SPECTRAL SUBTRACTION IN NOISY ENVIRONMENTS APPLIED TO SPEAKER ADAPTATION BASED ON HMM SUFFICIENT STATISTICS

Shingo Yamade[*1], *Kanako Matsunami*[*1], *Akira Baba*[*1*2], *Akinobu Lee*[*1],
Hiroshi Saruwatari[*1], *Kiyohiro Shikano*[*1]
[*1] Nara Institute of Science and Technology, Japan, [*2] Matsushita Electric Works, Ltd., Japan
shing-y, kanako-m, a-baba, ri, sawatari, shikano@is.aist-nara.ac.jp

ABSTRACT

Noise and speaker adaptation techniques are essential to realize robust speech recognition in real noisy environments. In this paper, we applied spectral subtraction to an unsupervised speaker adaptation algorithm in noisy environments. The adaptation algorithm consists of the following five steps. (1) Spectral subtraction is carried out for noise added database. (2) Noise matched acoustic models are trained by using noise added speech database. (3) HMM sufficient statistics for each speaker are calculated from noise added speech database, and stored. (4) According to one arbitrary utterance, speakers close to a test speaker are selected by using speaker GMMs. (5) Speaker adapted acoustic models are constructed from HMM sufficient statistics of the selected speakers. We evaluated our unsupervised speaker adaptation algorithm in noisy environments in the 20k dictation task. The recognition experiments show that our speaker adapted acoustic model can achieve 82% word accuracy in 20dB SNR, which is about 6% higher than that of the noise matched models trained by Forward-Backward algorithm.

We also investigated the robustness of the adapted models in various SNR conditions. Integration with the supervised MLLR is also examined.

1. INTRODUCTION

Speech recognition in noisy environments can be improved with noise reduction techniques such as spectral subtraction [1]. In this paper, our proposed unsupervised speaker adaptation algorithm [8] is combined with spectral subtraction and applied to the large vocabulary continuous speech recognition in noisy environments. Large vocabulary continuous speech recognition in real noisy environments requires a noise adaptation as well as a speaker adaptation [2][7]. There exist huge numbers of different noises. It is almost impossible to collect all kinds of environment noise data beforehand. Usually speaker adaptation and/or noise adaptation algorithms require for a user to utter several tens sentences.

The proposed speaker adaptation algorithm in noisy environments is evaluated in the 20k vocabulary newspaper dictation task [4] with spectral subtraction. We attain 72.3%, 81.8% and 87.2% word accuracy rates in 15dB, 20dB and 25dB SNR conditions, respectively. These word accuracy rates are better than those of the noise matched models trained by Forward-Backward algorithm using the noise added whole speech database.

We also evaluate the robustness of the spectral subtraction for the adapted acoustic models in different SNR conditions. The adapted acoustic models with spectral subtraction are robust even in the different SNR conditions. The adapted acoustic models are also useful as an initial model for the supervised MLLR adaptation.

2. SPECTRAL SUBTRACTION

Spectral subtraction is a technique to reduce noise from noisy speech by subtracting noise spectrum from noisy speech [1]. Spectral subtraction offers a computationally efficient technique for reducing noise by using the FFT. Assume that a speech signal $s(n)$ has been degraded by an uncorrelated additive noise $v(n)$. The corrupted noisy speech $x(n)$ can be expressed as

$$x(n) = s(n) + v(n).$$

Taking the DFT of $x(n)$ gives

$$X(k) = S(k) + V(k).$$

Assuming that $v(n)$ is zero-mean and uncorrelated with $s(n)$, the estimate of $|S(k)|$ can be expressed as

$$|\hat{S}(k)|^2 = |X(k)|^2 - a\,E|V(k)|^2,$$

where $E|V(k)|$ is the expected noise spectrum taken during the non-speech period, and a is a subtraction parameter. In this paper, $E|V(k)|$ is estimated from 300msec noise period of every utterance. To avoid negative speech spectrum, flooring operation is introduced as

$$|\hat{S}(k)| = |X(k)|\,A,$$

when $|\hat{S}(k)|^2 = |X(k)|^2 - a\,E|V(k)|^2 < 0$.
A is a flooring parameter. $a = 2.0$ and $A = 0.5$ are adopted according to preliminary experiments.

3. SPEAKER ADAPTATION ALGORITHM IN NOISY ENVIRONMENTS AND EVALUATION

The procedure for the proposed speaker and noise adaptation algorithm is shown in Figure 1. This algorithm requires only one arbitrary utterance and a few minutes of noise data. JNAS speech database [3] from 306 speakers are adopted as the algorithm implementation and evaluation.

3.1. Speaker adaptation algorithm in noisy environments

The adaptation procedure consists of the following five steps.

(Step 1) Spectral subtraction is carried out for noise added speech database JNAS. Noise spectrum is estimated from the 300msec noise-period at the beginning of every utterance. $a = 2.0$ and $A=0.5$ are used.

(Step 2) Noise matched speaker-independent acoustic models are trained based on noise added speech database with Forward-Backward algorithm.

(Step 3) HMM sufficient statistics for each speaker, which include average, variance and E-M count of each Gaussian distribution, are calculated from noise added speech database using noise matched speaker-independent acoustic models, and stored.

(Step 4) According to one arbitrary utterance, speakers close to a test speaker are selected by using speaker GMMs, where each speaker GMM with one-state 64 Gaussian mixture is beforehand trained from 150 sentence utterances. Speaker selection from 260 JNAS database training speakers is carried out based on one noise added arbitrary utterance and 260 GMM speaker models. To avoid the noise effects in the speaker selection, the likelihood values from the speaker GMMs are calculated from only the speech part frames by discarding the low power frames [7].

(Step 5) Speaker adapted acoustic models are constructed from HMM sufficient statistics from the selected speakers. This calculation procedure is equivalent to the one-iteration of the HMM Forward-Backward training algorithm from the SNR matched speaker-independent model.

3.2. Evaluation experiment in large vocabulary continuous speech recognition

The proposed speaker adaptation algorithm in noisy environments is evaluated with a large vocabulary continuous speech recognition task. We adopt two types of HMM acoustic models, simple monophone models and accurate and computationally efficient PTM (phonetic tied mixture models based on triphones) [5].

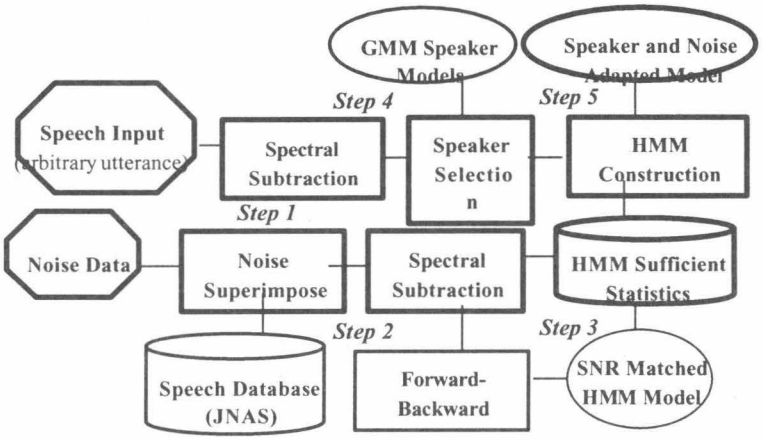

Figure 1: Speaker Adaptation Based on HMM Statistics in Noise

3.2.1. Evaluation task and conditions

The evaluation task is the JNAS newspaper dictation task with 20k vocabulary size [4][5]. The baseline speaker-independent acoustic models are trained from 260 training speakers' data in the JNAS speech database [3]. The training speech database includes 260 speakers (39,000 sentence utterances in total). The test set contains 46 speakers from JNAS. Each test speaker utters 4 or 5 newspaper article sentence utterances (200 test sentence utterances in total), according to the IPA '99 test set [4]. We also adopt the decoder JULIUS and the language model from the IPA dictation project [5]. The experiment conditions are summarized in Table 1. The noise added experiment data are prepared by superimposing the office noise on the JNAS clean database according to three SNR levels, 15dB, 20dB and 25dB SNR.

Table 1: Experiment conditions

Number of Speakers in JNAS Training Database	260 speakers (130 male speakers, and 130 female speakers)
Speaker GMM	64 Gaussian mixture
Number of Selected Speakers for Sufficient Statistics Adaptation	20 speakers for monophone model, 40 speakers for PTM
Speech Analysis and Feature Extraction	25 msec hamming window, 10 msec frame shift, CMN based on a sentence utterance, 12 MFCC, 12 delta-MFCC, and delta-power
Noise Data and Spectral Subtraction	Office environment (3 minutes). $a = 2.0$ and $A=0.5$

3.2.2. Evaluation experiments

First, HMM sufficient statistics for each training speaker are calculated from the noise matched speaker-independent model using the noise added training JNAS

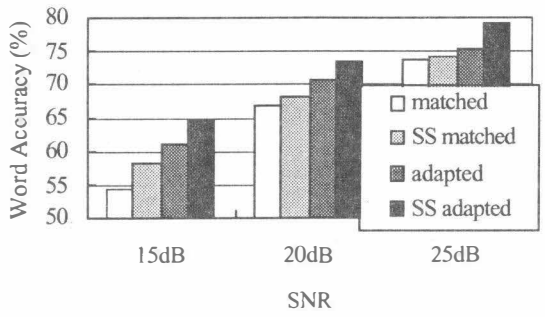

Figure 2: Spectral Subtraction Effects for Speaker Adapted Monophone Models

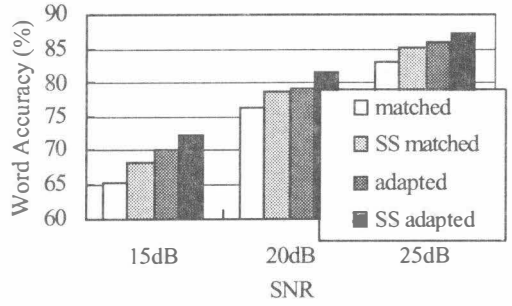

Figure 3: Spectral Subtraction Effects for Speaker Adapted PTM Models

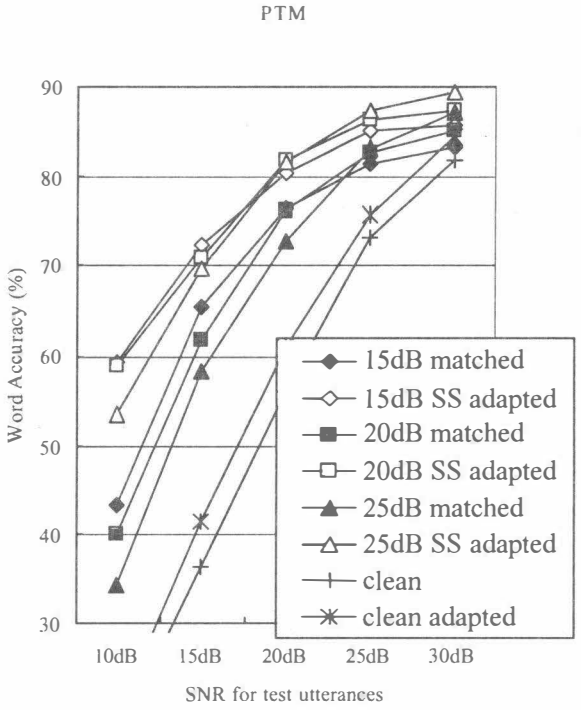

Figure 4: Word Accuracy in Different SNR Conditions for Noise Matched PTM and PTM Speaker and Noise Adapted PTM

database with spectral subtraction. This HMM sufficient statistics preparation is carried out off-line.

Second, numbers of selected speakers from the JNAS training data are 20 speakers for the monophone models and 40 speakers for PTM, according the previous report [8]. Speaker adapted HMM acoustic models are constructed from the HMM sufficient statistics from the selected speakers. This part can be carried out on-line.

The average word accuracy rates of 46 test speakers for the 20k dictation task are shown in Figure 2 for the monophone models, and Figure 3 for PTM. In Figure 2 and 3, "matched" indicates noise matched models, "SS matched" indicates noise matched models with spectral subtraction, "adapted" indicates speaker adapted models, and "SS adapted" indicates speaker adapted models with spectral subtraction.

The proposed speaker and noise adapted acoustic models consistently show 3 or 4 % better word accuracy rates than those of the matched models in every SNR.

4. ROBUSTNESS IN DIFFERENT SNR CONDITIONS

SNR usually fluctuates in noisy environments. The robustness in different SNR conditions is required even for adapted acoustic models. The same 20k vocabulary continuous speech recognition experiments are carried out in the different SNR conditions for the adapted models and the matched models with spectral subtraction. Experiment results are shown in Figure 4 for PTM. For example, "20dB matched" indicates 20dB noise matched PTM, and "20dB SS adapted" indicates 20dB noise matched speaker adapted PTM with spectral subtraction. Monophone shows the same tendency as PTM.

The noise and speaker adapted models and the noise matched models for PTM are relatively robust against the different SNR conditions. Especially, mismatched SNR speaker adapted PTMs with spectral subtraction (SS adapted) shows almost the same word accuracy as those of SNR matched adapted PTMs with spectral subtraction. The speaker adapted PTMs (white dots) with spectral subtraction show clearly higher word accuracy than those of SNR matched PTMs (black dots).

5. INITIAL MODEL FOR SUPERVISED MLLR

Supervised MLLR [6] is a popular speaker and noise adaptation algorithm. Of course, it requires a speaker to utter a lot of sentences correctly according to the specified

transcription. In the MLLR adaptation, an initial model is important to achieve the better performance [2][7]. We adopt the speaker adapted models with/without spectral subtraction and the matched models as an MLLR initial model. Supervised MLLR training data of 10 and 50 utterances for each test speaker are prepared from the JNAS database excluding the test utterances.

The same 20k vocabulary continuous speech recognition experiments are carried out in 15dB, 20dB and 25dB SNR conditions for MLLR supervised adaptation evaluation. Figure 5 shows the average word accuracy rates of 46 test speakers for monophone models, and Figure 6 shows the average word accuracy rates for PTM. Spectral subtraction effects are also investigated.

Comparing the word accuracy rates in the MLLR adaptation between the adapted initial models and the corresponding initial matched models, the initial adapted models always 2 to 4% better word accuracy in 10-utterance MLLR adaptation. These results show the usefulness of the adapted models for an MLLR initial model. In the case of 50-utterance MLLR adaptation, initial model effects become small.

6. CONCLUSION

An unsupervised speaker adaptation algorithm with spectral subtraction in noisy environments was investigated. The speaker adaptation algorithm based on HMM sufficient statistics from selected speakers was much improved with spectral subtraction.

We also showed the robustness of the adapted acoustic models in mismatched SNR conditions, and the usefulness for the supervised MLLR adaptation as an initial model.

References

[1] S.F.Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction", IEEE transaction on ASSP, ASSP-33, vol.27, pp.113-120, 1979

[2] Yuqing Gao, Mukund Padmanabhan, Michael Picheny, "Speaker Adaptation Based on Pre-Clustering Training Speakers", *Proceedings of EuroSpeech*, pp.2091-2094, 1999

[3] K.Itou, et al., "JNAS: Japanese Speech Corpus for Large Vocabulary Continuous Speech Recognition Research", *The Journal of the Acoustical Society of Japan (E)*, Vol.20, pp.199-206, 1999

[4] T.Kawahara, et al., "Free Software Toolkit for Japanese Large Vocabulary Continuous Speech Recognition", *Proceedings of ICSLP*, Ob(16)-V-07, pp.IV-476-479, 2000

[5] A.Lee, T.Kawahara, K.Takeda, K.Shikano, "A New Phonetic Tied Mixture Model for Efficient Decoding", *Proceedings of ICASSP*, pp.1269-1272, 2000

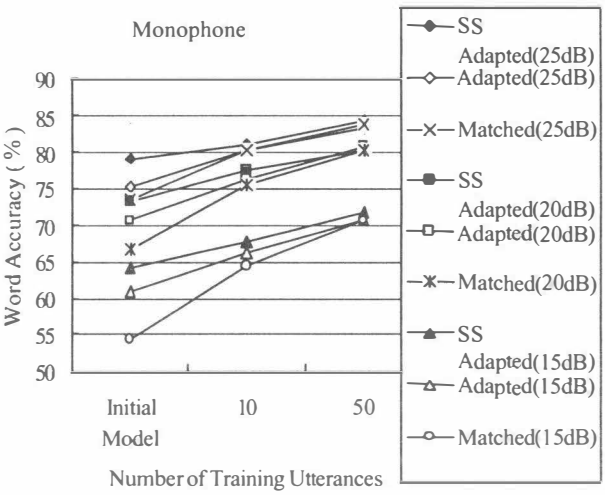

Figure 5: Initial Model Effects for MLLR Adaptation in Monophone

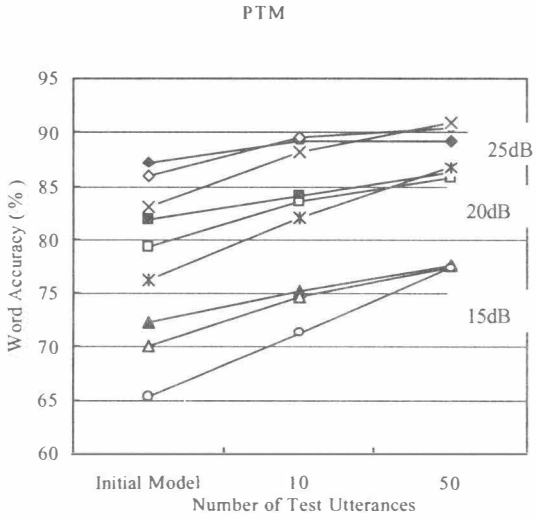

Figure 6: Initial Model Effects for MLLR Adaptation in PTM

[6] C.J.Leggetter, C.Woodland, "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models", *Computer Speech and Language*, Vol.9, pp.171-185, 1995

[7] M.Yamada, A.Baba, S.Yoshizawa, Y.Mera, A.Lee, H.Saruwatari, K.Shikano,"Unsupervised Noisy Environment Adaptation Algorithm Using MLLR and Speaker Selection", *Proceedings of EuroSpeech*, pp.869-872, 2001

[8] S.Yoshizawa, A.Baba, K.Matsunami, Y.Mera, M.Yamada, K.Shikano, "Unsupervised Speaker Adaptation Based on Sufficient HMM Statistics of Selected Speakers", *Proceedings of ICASSP*, pp.341-344, 2001