論文 / 著書情報
Article / Book Information

論題(和文)	
Title(English)	Noise discrimination using models with different structures
著者(和文)	篠田 浩一
Authors(English)	Agnieszka Betkowska, Koichi Shinoda, Sadaoki Furui
出典(和文)	日本音響学会 2005年春季講演論文集, Vol. , No. 2-Q-7, pp. 111-112
Citation(English)	, Vol. , No. 2-Q-7, pp. 111-112
発行日 / Pub. date	2005, 3

2—Q—7 Noise Discrimination Using Models with Different Structures*

©Agnieszka Betkowska, Koichi Shinoda, and Sadaoki Furui (Tokyo Institute of Technology)

1 Introduction

The goal of automatic noise recognition (ANR) is the classification of audio noises in acoustic environments. The ability of segregating noise is important for personal robots. This fact motivated us to investigate the noise discrimination problem in house environments. In our previous work [1], for each noise category, a different Hidden Markov Model (HMM) was defined. The choice of its topology depended on the amount of data available and the characteristics of the acoustic signals to be modeled. We proposed a method of adjusting the complexity of the HMM according to the training data available. The number of mixtures of each state was reduced by using the Maximum Description Length (MDL) criterion.

In this paper, we investigate noise discrimination methods based on the combination of the recognition results from more than one systems. The model topologies are different among the systems.

2 Method

Let N be the the number of recognition systems and K be the number of noise classes. The *a posteriori* probability (score) that the noise sample x belongs to class i in system j is expressed as follows:

$$a_{ij}(x) = \frac{p_{ij}(x)}{\sum_{k=1}^{K} p_{kj}(x)}, \quad i = 1,\ldots,K, \quad j = 1,\ldots,N,$$

where p_{ij} is the probability of noise i given by system j. In order to combine scores of N systems, the Mixtures of Experts (MoE) methodology is utilized. Instead of one global network, several expert networks are created. A gating network that decides which of the expert network should be used is also built [2] (Figure 1(a)).

A neural network for each expert consists of N input neurons, one hidden layer with M neurons and one output neuron (Figure 1(b)). The likelihood L_i for each class i is calculated as follows:

$$q_{im}(x) = g(\sum_{j=1}^{N} \omega_{mj} a_{ij}(x)), \quad i = 1,\ldots,K, \quad m = 1,\ldots,M,$$

$$L_i(x) = g(\sum_{m=1}^{M} v_m q_{im}(x)), \quad i = 1,\ldots,K,$$

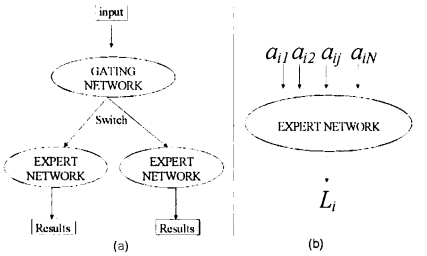

Figure 1. (a) An architecture of Mixtures Of Experts (MoE) (b) An expert network for noise i in an MoE

Table 1. List of noises and their labels in PaPeRo database.

Noise	Label	No. samples
TV	T	422
Human distant speech	A	465
Sudden noise	X	417
Footstep	F	44
Kitchen sounds	K	318

where $g(x)$ is a sigmoid function. The weights ω_{ij} and v_m are trained in such a way that L_i is close to one if the correct noise class is i and zero otherwise.

3 Experiments

3.1 Database

In our experiments, we used the database recorded by a personal robot PaPeRo developed by NEC[3], which was used into the houses of 12 families. The database contains 74640 sounds, each of which is detected by the speech detection algorithm equipped in PaPeRo. The samples recorded by PaPeRo were labelled manually. These sounds were classified into three kinds: speech without noise, noisy speech, and noise without speech. In this study, we used noises without speech. While there were various combinations of different noises in the database, for simplification we used data that contains only one source of noise (see Table 1).

3.2 Experimental conditions

We divided the data into three different sets: training set, development set, and test set. They were divided in the following way: data from eight families were used for training the noise recognition systems and neural networks, and the remaining data from the other four families were distributed between development and test set.

---

* 構造の異なる複数モデルを用いた雑音識別手法
アグニェシカ ベントコフスカ, 篠田浩一, 古井貞熙 (東工大)

Table 2. Topologies of noise HMMs.

Noise	Topology
TV	Type C, 11 states, 2 mix
Human distant speech	Type C, 4 states, 16 mix
Sudden noise	Type D, 3 states, 8 mix
Footstep	Type A, 2 states, 2 mix
Kitchen sounds	Type A, 2 states, 16 mix

The topology of each system was obtained by optimizing it for the noises in a particular noise class. Hence, each discrimination system performs in favor of one kind of noise.

The search for the optimal topology was conducted from among the following four types of HMMs:

Type A: Ergotic HMM

Type B: Left to right HMM

Type C: Left to right HMM where a skip of one state is allowed

Type D: Left to right HMM where a skip of more than one states is allowed

For each HMM type, the number of states was varied from 1 state to 11 states. The number of mixtures per state was chosen from 1,2,4,8,16,32,64. For each noise class, the topology of the model with the minimum description length was chosen. In this way we created five systems with the topologies given by Table 2.

The recognition of the development data and the test data was performed by all five discrimination systems. The best results among those five systems were taken as a baseline. We built two kind of MoE systems:

MoE A: One network for separating all noises.

MoE B: Two-step discrimination network. First a gating network separates verbal-sounds (TV and human distant speech) from non-verbal sounds (kitchen sound, sudden noise and footstep). If a verbal-sound is detected by the gating network, this sound goes to an expert network responsible for separating TV from human distant speech. Similarly, if a non-verbal sound is detected, this sound goes to an expert network responsible for separating kitchen sound, sudden noise and footstep.

3.3 Results

The results are shown in Figure 2. In the case of development set, the recognition accuracy was improved by about 2.4% in MoE A. The accuracy increased when the separation for non-verbal and verbal sound was performed first (MoE B). The segregation for these two groups was conducted with

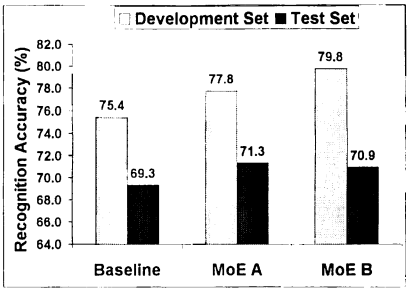

Figure 2. Results for Development Set and Test Set for the two kinds of Mixtures of Experts (MoE)

an error rate of 3.2%. The expert network responsible for separating TV from human distant speech had an accuracy of 78.6% and the network responsible for separating sudden noise, kitchen noise and remaining sounds had an accuracy of 81.0%.

When the test set was used, however, one-step segregation between noises (MoE A) performed better than two-step discrimination (MoE B) (see Figure 2). This might be because of the mismatch between the development set and the test set. For MoE B, the classification error between verbal and non-verbal sounds was 9.1%. The segregation of verbal sounds was performed with an accuracy of 65.4%. Sudden noise, kitchen sounds and footsteps were separated with an accuracy of 77.4%. All of these rates were worse than those for the development set.

4 Conclusions

We presented a noise discrimination method based on combination of results from more than one recognition systems. It was shown that the neural network for separating noise improved the performance by 2.0%. While the MoE architecture did not show any improvements in this study where the number of noise classes was small, it might be effective when there exists a large variety of noises. Researches in this direction seem promising.

Acknowledgment

This work is sponsored by NEC Corporation, and supported in part by 21th Century COE-LKR Program.

References

[1] A. Betkowska, K. Shinoda, and S. Furui, "A study of noise discrimination for personal robots," ASJ autumn meeting 2004,vol. 1,pp. 11-12.

[2] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton, "Adaptive mixtures of local experts," in Neural Comput.,vol 3, pp. 79-87, 1991.

[3] 岩沢透, 大中慎一, 藤田善弘, "状況検知を利用したロボット用音声認識インターフェースの一手法とその評価," 人工知能学会第 16 回 AI チャレンジ研究会, pp. 33-38, 2002.