Hybrid model decomposition of speech and noise in a radial basis function neural model framework

Sørensen, Helge Bjarup Dissing; Hartmann, Uwe

Published in:
Proceedings og the IEEE International Conference on Acoustics, Speech, and Signal Processing

Link to article, DOI:
10.1109/ICASSP.1994.389570

Publication date:
1994

Document Version
Publisher's PDF, also known as Version of record

HYBRID MODEL DECOMPOSITION OF SPEECH AND NOISE IN A RADIAL BASIS FUNCTION NEURAL MODEL FRAMEWORK

Helge B.D. Sorensen ¤ and Uwe Hartmann #

¤ Department of Electronics and Electrical Engineering
The Engineering Academy of Denmark, DK-2800 Lyngby, DENMARK

Speech Technology Centre, Institute of Electronic Systems
Aalborg University, DK-9220 Aalborg, DENMARK

ABSTRACT

This paper focus on a new approach to automatic speech recognition in noisy environments where the noise has either stationary or non-stationary statistical characteristics. The aim is to perform automatic recognition of speech in the precence of additive car noise. The technique applied is based on a combination of the Hidden Markov Model (HMM) decomposition method [1], for speech recognition in noise, developed by Varga and Moore from DRA and the hybrid (HMM/RBF) recognizer [2], containing Hidden Markov Models and Radial Basis Function (RBF) Neural Networks, developed by Singer and Lippmann from MIT Lincoln Lab. We modified the hybrid recognizer to fit into the decomposition method to achieve high performance speech recognition in noisy environments. Our approach has been denoted the Hybrid Model Decomposition method and it provides an optimal method for decomposition of speech and noise by using a set of speech pattern models and a noise model(s), each realized as an HMM/RBF pattern model.

1. INTRODUCTION

One of the most promising methods for dealing with noise contamination in the recognition phase is the HMM decomposition method [1] developed by Varga and Moore from DRA. The components of the decomposition are speech HMM models of for example whole words and a second concurrent set of noise HMM models. The input to the HMM decomposition is a noisy speech pattern represented by a sequence of observation vectors each containing e.g. the log energy level outputs from a filter bank. The observation vector probabilities are calculated on the basis of the output from a speech HMM model combined with the output from a noise HMM model. Recognition of the speech pattern is performed by the speech model and the noise is recognized by the noise model simultaneously. HMMs can model dynamically varying signals making it possible to handle noises that have statistical characteristics ranging from stationary to highly time varying noise e.g. background talkers or machine noise. Another advantage of the HMM decomposition over previous approaches is that it provides an optimal method for recognizing speech and noise simultaneously.

The above mentioned decomposition method applies HMMs to model speech and noise patterns. Alternatives to HMMs are hybrid recognizers e.g. a combination of HMMs and Neural models. The main objective by this combination is to utilize the temporal modeling capability of HMMs and the discriminative power of Neural models. Recently several hybrid models have appeared for example "Context-Dependent Neural Networks", CDNN, by Bourlard, Morgan, Wooters and Renals [5], "Linked Predictive Neural Networks", LPNN, by Tebelskis and Waibel [6], "Self-structuring Hidden Control neural models", SHC models, by Sorensen and Hartmann [7], and "HMM/RBF recognizers", by Singer and Lippmann [2]. These hybrid recognizers typically have better or equal performance compared to pure HMMs. We have selected the latter hybrid models as a starting point for the design of the Hybrid model decomposition method, because these models have few model parameters, discriminative training, fast training (one pass matrix inversion) and as presented in our paper high performance in noise compared to pure HMMs.

An HMM/RBF pattern recognizer consist of one or more RBF neural models combined with an HMM e.g. a left-to-right state model. The initial and transition probabilities in the state model are estimated using standard HMM reestimation techniques. The observation probabilities

0-7803-1775-0/94 $3.00

for the state model are estimated by the RBF models. The motivation for using RBFs is the desire to apply the discriminative properties of neural networks [2]. Discriminative training of an RBF model tries to produce estimates of all Bayes probabilities simultaneously.

The Hybrid decomposition method based on a combination of the Varga-Moore decomposition method and the Singer-Lippmann recognizer is presented in section 2 followed by a decription of the results in section 3.

2. HYBRID MODEL DECOMPOSITION USING RADIAL BASIS FUNCTION NEURAL MODELS

The application in this paper is recogntion of digits contaminated by additive car noise. For each digit a left to right state model is trained and for the noise one ergodic state model is trained. All the models are initially HMM models and after the HMM training RBF models are trained and added to the HMM models. In the following three subsections, preprocessing, training and application of the Hybrid model decomposition system are described.

2.1 Preprocessing

The input to the Hybrid model decomposition system is a sequence of observation vectors representing a noise contaminated speech pattern. Each element in an observation vector is calculated as the log energy level of the corresponding channel from a filter bank front end similar to the filter bank in [1]. One observation vector at time t is denoted as:

$$o_t^{SN}=\log(o_t^S+o_t^N) \qquad (1)$$

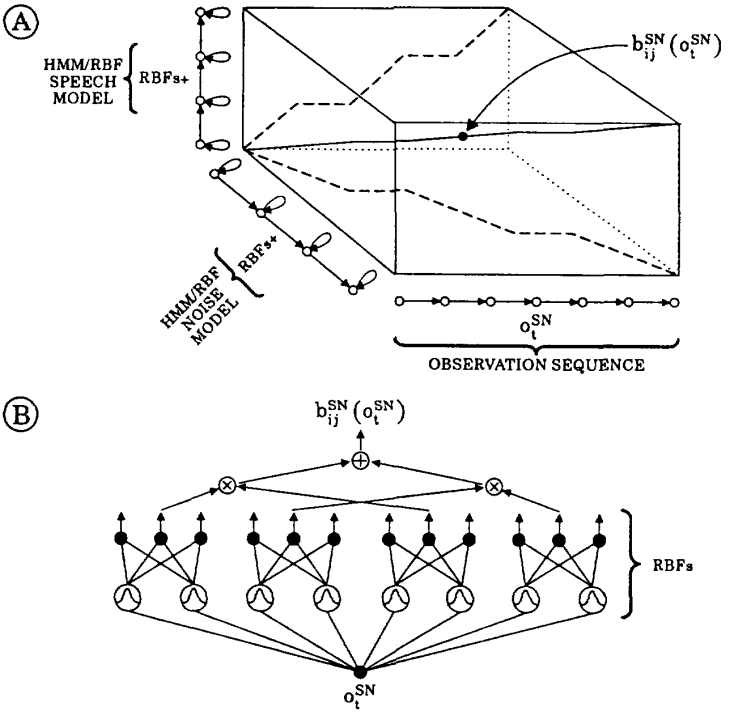

Figure 1 (A) Hybrid Model Decomposition system based on a combination of an HMM/RBF speech pattern (e.g. a digit) model and an HMM/RBF noise model. Initial, transition and output probabilities $b_{ij}^{SN}(o_t^{SN})$ (only the latter is indicated on the figure) are the foundation for a 3-dimensional Viterbi search for the optimal state sequence in the state-space, thus carrying out recognition of simultaneous speech and noise signal components. (B) Four RBF neural models estimate the output probabilities $b_{ij}^{SN}(o_t^{SN})$, see equation (3). For simplicity the number of inputs, basis functions and connections in the RBFs, states etc. are minimized in (A) and (B).

where the sum indicate that the observed log energy level is the log of a sum of the speech energy level o_t^S and the noise energy level o_t^N. An observation vector is calculated every 32 msec and the frame overlap is 16 msec.

2.1 Training of speech and noise models

For every digit in the vocabulary a Continuous HMM (CHMM) based on a left-to-right state model is trained using standard HMM reestimation techniques. The CHMMs use continuous density, unimodal diagonal-covariance Gaussian classifiers for each digit state. A CHMM based on an ergodic state model is trained on the available noise signal. The next phase of the training is to expand these speech and noise HMMs into HMM/RBF models as indicated on Fig. 1 by training a set of RBF models. The initial and transistion probabilities from the HMM models are reused in the HMM/RBF models. The purpose of the RBF models is to estimate the output vector probabilities which are based on the following expression:

$$b_{ij}^{SN}(o_t^{SN}) = \int_C b_{ij}(o_t^S, o_t^N)\, d(o_t^S)\, d(o_t^N) \quad (2)$$

$$C = \{\, o_t^S, o_t^N \mid o_t^{SN} = \log(o_t^S + o_t^N) \,\}$$

From equation (2) it is possible to derive the following equation if a few assumptions are made e.g. that the speech and noise signals are independent. The output probabilities can be approximated by:

$$b_{ij}^{SN}(o_t^{SN}) \approx b_i^S(o_t^{SN}) \cdot F_j^N(o_t^{SN}) + b_j^N(o_t^{SN}) \cdot F_i^S(o_t^{SN}) \quad (3)$$

Each term on the right side contains a probability density function and a probability distribution function and each of the four factors is implemented using an RBF model. Thus four RBF models are necessary, see Fig. 1. To calculate the distribution functions integrations of basis functions (Gaussian functions) are necessary and these are performed by applying Taylor series expansions. It is easy to show that each term in equation (3) can be realize using a weighted sum of basis functions, no matter if the term contains a density function or a distribution function. The only difference turns out to be the type of basis functions.

It is important to mention that each of the RBF models either is connected to all speech models or to all noise models. The training of an RBF model is thus discriminative because the model tries to produce all probabilities for all speech (or noise) pattern models simultaneously.

The training of the RBF models is described in details in [2]. The most important feature of the training is that the weights in the neural models can be trained using a one-pass matrix inversion technique in stead of the popular but very slow gradient descent techniques.

2.2 Application of the Hybrid decomposition method

After training the HMM/RBF speech models and the HMM/RBF noise model the next step is to apply these in the Hybrid model decomposition system as indicated on Fig. 1. The initial and transition probabilities for the speech and noise state models are available from the HMM training and the output probability $b_{ij}^{SN}(o_t^{SN})$ for state i in the speech model and for state j in the noise model is calculated from equation (3) which is based on the four RBF models illustrated on Fig. 1.

The above mentinoned probabilities are the foundation for a 3-dimensional Viterbi search for the optimal state sequence in the state-space shown in Fig. 1, thus carrying out recognition of simultaneous speech and noise signal components. For each combination of a speech model and a noise model the likelihood for the optimal state sequence is calculated. The combination having maximum likelihood indicate the recognition of the speech pattern (digit) hidden in the noisy observation vector sequence.

3. EXPERIMENTS

The HMM/RBF models and the Hybrid model decomposition system have been tested under noise-free and noisy conditions applying the TIDIGITS [3] and NATO RSG-10 [4] databases. Speech signals and car noise signals were added together. The speech signals were sampled at 8 kHz and limited to telephone bandwidth. The first set of experiments were performed using the HMM/RBF models only. The performance of these models was compared to the performance of the HMMs that were used to initialize the HMM/RBF models, see Table 1. The next experiment was performed under

noisy conditions applying the Hybrid model decomposition system at - 5 dB, see Table 1. Future experiments should naturally include tests at other SNRs.

HMM/RBF models are more noise robust than the applied HMMs probably due to the discriminative training of the RBF models and the parallel processing in the models. The preliminary tests in Table 1 indicate that the Hybrid model decomposition system is a relevant noise reduction system.

SNR	Infinity	15	0	-5
HMM	94.3	87.7	82.5	68.0
HMM/RBF MODELS	97.5	91.6	87.5	78.8
HYBRID DECOMP.	*	*	*	83.0

Table 1 The performance of three types of recognizers.

4. CONCLUSION

Hybrid decomposition for robust speech recognition has been proposed, defined and tested. The efficient Singer-Lippmann (HMM/RBF) recognizer has been combined with the Varga-Moore decomposition method with the purpose of achieving high performance recognition of noise contaminated speech. A theoretical analysis of RBF models has given the foundation for a definition of four RBF models for the estimation of the output probabilities necessary in the Hybrid decomposition system. The recognition results presented are promising and these can be improved e.g. by using other features, increasing the number of basis functions in the RBF models and by fine-tuning of the RBF parameters. Finally we showed that HMM/RBF recognizers can be more noise robust than HMMs. Hybrid recognizers can outperform HMM recognizers in terms of fewer parameters, training time and performance in noise as indicated in [2] and in this paper justifying the merging of a hybrid recognizer and the decomposition method. We conclude that neural models can improve advanced noise reduction systems.

ACKNOWLEDGEMENTS

Thanks to Professor Paul Dalsgaard, Center for PersonKommunikation, Institute of Electronic Systems, Aalborg University for support. The support and excellent noise reduction work by M.Sc.E.E.s Claus D. Jensen, Thomas H. Sørensen and Lars L. Andresen are greatly appreciated.

REFERENCES

[1] A. Varga, R.K. Moore, "Hidden Markov Decomposition of Speech and Noise", in IEEE Proceedings ICASSP90, Albuequerque, USA, 1990.

[2] E. Singer, R.P. Lippmann, "A Speech Recognizer Using Radial Basis Function Neural Networks in an HMM Framework", in IEEE Proceedings ICASSP92, San Francisco, USA, March 1992.

[3] Texas Instruments and National Institute of Standards and Technology, "Studio Quality Speaker-Independent Connected-Digit Corpus (TI-DIGITS)", NIST Speech Discs 4-1, 4-2 and 4-3, February 1991.

[4] NOISEX_0 AND NOISEX_1, NATO: AC243/(Panel 3)/RSG.10, ESPRIT: Project No. 2589 - SAM, SRU, Defence Research Agency, Malvern, Great Britain, June 1992.

[5] H. Bourlard, N. Morgan, C. Wooters and S. Renals, "CDNN: A Context-Dependent Neural Network for Continuous Speech Recognition", in IEEE Proceedings ICASSP92, San Francisco, USA, March 1992.

[6] J. Tebelskis, A. Waibel, B. Petek and O. Schmidbauer, "Continuous Speech Recognition Using Linked Predictive Neural Networks", in IEEE Proc. ICASSP91, Toronto, Canada, 1991.

[7] H.B.D. Sorensen, U. Hartmann, "Self-structuring Hidden Control Neural Models for Speech Recognition", in IEEE Proceedings ICASSP92, San Francisco, USA, March 1992.