorted by noise at different SNRs. The noise originates from the RSG-10 database [5] and consists of car noise recorded at the constant speed of 120 km/h and operation room noise recorded in a destroyer. The relatively non-stationary operation room noise has a broadband spectrum whereas the relatively stationary car noise has its main energy present below 1 kHz. Log energy mel scale filterbank coefficients (FBANKs) are used for representing speech since this parameter type is appropriate for analyzing triphone HMMs with respect to formants and it relates well to the human auditory system.

The 18 FBANK mean values of the HMM covering the frequency range from 200 Hz to 4 kHz are shown in figure 1. It is observed how the FBANK mean values

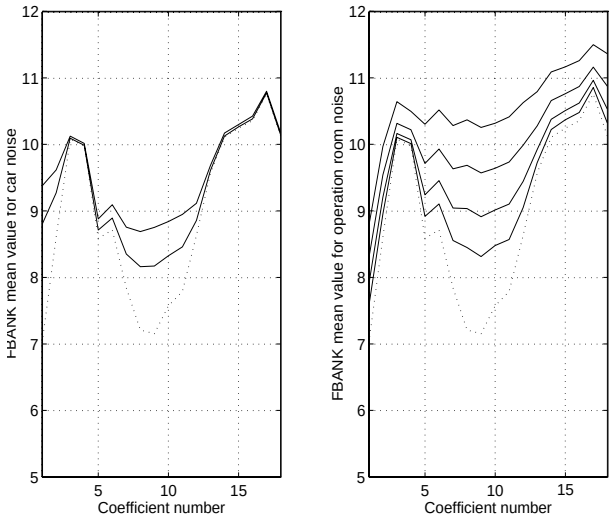


Fig. 1. FBANKs of the Danish vowel / i / distorted by car noise and operation room noise. The dotted line is clean speech. The solid lines are 18 dB, 12 dB, 6 dB, and 0 dB SNR where 6 dB and 0 dB are only illustrated for the operation room noise.

around the formants are less distorted by the two different noise types than in the regions between the formants. This is explained by the log function approximation: $\log(x+y) \approx \log(max(x,y))$.

Thus, for noisy speech representing sonorant sounds the FBANKs are expected to be dominated by the speech signal in the regions around the formants and by the noise signal in the regions between the formants for many noise types at signal-to-noise ratios (SNRs) above 0 dB.

## 2. FEATURE SELECTIVE MODELING

The aim of the FSM approach is to avoid the pre-training (or adaptation) of noise models by excluding or de-weighting the features not being found noise robust and discriminative for the particular phoneme.

In the FSM approach HMMs are trained on clean speech. The HMMs are then analyzed and the lowest *n* mean values of each density function are marked as being sensitive to noise. That is, each density function will have different mean values marked according to the data it models. The strength of FSM is that it is not based on a global optimization or selection of a set of parameters, but it is optimized on a model basis where the noise robust and phonetically descriptive parameters of a particular HMM can be set in focus of the modeling.

Although the FSM approach appears to be fundamentally different in its way to obtain noise immunity, it is shown in [1] that it can be considered as a non-noise specific version of the noise model based approaches presented in [2], [3] and [6].

Two methods are suggested for de-emphasizing the marked dimensions: 1) mean value masking and 2) dimensional reduction. For reasons of simplicity diagonal covariance matrices are assumed.

### 2.1 Mean Value Masking (MVM)

This method is related to noise masking [3] as the marked dimensions of the density functions are masked by the respective mean values. The probability $P(\boldsymbol{o})$ of an observation vector $\boldsymbol{o}$ with dimension D is then calculated as illustrated in equation 1.

$$\begin{aligned} P(\boldsymbol{o}) &= \prod_{d=1}^{D} P_d(o_d) \\ P_d(o_d) &= \begin{cases} N(\mu_d, \mu_d, \sigma_d), \text{ if marked} \\ N(o_d, \mu_d, \sigma_d), \text{ else} \end{cases} \end{aligned} \quad (1)$$

It is ensured that the noise sensitive parts of the density functions do not result in a low probability score, which then can only be obtained if a mismatch occurs around the formants (noise robust region).

When a dimension is masked it will emit an optimal probability score for the dimension in question. This is a general disadvantage of the masking approach because the total weight of the unmasked dimensions become more and more reduced as the number of masked dimensions increases. Information left to model the different sounds may thus disappear in all the probability scores from the masked dimensions. The second method presented aims at alleviating this problem.

### 2.2 Dimensional Reduction (DR)

The aim of the second method is to remove the probability contribution from the marked dimensions as illustrated in equation 2.

$$P(\boldsymbol{o}) = \prod_{d \in \{\text{un-marked}\}} \frac{1}{\sqrt{2\pi}\sigma_d} \exp\left( \frac{(o_d - \mu_d)^2}{2\sigma_d^2} \right) \quad (2)$$

The integral of the dimensional reduced multi-dimensional density function will still be equal to one. This will ensure that the dynamic range of the probability score is preserved even in the case of many marked mean values. As with the first approach it is ensured that low scores can only be obtained if mismatches occur around the formants (noise robust region).

### 2.3 Adding Delta Coefficients

Usually, in ASR-applications, time-derivative information (delta coefficients) is used in the speech parameterisation and estimated using a standard regression formula. However, delta coefficients based on FBANK are highly sensitive to stationary additive noise.

It was found in [1] that it is more sensible to calculate delta coefficients on the basis of linear energy filterbank output since stationary additive noise is cancelled out in the linear regression.

In order to focus on the high valued coefficients the absolute values of the linear energy based delta coefficients are subsequently transformed by using the log operation as given in equation 3

$$d_t^{\text{Log Linear}} = \log\left( \left| \frac{\sum_{\tau=1}^{N} \tau (E_{t+\tau} - E_{t-\tau})}{2\sum_{\tau=1}^{N} \tau^2} \right| \right) \quad (3)$$

where $E_t$ is the energy in a time-frame, *t*, and $N$ is the number of neighboring static features used for estimating the derivative. In the following, the coefficients obtained are denoted LL delta coefficients.

## 3. EXPERIMENTS

A number of experiments are conducted both using the MVM- and the DR-method. These are evaluated using static features only. Further the performance of the DR method is analysed when using combined masking of both static and LL delta coefficients. Finally, the latter results are compared to the results obtained using a standard PMC-algorithm [2] based on static features only.

The speech material applied in the experiments consists of the T0 and the U0 minimal pair lists from the Danish part of the EUROM.1 database [4]. The transcription in SAMPA notation is / t_d@ / (t_de in orthography) where the '_' denotes one of the following 11 Danish vowels / i, e, E, a, A, y, 2, u, o, O, Q /. The noise data is the car and the operation room noise presented previously. Noisy speech is generated artificially in a manner

similar to the one used for generating the NOISEX-92 database [7]. The speech signal is divided into frames using a 20 msec Hamming window. A frame overlap of 50% is used. The signal is band limited from 200 Hz to 4KHz. 18 FBANKs are calculated from each frame and they are extended with delta coefficients and delta energy.

Triphone HMMs are used for modeling the speech in an isolated word recognition mode. Speakers in the training part do not appear in the test part.

## 4. RESULTS

The results obtained by using the MVM-method are given in figure 2 and 3 for the car noise and operation room noise, respectively. The results of the DR-method are given in figure 4 and 5. The curves denoted *0 coeff. marked* correspond to recognition without using FSM.

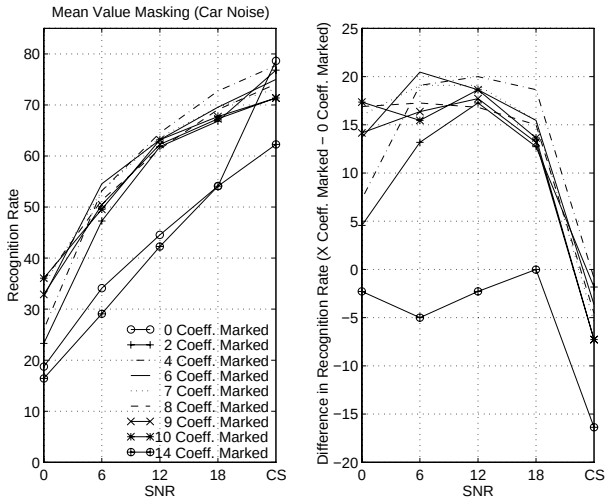


Fig. 2. The Mean Value Masking method for the car noise.

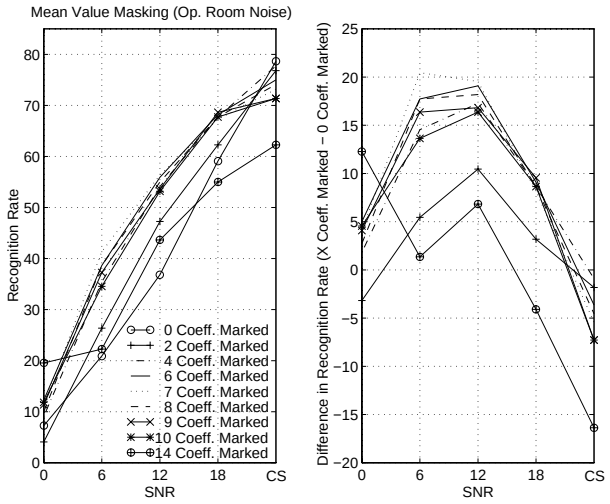


Fig. 3. The Mean Value Masking method for the operation room noise.

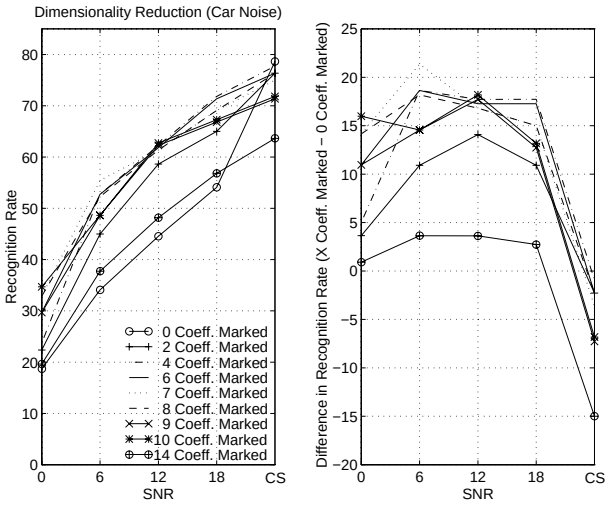


Fig. 4. The Dimensional Reduction method for the car noise.

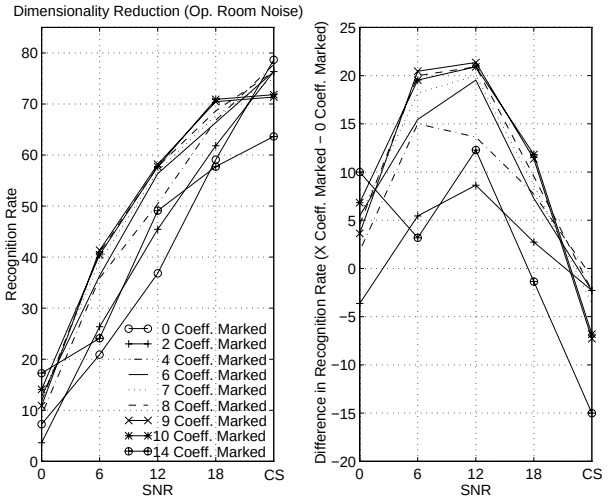


Fig. 5. The Dimensional Reduction method for the operation room noise.

The results in figure 2, 3, 4 and 5 clearly show an increase in the noise robustness using either the MVM or the DR-method. The maximum recognition score is dependent on the noise type, the SNR and the number of marked dimensions. It is, however, not difficult to select a number of coefficients to be marked, which gives a high degree of robustness for both noise types, since the performance seems to be relatively insensitive to this number.

The best number of marked coefficients seems to be higher for the DR method. Furthermore, the DR method performs better that the MVM method for 14 marked coefficients. This is probably because the DR method does not have the problem of decreasing the dynamic range of the probability score when many dimensions are marked.

The DR method is therefore preferred over the MVM-method and is used in all subsequent FSM experiments.

Figure 6 and 7 show the results of applying combined marking of both static and LL delta coefficients. Results are obtained for both car and operation room noise and are also obtained for the standard PMC-algorithm based on static features only. It is observed that the suggested

approach for calculating LL delta coefficients further improves the noise robustness.

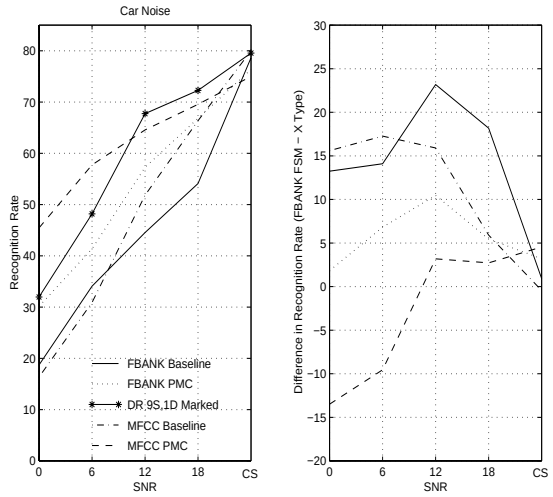


Fig. 6. Results of DR-method masking nine static and one LL delta coefficient. Results are also presented for FBANK and MFCC (Baseline and PMC). All results are for car noise.

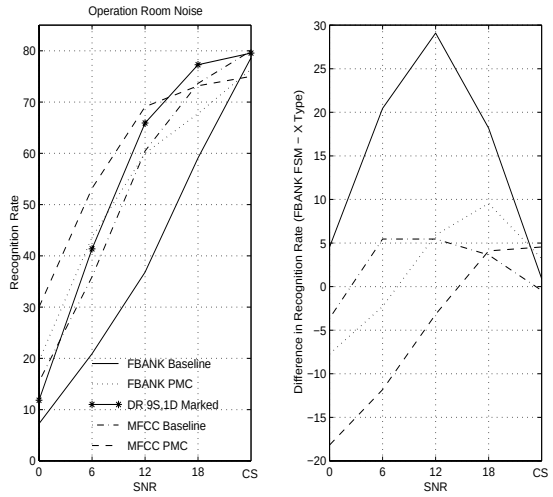


Fig. 7. Results of DR-method masking nine static and one LL delta coefficient. Results are also presented for FBANK and MFCC (Baseline and PMC). All results are for operation room noise.

## CONCLUSION

The methods proposed in this study are expected to perform noise robust modelling of all phonemes, but they are optimized to perform robust modelling of sonorant sounds and in particular vowels only. However, the general idea behind the FSM approach is to develop and use different methods for performing noise immune modelling of individual phoneme groups.

Two methods are suggested for de-emphasizing the marked dimensions of a mean vector within a density function of each individual HMM. These are mean value masking, MVM, and dimensional reduction, DR.

The results show a significant improvement in the recognition performance by focusing the modeling at the phonetically descriptive parameters even though no noise specific information is used. The FSM algorithm seems to be relatively insensitive to the number of marked coefficients within the range from 4 to 10.

The FSM approach obtains results which are competitive to the PMC results at the higher SNRs for vowels without using noise specific information.

To apply the FSM algorithm in an optimal way to vocabularies which contain a broader phonemic content it is important to develop a similar strategy for at least the obstruents.

## REFERENCES


[1] M. Brendborg. Towards Noise Immune Automatic Speech Recognition using Phoneme Models. PhD-Thesis, Center for PersonKommunikation, Aalborg University, Denmark, October 1996.

[2] M. Gales and S. Young. Cepstral parameter compensation for HMM recognition in noise. *Speech Communication*, 12(3):231–239, July 1993.

[3] D. Klatt. A digital filter bank for spectral matching. In *Proc. IEEE International Conference on Acoustics, Speech, Signal Processing*, pages 573–576, 1976.

[4] B. Lindberg and H. Christensen. Documentation of the danish EUROM.1 database. ESPRIT project 2589 SAM. Technical report, Institute of Electronic Systems, Aalborg University, Denmark, 1995.

[5] H. Steeneken and F. Geurtsen. Description of the RSG-10 noise database. Technical report, Institute for Perception TNO, Soesterberg, Kampweg 5,The Netherlands, 1988.

[6] A. Varga and R. Moore. Hidden markov model decomposition of speech and noise. In *Proc. IEEE International Conference on Acoustics, Speech, Signal Processing*, pages 845–848, 1990.

[7] A. Varga, H. Steeneken, M. Tomlinson, and D. Jones. The NOISEX-92 study on the effect of additive noise on automatic speech recognition. Technical report, Speech Research Unit, IS2, Defence Research Agency, Electronics Division and TNO Institute for Perception, RSRE, St. Andrews, Great Malvern, England, TNO Institute for perception, P.O. Box 23, 3769 Zg Soesterberg, The Netherlands, 1992.

[8] S. Young, P. Woodland, and W. Byrne. *HTK - Hidden Markov Model Toolkit v. 1.5*. Entropic Research laboratory, Inc., December 1993.