T2R2 東京科学大学 リサーチリポジトリ
Science Tokyo Research Repository

論文 / 著書情報
Article / Book Information

Title	Noise Adaptation of HMMs Using Neural Networks
Authors	Sadaoki Furui, Daisuke Itoh
Citation	ISCA Workshop on Automatic Speech Recognition, Vol. , No. , pp. 160-167
Pub. date	2000, 9

Powered by T2R2 (Science Tokyo Research Repository)

NOISE ADAPTATION OF HMMS USING NEURAL NETWORKS

Sadaoki Furui and Daisuke Itoh

Department of Computer Science, Tokyo Institute of Technology
2-12-1 Ookayama, Meguro-ku, Tokyo, 152-8552 Japan
furui@cs.titech.ac.jp

ABSTRACT

This paper proposes a new method, using neural networks, of adapting phone HMMs to noise added speech. The network is designed to map clean speech HMMs to noise-adapted HMMs using inputs of clean speech phone HMMs, noise HMMs and signal-to-noise ratios (S/N). The network is trained to minimize the mean squared error between the output HMMs and the target noise-adapted HMMs. Noisy broadcast-news speech was recognized in speaker-dependent and speaker-independent network training conditions, and the trained networks were confirmed to be effective in the recognition of new speakers and under new noise and S/N conditions.

1. INTRODUCTION

Increasing the robustness of speech HMMs (hidden Markov models) in terms of additive noise is one of the most important issues in state-of-the-art speech recognition. HMMs are usually represented by Gaussian mixtures in the multi-dimensional space of cepstral coefficients, in other words, by parameters in the logarithmic spectral domain. However, noise is added to speech in the waveform or in the linear spectral domain, so the incorporation of additive noise into HMMs is not straightforward. Parallel model combination (PMC, also called HMM composition) [1][2] is one of the most useful methods used to handle additive noise. PMC can be used to derive noisy speech HMMs by combining clean speech HMMs, a noise HMM and a signal-to-noise ratio (S/N). However, this method requires nonlinear conversions of the distribution parameters between cepstral and linear spectral domains, and also some approximations in the computational process.

This paper proposes a method using mapping functions of neural networks to represent the total effect of additive noise on HMMs including nonlinear computations. In this method, the neural networks are trained to map clean speech HMMs to noise-added speech HMMs, using noise HMM and S/N ratio as additional input signals. The noise-added speech HMMs used as the target/ideal output signals for the neural networks training are made using noise-added speech produced by summing various clean speech and noises with various S/N ratios. Once the network is trained under various conditions of speech, noise and S/N, the network is expected to produce noise-added speech HMMs under a new condition of speech, noise and S/N within the generalization capability of the neural networks. In the present framework, only the mean vectors of Gaussian mixtures are adapted. Covariance values are preserved unchanged for simplicity.

This paper will explain the principal methods employed, and report on two experiments. The first experiment numerically adds noises by computer to the utterances of a limited number of speakers. Actual utterances by many speakers under various noises are used in the second experiment. The paper will conclude with a general discussion, and discuss issues related to future research.

2. PRINCIPLES OF NOISE-ADAPTED HMMS USING NEURAL NETWORKS

2.1 Fundamental Principles

Figure 1 shows the structure of the neural network that converts clean speech HMMs (Speech HMMs) into HMMs adapted to noise-added speech. As previously mentioned, only the mean vectors are converted, keeping the covariance matrices unchanged. HMMs that model noises (Noise HMMs) are presumed to be represented by a single state with a single Gaussian distribution. S/N values are also required to estimate the HMMs of noise-added speech. Therefore, the inputs of the neural networks are the mean vectors of Gaussian mixtures consisting of Speech HMMs, a Noise HMM, and a S/N ratio. The S/N ratio is implicitly computed in the network by providing mean energy values of both speech and additive noise as inputs to the network.

The Speech HMMs are tied-state context-dependent phone HMMs with 2,106 states in total, and each state has four Gaussian mixtures. Feature vectors have 34 parameters consisting of 16-dimensional LPC (linear predictive coding) cepstra, delta LPC cepstra, log-energy and delta log-energy. Only male utterances are used in the experiments described below. The Speech HMMs were estimated from 13,270 training sentence utterances, spoken by 53 males in a quiet environment. The utterances were sampled at 12kHz, quantized into 16bits, and analyzed with a frame length of 32ms and a frame period of 8ms. Cepstral mean normalization/subtraction is applied sentence by sentence in order to handle the effects of voice individuality and the variation of microphones used in training and testing.

Evaluation experiments were conducted on the recognition of broadcast-news utterances, using a system with a vocabulary size of 20,000 words. Statistical language models were calculated using approximately 400,000 sentences of broadcast news text collected over five years, to which filled pauses were incorporated as pronunciations of punctuation marks [3].

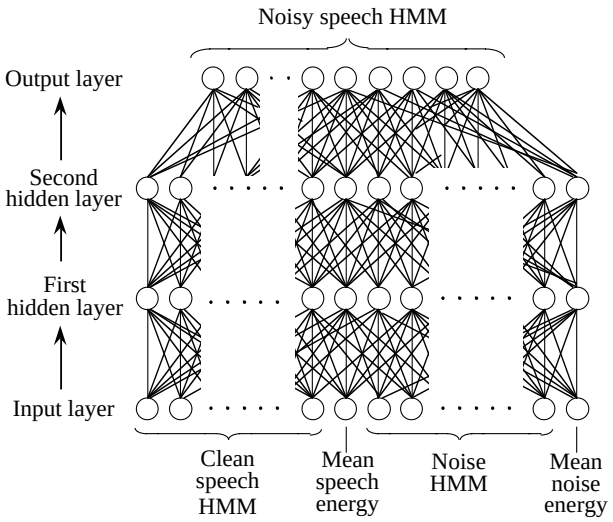

Fig. 1: Structure of neural networks used for HMM noise adaptation.

2.2 Neural Network Training

The neural networks are trained to produce mean vectors of HMMs adapted to noise-added speech (Noisy speech HMMs) as outputs, from inputs of the mean vectors of the Gaussian mixtures of (Clean) Speech HMMs, the mean vectors of Noise HMMs, and the mean energy values of clean speech and noise. Based on preliminary experiments, the neural networks are constructed for each of the 2,106 states comprising the Speech HMMs.

A feed-forward neural network with two hidden layers was employed, and is shown in Figure 1. The (Clean) Speech HMM and Noise HMM each has 35 (34+1) input nodes. The output HMM has 34 output nodes (energy is not included in the output). Training was performed using the error back-propagation method. Evaluation was conducted by mean square error (MSE), the most common measure used for evaluation. Each time an input-output pair is given to the network, the weighting factors of the network are incrementally updated, so that the local MSE between the mean vectors of the target HMMs and the actual network output is minimized.

There are many ways to obtain the target or ideal output of networks, including HMMs made by PMC (HMM composition). In this paper, the target HMMs are made by the combination of MLLR, MAP and VFS methods [4]. Supervised adaptation is performed using the correct transcription of training utterances. An HMM adapted in this way can easily achieve likelihood maximization for short speech data, so that sentences of noise-added speech in a real environment of frequently varying noise can be used individually for training. In the MLLR method, phonemes are clustered into seven classes corresponding to noise, consonants, and each of the five Japanese vowels. A linear transformation is applied to each class.

3. EXPERIMENTS USING ARTIFICIALLY CREATED NOISY SPEECH

3.1 Methods of Experiments

Fourteen broadcast-news sentence utterances by a male speaker (S1) recorded in a quiet studio were used, of which four sentences were used for adapting speaker-independent HMMs to the speaker's voice. The MAP-MLLR-VFS method that was used for making target noisy speech HMMs was also used for speaker adaptation. The remaining 10 sentences were used for evaluation. The following six noises were added to the utterances for training and the utterances for testing, with seven S/N ratios: 0, 2, 5, 7, 10, 12, and 15dB.

- Noise A: Elevator hall noise
- Noise B: Crowd noise
- Noise C: Computer room noise
- Noise D: Street noise
- Noise E: Noise in cars

- Noise F: Exhibition hall noise

In order to shorten computational time, only word unigrams and bigrams were used as language models for speech recognition in the experiments in this section.

The four following neural networks were trained.

- NN9-1: Trained in nine conditions: combinations of three noises (Noise A, C, E) and three S/N values (2, 7, 12dB)
- NN9-2: Trained in nine conditions: combinations of three noises (Noise B, D, F) and three S/N values (2, 7, 12dB)
- NN15-1: Trained in 15 conditions: combinations of three noises (Noise A, C, E) and five S/N values (2, 5, 7, 10, 12dB)
- NN15-2: Trained in 15 conditions: combinations of three noises (Noise B, D, F) and five S/N values (2, 5, 7, 10, 12dB)

3.2 Experiments on the Training Set

The following experiments were conducted using the training utterances to examine whether the output HMMs of trained networks achieved a similar level of performance as the target HMMs.

(1) Under the same noise and S/N conditions as training

In order to examine whether the networks were well trained for the input-output relationships applied in training, the same noise-added speech that was used in training was recognized using the obtained output HMMs (NN-HMMs), under the same noise and S/N conditions as training. The four sentence utterances were recognized using the NN-HMMs of the two neural networks, NN9-1 and NN9-2. Word accuracy averaged over all noise and S/N conditions is shown in Fig. 2. For comparison, results of speaker-independent HMMs (SI-HMM), speaker-adapted HMMs (SA-HMM) before noise adaptation, and the target HMMs used in network training are also shown in the figure. These results show that word accuracy close to that of the target HMMs was achieved by the NN-HMMs produced by the networks, indicating that the networks were well trained.

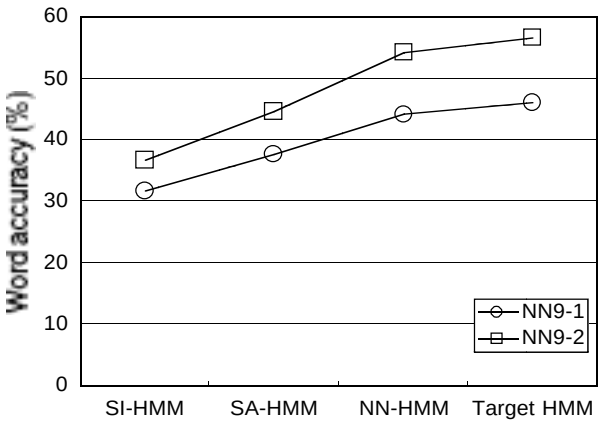

Fig. 2: Word accuracy for same noise and S/N as training.

(2) Conditions of same noise but different S/N to training

The performance of NN-HMMs for the training utterances under the same additive-noise conditions as training, but with different S/N ratios was investigated. Specifically, networks NN9-1 and NN9-2 were used, and the same three noises were added to the clean utterances, this time with S/N ratios of 0, 5, 10, and 15 dB (different from the training conditions 2, 7, 12 dB). The noise-added utterances were recognized using NN-HMMs. The results, averaged over all three noises, are shown in Fig. 3. The word accuracy of NN-HMMs is close to that of the target HMM, which indicates that the networks trained with S/N ratios of 2, 7, and 12 dB are applicable to new S/N conditions, including S/N ratios of 0 and 15 dB which are outside of the range of trained conditions.

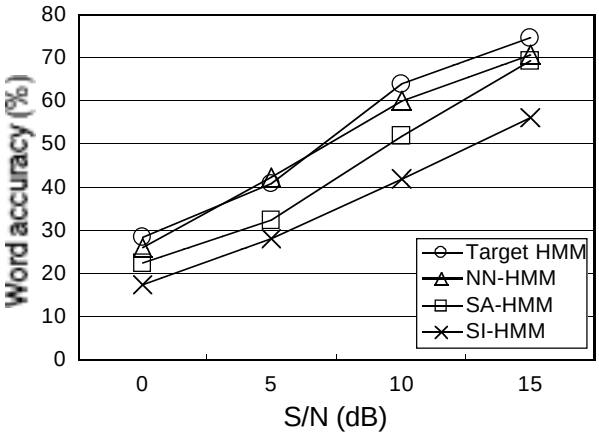

Fig. 3: Word accuracy for same noise conditions as training with different S/N.

(3) Experiments using noises different from training

The following NN-HMMs were examined in terms of their applicability to noises different from those applied in training.

- NN-HMMs obtained as the output of NN9-1 and NN15-1 (trained by noises A, C, and E) when noise B, D, or F is added to speech and used as noise input to the neural networks with S/N of 0, 2, 5, 7, 10, 12 or 15 dB.
- NN-HMMs obtained as the output of NN9-2 or NN15-2 (trained by noises B, D, and F) when noise

A, C, or E is added to speech and used as noise input to the neural networks with S/N of 0, 2, 5, 7, 10, 12 or 15 dB.

The experimental results are shown in Fig. 4. NN-HMMs achieved word accuracy close to that of the target HMM, which indicates that the networks are applicable to noises different from those used in training.

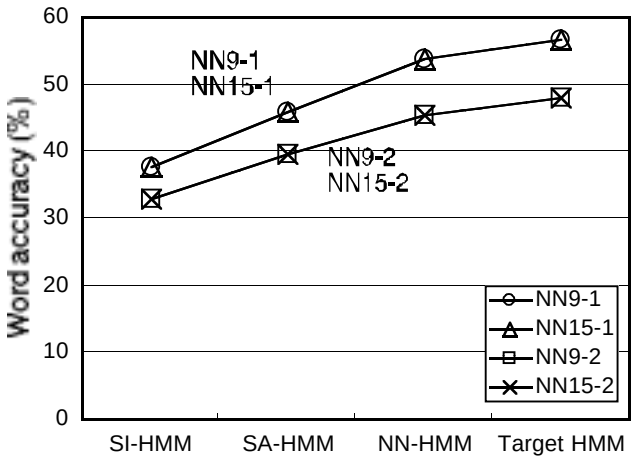

Fig. 4: Word accuracy for noises different from training.

3.3 Experiments on the Test Set

The following experiments were performed using the noise-added test sets each containing 10 sentences.

(1) Under the same noise and S/N conditions as training

Noise-added test set utterances were recognized by NN-HMMs produced by NN9-1 and NN9-2, under the same additive-noise and S/N conditions as training. The results are shown in Fig. 5. The NN-HMMs achieved even better results than the target HMM used to train the networks. This indicates that the target HMMs made by the MAP-MLLR-VFS method are over-tuned to the training sets and the NN-HMMs are more effective for new input utterances.

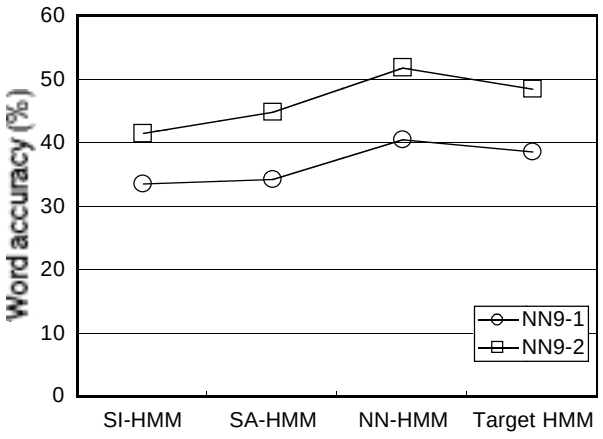

Fig. 5: Word accuracy for test set with same noise and S/N as training.

(2) Conditions of same noise but different S/N to training

Recognition experiments were performed using NN9-1 and NN9-2 under the same noise conditions as training, but with different S/N ratios. The results shown in Fig. 6 show that NN-HMMs achieve performance equal to or better than that of the target HMM, even in the 0 and 15 dB conditions that exceed training conditions. This indicates that the networks are applicable to new S/N conditions.

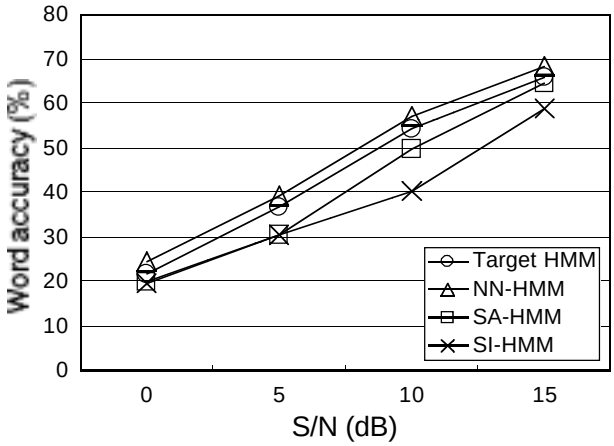

Fig. 6: Word accuracy for test set with same noise as training but different S/N.

(3) Experiments using noises different to training

Experiments using the test set with noises different to those used in training were performed. The results in Fig. 7 show that NN-HMMs achieve better results than the target HMM. This indicates that the networks are effective for new noises.

3.4 Experiments Using New Speakers

A supplementary experiment was performed, employing new speakers (S2, S3, and S4) in addition to the speaker (S1) who was used in the previous experiments, to check the effectiveness of the networks in recognizing new speakers. Table 1 shows the number of sentence

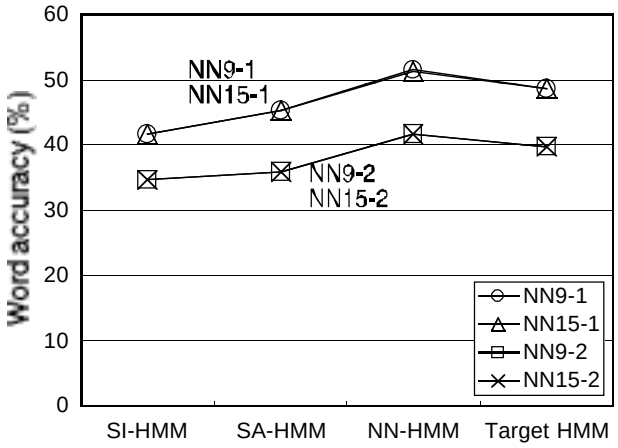

Fig. 7: Word accuracy for test set with noises different from training.

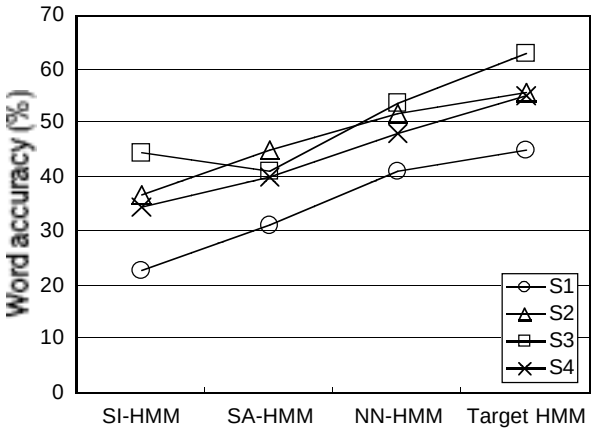

Fig. 8: Word accuracy for noise-added speaker adaptation utterances.

utterances used in the experiment. The speaker-independent HMM was first adapted to each speaker using three or four clean utterances, as shown in the table. The speaker-adapted HMMs were then adapted to additive-noise using the NN9-1 network trained with the utterances of S1. Noise F, which was not included in the training of NN9-1, was used for evaluation with S/N ratios of 0, 2, 5, 7, 10, 12, and 15 dB. For comparison, target HMMs were directly produced by adapting the speaker-independent HMM to noisy utterances of each speaker, where the noisy utterances were made by adding noise to the utterances used for speaker adaptation.

Table 1: Number of sentence utterances used in experiment

Speaker	For speaker adaptation	For evaluation	Total
S1	4	10	14
S2	4	10	14
S3	4	4	8
S4	3	3	6

(1) Recognition results for the noise-added speaker adaptation utterances

Figure 8 shows recognition results for the noisy utterances, which were made by adding noise to the utterances used for adapting speaker-independent HMMs to each speaker. The results show that improvement comparable to that for S1 is achieved for speakers S2, S3 and S4 by using the NN-HMMs. This indicates that the neural networks trained using the utterances of S1 are also effective for other speakers, even if new noise is added to the speech.

(2) Recognition results for test set utterances

Recognition results for test set utterances are shown in Fig. 9. Although the performance of the NN-HMM exceeds that of the target HMM for speaker S1, to whom the network was trained, performance is lower for speakers S2, S3 and S4. However, the performance of the NN-HMM for the latter three speakers is still very close to that of the target HMM. This indicates that the network is effective in recognizing new utterances by new speakers with new noises.

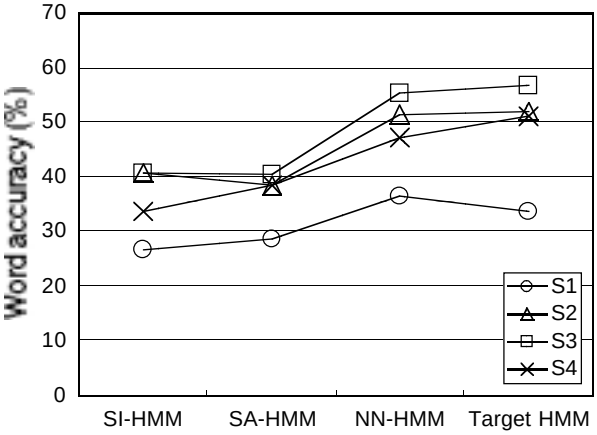

Fig. 9: Word accuracy for the test set utterances.

4. EXPERIMENTS USING REAL NOISE-ADDED SPEECH SIGNALS OF MANY SPEAKERS

4.1 Experimental Method

Fifty sentence utterances by many speakers, superimposed

with noise and music, were extracted from real broadcast-news speech, such as reports, and used for the experiments. From the 50 utterances, two sets of 10 utterances (Case 1 and Case 2) were randomly chosen for neural network training, and the remaining 40 utterances were used for testing each case. The distribution of S/N between the training and test sets of utterances is shown in Table 2. A multiple search decoder was used in this experiment, and bigrams and trigrams were used as language models in the first and second passes of the search, respectively.

Since clean speech for each speaker was not available, target HMMs were produced by adapting speaker-independent HMMs trained by clean speech (Speech HMMs) to training utterances (noisy utterances by multiple speakers). A combination of MAP-MLLR and VFS methods was used for noise adaptation in the same way as in the previous experiments.

Table 2: Distribution of S/N among training and test sets of utterances

		Number of sentences	Min-Max	Average	Standard deviation
Case 1	Training set	10	13.8-19.8 dB	17.1 dB	2.0 dB
	Test set	40	8.2-26.3 dB	17.3 dB	5.3 dB
Case 2	Training set	10	9.3-23.1 dB	14.1 dB	3.9 dB
	Test set	40	8.2-26.3 dB	18.0 dB	4.7 dB

The Noise HMM, having a single state with a single Gaussian distribution, is trained using non-speech segments of each utterance in the training set. Neural networks were trained using the training sets of 10 sentence utterances. Mean vectors of the Speech HMM and the Noise HMM as well as mean energy values were supplied to the network as inputs, and the mean vectors of the target HMM provided as outputs. There were 40 input-output pairs (10 sentences by 4 mixtures) for each training set.

4.2 Results for the Training Sets

An experiment was conducted to determine whether the trained networks achieved performance close to the target HMM for the training sets of broadcast-news speech. The training set utterances were recognized using the NN-HMM obtained from the trained network. Word accuracy is shown in Fig. 10. Speaker-independent HMMs, before noise adaptation, are represented by SI-HMM. The figure shows that the performance of NN-HMM is much better than SI-HMM, and is close to that of the target HMM. This indicates that the network is well trained, at least for the training set.

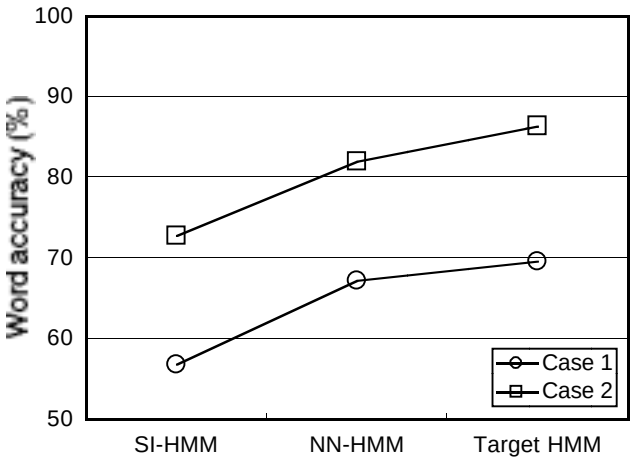

Fig. 10: Word accuracy for training sets.

4.3 Results for the Test Sets

The corresponding test set experiment was then performed. Additionally a target HMM was produced from each test utterance by supervised adaptation using the MAP-MLLR-VFS method, and its performance was compared with that of an NN-HMM produced by a neural network made from the training set. The results are shown in Fig. 11. The figure shows that the NN-HMM, produced by the network driven by the mean vectors of speech and noise HMMs and corresponding energies, is useful for improving recognition performance. However, the improvement over the SI-HMM by using the NN-HMM on the test set is smaller than the improvement seen in the training set experiment, and performance is significantly below the target HMM derived from the noise-added test utterance itself.

The ratio between improvement and the difference between the performance of the target HMM and the speaker-independent clean speech HMM, is approximately 40% for the test set, though it was about 80 % for the training sets.

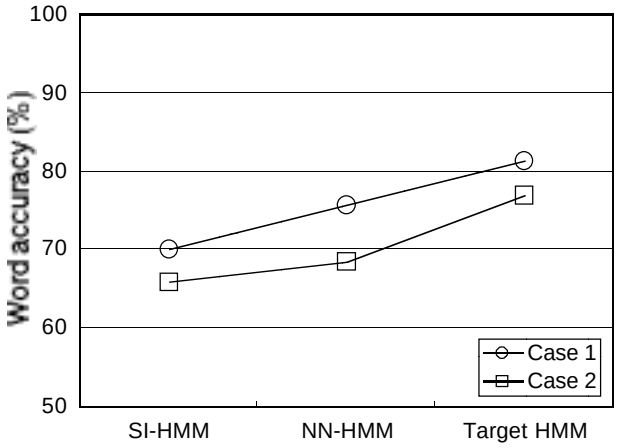

Fig. 11: Word accuracy for test sets.

The main cause of this discrepancy between test set and training set results is believed to be the fact that the speakers of the utterances for neural network training and testing were the same in the training-set case and different in the test-set case. In the test-set case, the NN-HMMs were adapted to only noise effects, whereas the target HMMs were adapted to both noise and speaker effects. Thus the improvement by the NN-HMMs in the test-set case was only half of the target HMMs.

4.4 Analysis of the Method's Effectiveness as a Function of S/N

The 40 utterances in cases 1 and 2 exhibit a wide range of S/N ratios, as shown in Table 2. Figure 12 shows a breakdown of word accuracy for the test set according to levels of S/N. This figure shows that relatively good results are obtained by using the NN-HMM at S/N ratios below 20dB, while no improvement is observed above 20dB. This is probably because the noise level in the

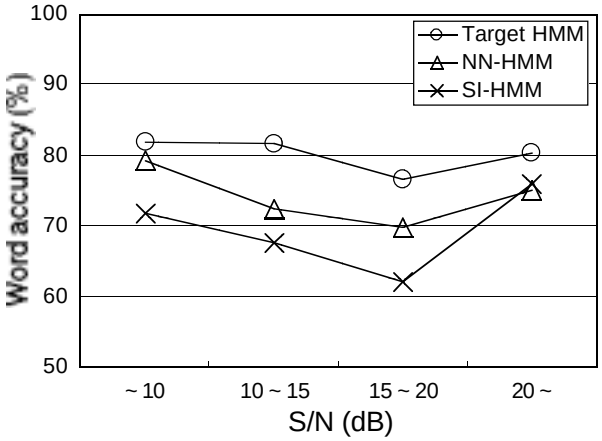

Fig. 12: Word accuracy for test set according to levels of S/N.

S/N range above 20dB is very low, and therefore the effect of noise on speech is low, so the target HMMs are mainly adapting to the voice individuality of the speaker. As a result, the network trained by using different speakers' utterances is not effective in such condition.

5. CONCLUSION

This paper has reported the investigations of HMM adaptation using neural networks, with the intent of improving large-vocabulary continuous-speech recognition accuracy for noise-added speech. The networks were trained by applying the mean vectors of phone HMMs estimated from clean speech, the mean vectors of noise HMMs estimated from noise extracted from input utterances, and the mean energies of both clean speech and noise as input values, and applying the mean vectors of HMMs adapted to noise-added speech as output values. Once trained, the network produces noise-adapted HMMs that use clean speech phone HMMs, a noise HMM, and the mean energies of speech as inputs into the network under varying noise conditions.

The first experiment involved using the speech signals of clean speech uttered by a single speaker with numerically-added noise for neural network training and various evaluation experiments. The results of the experiment showed that the output HMMs of the neural networks achieved almost the same word accuracy as the HMM made by supervised adaptation using noise-added speech. These results are indicative of the fundamental effectiveness of the method proposed. Neural networks trained by a single speaker's utterances were then applied to noise-added speech uttered by different speakers, and it was confirmed that the neural networks could be successfully applied to new speakers and new noises.

The second experiment was performed using real broadcast news speech distorted by relatively high-level noises, including real reports from various fields. Clean speech was unavailable for the speakers in this experiment. Consequently, speaker-independent HMMs, instead of speaker-adapted HMMs, were used as input vectors for the neural networks, in conjunction with noise HMMs and S/N ratios. The neural networks were trained such that the mean squared error between the outputs of the network and the HMMs adapted to noise-added speech was minimized. The training process learns not only noise effects but also voice individuality included in the training utterances. This means that the training process for noise

effects is contaminated by voice individuality. Therefore the degree of noise adaptation in the speaker-independent recognition experiments was approximately half that of the speaker dependent neural network training case. Nevertheless the effectiveness of the proposed method for noise-added speech under real conditions was confirmed.

It may be beneficial to attempt to apply the neural networks trained to the noise-added single speaker utterances in the previous experiments, to the adaptation of speaker-independent HMMs and use the corresponding output HMMs for recognizing noisy broadcast news speech. Future work will include neural network training for additive noise using a large database consisting of pairs of clean speech and noise-added speech by many speakers, after separating the effects of voice individuality.

So far, experiments have used the MAP-MLLR method combined with the VFS technique to produce target HMMs with short utterances. This method is not necessarily the best for adapting HMMs for additive noise, and better techniques such as PMC (HMM composition) with the likelihood maximization framework [5] should be considered.

Future investigations include trials of using Gaussian mixture noise modeling instead of using single Gaussian modeling and the adaptation of covariance matrix components of speech HMMs. It is also important from a practical perspective to establish an automatic and efficient method for separating speech and noise periods. Although this paper investigated only the influence of additive noise, actual speech usually contains various combination of distortions including multiplicative (convolutional) distortions. It is therefore important to investigate new methods to handle the components of complex distortions at the same time.

ACKNOWLEDGMENTS

The authors wish to express their appreciation of Professor Itsuo Kumazawa at Tokyo Institute of Technology for several fruitful discussions on neural networks. The authors would like to thank NHK (Japan Broadcasting Corporation) for providing the broadcast news database. This work is supported in part by the International Communications Foundations.

REFERENCES

[1] M. J. F. Gales and S. J. Young: “An improved approach to the hidden Markov model decomposition of speech and noise”, Proc. ICASSP, pp. 233-236 (1992)

[2] F. Martin, K. Shikano and Y. Minami: “Recognition of noisy speech by composition of hidden Markov models”, Proc. Eurospeech, pp. 1031-1034 (1993)

[3] K. Ohtsuki, S. Furui, N. Sakurai, A. Iwasaki and Z.-P. Zhang: “Recent advances in Japanese broadcast news transcription”, Proc. Eurospeech, pp. 671-674 (1999)

[4] S. Furui, Z.-P. Zhang and K. Ohtsuki: “On-line incremental speaker adaptation for broadcast news transcription”, Proc. IEEE Automatic Speech Recognition and Understanding Workshop, pp. 165-168 (1999)

[5] Y. Minami and S. Furui: “A maximum likelihood procedure for a universal adaptation method based on HMM composition”, Proc. ICASSP, pp. 129-132 (1995)