NEURAL-NETWORK-BASED HMM ADAPTATION FOR NOISY SPEECH

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 noisy speech. The neural networks are designed to map clean speech HMMs to noise-adapted HMMs, using noise HMMs and signal-to-noise ratios (SNRs) as inputs, and are trained to minimize the mean square error between the output HMMs and the target noise-adapted HMMs. In evaluation, the proposed method was used to recognize noisy broadcast-news speech in speaker-dependent and -independent modes. The trained networks were confirmed to be effective in recognizing new speakers under new noise and various SNR conditions.

1. INTRODUCTION

Increasing the robustness of speech HMMs (hidden Markov models) to additive noise is one of the most important issues in state-of-the-art speech recognition. HMMs with Gaussian mixtures are usually used to model speech represented by cepstral coefficients, meaning that speech is modeled in the logarithmic spectral domain. However, noise is often additive 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 practically useful methods used to handle additive noise. PMC can derive noisy speech HMMs by combining clean speech HMMs, a noise HMM and a signal-to-noise ratio (SNR). However, this method requires numerical conversion of the distribution parameters between cepstral and linear spectral domains.

This paper proposes a method using neural network mapping functions to learn the effects of additive noise on HMMs. The neural network is trained using an input consisting of a clean speech HMM, noise HMM and the SNR. The output of the neural network is a noisy speech HMM which, during training, is obtained by a combination of MLLR, MAP and VFS adaptation techniques. The neural network learns the mapping between the input and output. During testing, the mapping is used to obtain the noisy speech HMM from the inputs. Once the network is trained under various conditions of speech, noise and SNR, the network is expected to produce noise-added speech HMMs under new speech, noise and SNR conditions within the bounds of generalization capabilities of neural networks. In the present framework, only the mean vectors of Gaussian mixtures are adapted, and covariance values are preserved unchanged for simplicity.

This paper explains the principal methods employed, and then reports on two experiments. The first experiment numerically adds noises to utterances by a limited number of speakers. Actual utterances by a wide range of speakers under various noise conditions are used in the second experiment. The paper concludes with a general discussion and issues related to future research.

2. PRINCIPLES OF NOISE-ADAPTED HMMS USING NEURAL NETWORKS

2.1 Fundamental Principles

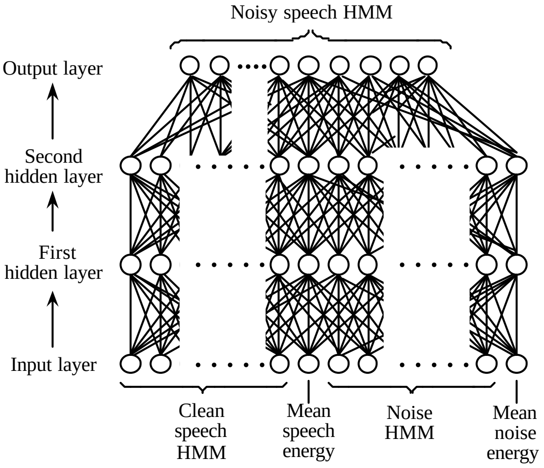

Fig. 1: Structure of the neural network used for HMM noise adaptation.

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 different noise types (Noise HMMs) are presumed to be represented by a single state with a single Gaussian distribution. Since SNR values are also necessary to estimate the HMMs of noise-added speech, inputs of the neural networks are mean vectors of Gaussian mixtures included in a Speech HMM and a Noise HMM, and an SNR. The SNR 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, 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 with 16 bits, 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 the pronunciation of punctuation marks [3].

2.2 Neural Network Training

The neural network is trained to output mean vectors of the HMM adapted to noise-added speech (Noisy speech HMM), using as inputs the mean vectors of the Gaussian mixtures of (Clean) Speech HMM and Noise HMM, and the mean energy values of clean speech and noise. Based on preliminary experiments, one neural network is constructed for each of the 2,106 states in the Speech HMM.

A feed-forward neural network with two hidden layers as shown in Figure 1 is employed. The (Clean) Speech HMM and Noise HMM each have 35 (34+1) input nodes, whereas the output HMM has only 34 output nodes, since energy is not used in recognition. Training is performed using the error back-propagation method. Evaluation was conducted by mean square error (MSE), the most commonly used evaluation measure. 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 HMM, 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 individually used for training. In the MLLR method, phonemes are clustered into seven classes corresponding to noise, consonants, and each of the five Japanese vowels , and a linear transformation is applied to each class.

3. EXPERIMENTS USING ARTIFICIALLY CREATED NOISY SPEECH

3.1 Experimental Method

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 a speaker-independent HMM 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 noise types were added to training and testing utterances with seven SNRs: 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 bigrams were used as the language model for speech recognition experiments in this section.

The following neural networks were trained.

- NN-1: Trained under nine conditions: combinations of three noise types (Noise A, C, E) and three SNR values (2, 7, 12dB)
- NN-2: Trained under nine conditions: combinations of three noise types (Noise B, D, F) and three SNR values (2, 7, 12dB)

3.2 Results

(1) Same noise and SNR conditions as training

Noise-added test utterances were recognized by two HMMs obtained as the output of neural networks (NN-HMMs) produced by NN-1 and NN-2, respectively, under the same additive-noise and SNR conditions as training. Word accuracy averaged over all noise and SNR 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. The NN-HMMs achieved even better results than the target HMM used to train the networks. This indicates that the target HMMs may be over-tuned to the training sets and the NN-HMMs are more effective for new input utterances.

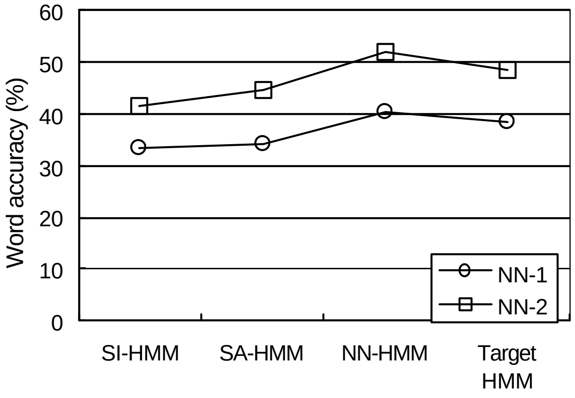

Fig. 2: Word accuracy for same noise and SNR conditions as training.

(2) Same noise but different SNR from training

Recognition experiments were performed using NN-1 and NN-2 under the same noise conditions as training, but with different SNRs. That is, the same three noise types were respectively added to the clean test utterances, this time with SNRs of 0, 5, 10, and 15 dB (different from the training conditions 2, 7, 12 dB). The results in Fig. 3, averaged over all three noise types, show that NN-HMMs trained with SNRs of 2, 7, and 12 dB achieve performance equal to or better than that of the target HMM in new SNR conditions, even at 0 and 15 dB which are outside the range of trained conditions.

(3) Different noise types from training

The following experiments were performed to test the applicability of NN-HMMs to new noises.

- NN-HMMs obtained as the output of NN-1 (trained on noises A, C, and E) were tested on utterances with noise B, D, or F with SNR of 0, 2, 5, 7, 10, 12 or 15 dB.
- NN-HMMs obtained as the output of NN-2 (trained on noises B, D, and F) were tested on utterances with noise A, C, or E with SNR of 0, 2, 5, 7, 10, 12 or 15 dB.

The results in Fig. 4 show that NN-HMMs achieve better results than the target HMM. This indicates that the networks are effective for new noise types.

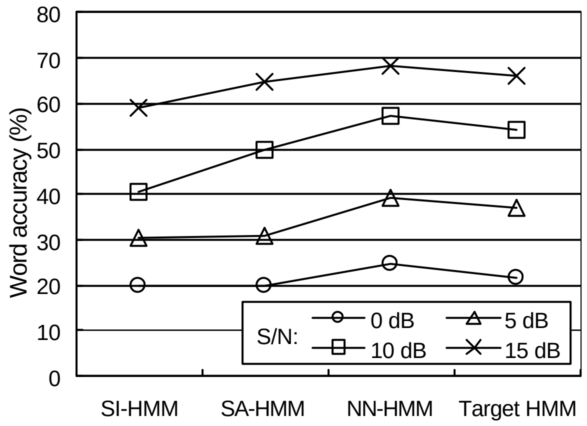

Fig. 3: Word accuracy for same noise conditions as training but with different SNR.

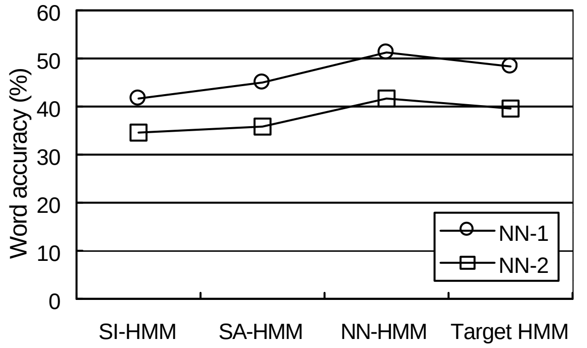

Fig. 4: Word accuracy for different noise types from training.

3.3 Experiments Using New Speakers

A supplementary experiment was performed, employing new speakers (S2, S3, and S4) in addition to the speaker used in the previous experiments (S1), to check the effectiveness of the neural networks in recognizing new speakers. Table 1 shows the number of sentence utterances used in the experiment. The speaker-independent HMM was first adapted to each new speaker using either three or four clean utterances, as shown in the table. The speaker-adapted HMMs were then adapted to additive-noise using the NN-1 trained with the utterances of S1. Noise F, which was not included in the training of NN-1, was used for evaluation with SNRs 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 made by adding noises to the utterances used for speaker adaptation.

Table 1: Number of sentence utterances used in experiment

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

Recognition results are shown in Fig. 5. The performance of the NN-HMM for the new 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 noise types.

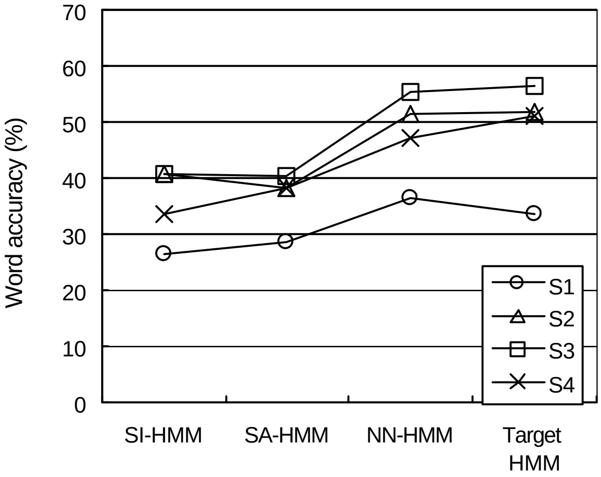

Fig. 5: Word accuracy for new speakers and new noise.

4. EXPERIMENTS USING REAL NOISE-ADDED SPEECH BY A WIDE RANGE OF SPEAKERS

The technique was tested using real broadcast-news speech that included reports from remote sites. Fifty sentence utterances by a wide range of speakers, superimposed with noise and music, were extracted from the real broadcast-news speech and used for the experiments. This represents a difficult task, since the noise are relatively unstationary and cannot realistically be modeled using only non-speech segments of each utterance. 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 in each case. The distribution of SNR 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 the language model in the first and second passes of the search, respectively.

Since clean speech was not available for each speaker, HMM speaker adaptation could not be performed in this experiment. Consequently, a target HMM for neural network training was produced by adapting the SI clean speech HMM (SI-HMM) to each noisy training utterance. The MAP-MLLR-VFS method was used for adaptation in the same way as in the previous experiments.

The noise HMM, consisting of a single state with a single Gaussian distribution, is trained using non-speech segments of each utterance in the training set. The neural networks are then trained so that SI-HMM, instead of SA-HMM, is mapped to the target noisy speech HMM. Therefore, not only noise but also speaker effects are learned by the neural network. This means that the training process for noise effects is contaminated by voice individuality. Two neural networks were trained using each training set of 10 sentence utterances. There were 40 input-output pairs (10 sentences by 4 mixtures) for each training set.

Results broken down into different SNR ranges are given in Fig. 6. Here, for comparison, a target HMM was produced for each test utterance by supervised adaptation using the MAP-MLLR-VFS method. The figure shows that the NN-HMM is useful for improving recognition performance at

SNRs below 20dB, while no improvement is observed above 20dB. This is probably because the noise level in the SNR 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 conditions. The overall gap between performance of the NN-HMM and the target HMM is considered to be attributable to speaker adaptation.

A supplementary experiment was performed to compare the results of NN-based noise adaptation with that of the PMC (HMM composition) method. Experimental results showed that NN-based method performed significantly better than PMC. This may be because baseline performance prior to adaptation for the latter method was significantly worse than the former method. The reason for the poor baseline performance of PMC is due to the fact that cepstral mean normalization/subtraction cannot be combined with PMC.

Table 2: Distribution of SNR among training and test sets of utterances.

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

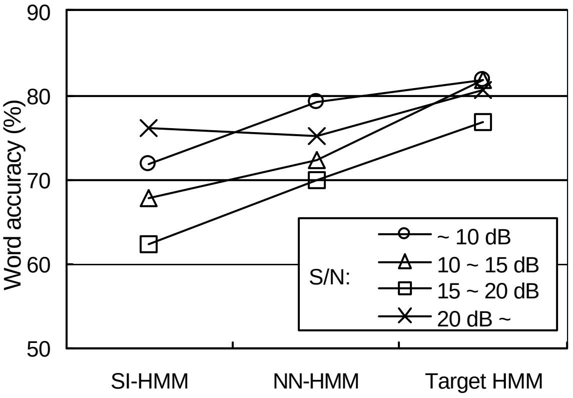

Fig. 6: Word accuracy for different SNR levels.

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 by a single speaker with noise for neural network training under various evaluation conditions. 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 indicate 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 noise types.

The second experiment was performed using real broadcast news speech distorted by varioustypes of noise, including reports from remote sites. 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 first experiment, to adapting speaker-independent HMMs and use the 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 out the effects of voice individuality.

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

Future investigations will include trials using Gaussian mixture noise modeling instead of 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 involves the combination of various distortions including multiplicative (convolutional) distortions. It is therefore important to investigate new methods to simultaneously handle the various components of complex distortions.

ACKNOWLEDGMENTS

The authors wish to express their appreciation of Professor Itsuo Kumazawa of 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)