Unsupervised Noisy Environment Adaptation Algorithm
Using MLLR and Speaker Selection

Miichi Yamada*, Akira Baba**, Shinichi Yoshizawa**, Yuichiro Mera*,
Akinobu Lee*, Hiroshi Saruwatari*, Kiyohiro Shikano**

*Graduate School of Information Science, Nara Institute of Science and Technology, Japan
**Laboratories of Image Information Science and Technology, Japan
shikano@is.aist-nara.ac.jp

Abstract
An unsupervised acoustic model adaptation algorithm using
MLLR and speaker selection for noisy environments is
proposed. The proposed algorithm requires only one arbitrary
utterance and environmental noise data. The adaptation
procedure is composed of the following four steps. (1)
Speaker selection from a large number of database speakers is
carried out using GMM speaker models based on one
arbitrary utterance. (2) Initial speaker adapted HMM acoustic
models are calculated from the HMM sufficient statistics of
the selected speakers, where the sufficient HMM statistics are
pre-calculated and stored. (3) A small subset of the clean
speech database from the selected speakers and the
environment noise data are superimposed. (4) MLLR
adaptation is carried out using the noise-superimposed speech
database from the selected speakers. The proposed
algorithm is evaluated in a 20k vocabulary
dictation task for newspaper in noisy environments. We attain
85.7% word correct rate in 25dB SNR, which is slightly better
than the matched model by the E-M training using noise
superimposed whole speech database. The proposed
algorithm is also 7% better than the HMM composition
algorithm.

1. Introduction
Large vocabulary continuous speech recognition systems in
real environments require a speaker adaptation and
environment noise adaptation method [8]. There exist various
kinds of environment noises. It is almost impossible to collect
two kinds of environment noise data beforehand. Usual speaker
adaptation or environment noise adaptation algorithms usually
require for a user to speak several ten sentence utterances
before the speech recognition starts.

We proposed an unsupervised environment noise adaptation
algorithm, which requires only one arbitrary utterance and
several ten seconds of noise recording data. This proposed
environment noise adaptation algorithm has the following
features,
(1) Speaker selection from speech database based on speaker
GMMs (Gaussian mixture models) and one arbitrary
speech utterance.
(2) Initial speaker adapted HMM models from the selected
speakers using their HMM sufficient statistics for clean
(noiseless) speech database of each speaker [9].
(3) Superimposing the recorded noise data on a small subset
of clean speech database from the selected speakers.
(4) Supervised MLLR [6] environment noise adaptation
using the noise superimposed speech database and its
transcription.

The proposed environment noise adaptation algorithm
evaluation is also carried out in the large vocabulary
newspaper dictation task [4]. Evaluation database is JNAS
(Japanese Newspaper Article Speech database)[3], where each
of 306 speakers utters 50 phonetic balanced sentences and 100
newspaper articles. We recorded 15dB to 25dB SNR office
noise. We also compare the proposed algorithm with the E-M
trained speaker-independent model (matched model) based on
the noise-superimposed whole JNAS database, the supervised
MLLR [6], and the HMM composition algorithm [1][7]. The
proposed algorithm shows slightly better word recognition
rates than those of the matched model. The proposed
algorithm is also much better than the HMM composition
algorithm in word recognition rates.

2. Unsupervised Environment Noise Adaptation
Algorithm
The procedure for the proposed unsupervised environment
noise adaptation algorithm is shown in Figure 1. This
algorithm requires only one arbitrary utterance and a few ten
seconds of noise data. JNAS speech database [3] from 306
speakers are adopted as algorithm implementation and
evaluation. The adaptation procedure is composed of the
following four steps.
(Step 1) Speaker Selection [9]
Speaker selection from 305 JNAS database speakers,
xcluding a test speaker, is performed based on one arbitrary
utterance and 305 GMM speaker models. Each GMM speaker
model with one-state 64 Gaussian mixtures is beforehand
trained using 140 sentence utterances. The speaker selection is
done according to the GMM likelihood values for the one
arbitrary utterance.
(Step 2) Initial Speaker Adapted HMM Acoustic Models [9]
According to the speaker selection result, initial speaker
adapted HMM acoustic models are generated from the
sufficient HMM statistics of the selected speakers. The HMM
sufficient statistics for each speaker, which include average,
variance and E-M counts for each Gaussian mixture, are also
calculated using speech database and speaker-independent
HMM acoustic models beforehand.
(Step 3) Noise Superimposed Speech Database Generation
A small subset of JNAS clean speech database from the
selected speakers in Step 1 and a few ten seconds of the noise
data are superimposed.
(Step 4) Supervised MLLR Adaptation [6]
The supervised MLLR adaptation algorithm is carried out
using the noise superimposed speech database from the
selected speakers and its phoneme transcription.
3. Evaluation Experiments in Large Vocabulary Continuous Speech Recognition

The proposed environment noise adaptation algorithm is evaluated with a large vocabulary continuous speech recognition task. We adopt two types of HMM phoneme models, monophone models and PTM (phonetic tied mixture models) [5].

3.1 Evaluation Task and Conditions

The evaluation task is the JNAS newspaper dictation task with the 20k vocabulary size [4][5]. The baseline phoneme models are trained from the JNAS speech database [3], which includes 306 speakers and 45,000 sentence utterances in all. The test set contains 46 speakers from JNAS. Each test speaker utters 4 or 5 sentence utterance. Totally 200 sentence utterances are used as a test set, according to the IPA '99 test set [4]. We also adopt the decoder JULIUS and the language model from the JNAS dictation project. The experiment conditions are summarized in Table 1.

Table 1: Experimental Conditions

| Number of Speakers in JNAS database | 306 speakers (153 male speakers and 153 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 ms second hamming window, 10 ms second frame shift, CMN based on a sentence utterance, 12 MFCC, 12 delta-MFCC, and delta-power |
| Number of Selected Speakers for MLLR | 20 speakers |
| Noise Data | Office environment (180 seconds) |

The JNAS database has 150 sentence utterances for each speaker. The speaker GMM models and the sufficient HMM statistics are calculated from 140 sentence utterances for each speaker. The other 10 sentence utterances are used as adaptation utterances. The noise data of 180 seconds is recorded in the office environment. The other 10 sentence utterances are superimposed with the noise data according to the specified SNR level. One arbitrary sentence utterance for each of 46 test speakers, which is a part of the other 10 sentence utterances, is also superimposed with noise data, and used for the unsupervised adaptation.

3.2 Evaluation Experiments

In the first speaker selection step, the likelihood values from the speaker GMM models are calculated from only the speech part frames of the arbitrary sentence utterance by ignoring the low power frames. According to the likelihood values, 20 speakers for the monophone models and 40 speakers for PTM are selected from 306 speakers excluding a test speaker [9].

In the second step [9], the initial speaker adapted models are calculated from the sufficient HMM statistics of the selected speakers as a speaker adapted model for a clean (noiseless) input.

In the third step, the clean speech database from the nearest 20 selected speakers is superimposed with the office noise data. The SNR levels of the superimposed noise data are controlled at 15dB, 20dB, and 25dB SNR levels.

In the fourth step, the supervised MLLR adaptation [6] is carried out using the noise superimposed speech data from the selected speakers together with the corresponding phoneme transcription. The variances and averages are estimated by more than twice MLLR E-M iterations, while speaker adaptation usually estimates only the averages by one MLLR E-M iteration. The numbers of MLLR adaptation sentence utterances are 20 utterances (one utterance for each selected speaker), 60 utterances (three utterances for each), 100 utterances (five utterances for each), and 200 utterances (ten utterances for each).

The above environment noise and speaker adapted phoneme models are evaluated in the 20k dictation task described in 3.1 for the 46 test speakers' sentence utterances. The average word recognition rates (word correct rate) of the 46 test speakers are shown in Table 2 and Figure 1.

As for the monophone model in 25dB SNR, the word recognition rates of the clean speaker independent model and the clean speaker adapted model are 64.3% and 70.0% respectively. The proposed noise adaptation algorithm improves the word recognition rate to 75.4% (20 utterances), and 78.4% (200 utterances). The word recognition rate of the matched model, which is an E-M trained monophone model from noise superimposed whole JNAS database, is 76.1%.

As for the PTM in 25dB SNR, the proposed noise adaptation algorithm improves from 76.0% of the clean...
The proposed noise adaptation algorithm is compared with the Noise Adapted Monophone 53.0% most popular noise adaptation algorithm [1][7]. Here a noise MLLR HMM Composition 43.1%: Monophone other adaptation algorithms. The HMM composition between Clean Speaker Adapted Model 36.6%: Monophone, Clean Speaker Independent 31.5%: Monophone, Noise Adapted PTM 63.3%: PTM Clean Speaker Adapted Model 39.7%: PTM, Noise Adapted Monophone 53.0%: Monophone Noise Adapted PTM 63.3%: PTM Clean Speaker Adapted Model 43.1%: Monophone, Noise Adapted PTM 84.9%: PTM Clean Speaker Independent 48.3%: Monophone, Noise Adapted PTM 84.9%: PTM Clean Speaker Independent 54.5%: Monophone, Noise Adapted PTM 84.9%: PTM Clean Speaker Adapted Model 66.4%: Monophone, Noise Adapted PTM 84.9%: PTM Clean Speaker Adapted Model 68.9%: Monophone, Noise Adapted PTM 84.9%: PTM Clean Speaker Independent 70.5%: Monophone, Noise Adapted PTM 84.9%: PTM Clean Speaker Adapted Model 70.5%: Monophone, Noise Adapted PTM 84.9%: PTM Clean Speaker Independent 77.0%: Monophone, Noise Adapted PTM 84.9%: PTM Clean Speaker Adapted Model 77.0%: Monophone, Noise Adapted PTM 84.9%: PTM Clean Speaker Independent 80.3%: Monophone, Noise Adapted PTM 84.9%: PTM Clean Speaker Adapted Model 80.3%: Monophone, Noise Adapted PTM 84.9%: PTM Clean Speaker Independent 83.2%: Monophone, Noise Adapted PTM 84.9%: PTM Clean Speaker Adapted Model 83.2%: Monophone, Noise Adapted PTM 84.9%: PTM Clean Speaker Independent 91.2%: Monophone, Noise Adapted PTM 84.9%: PTM Clean Speaker Adapted Model 91.2%: Monophone, Noise Adapted PTM 84.9%: PTM

3.3 Comparison with HMM Composition and Supervised MLLR

The proposed noise adaptation algorithm is compared with other adaptation algorithms. The HMM composition between speaker-independent models and a noise HMM model is the most popular noise adaptation algorithm [1][7]. Here a noise HMM is represented as one state HMM. The word recognition rates of the HMM composition are also included in Table 2. For example, the HMM composition word recognition rate of PTM at 25 dB shows 79.7%, while the proposed noise adaptation algorithm shows the word recognition rate of 86.6%. Our proposed noise adaptation algorithm recognition performance is much better than that of the HMM composition. The computation amount for adaptation of the proposed algorithms with 20 sentence utterance database is almost the same as that of the HMM composition algorithm.

Next, supervised MLLR [6] is also a popular algorithm. However, it requires a speaker to utter a lot of sentences correctly according to the specified transcription. The word recognition rates of the supervised MLLR at 20 dB SNR are shown in Table 3. Our proposed noise adaptation algorithm shows almost the same word recognition rate as the supervised MLLR with 10 sentence utterances.

Table 3: Comparison of word correct rates between supervised MLLR and unsupervised proposed adaptation algorithm at 20 dB SNR noise condition

<table>
<thead>
<tr>
<th>Training Type</th>
<th>Adaptation Algorithm</th>
<th>Number of Utterances</th>
<th>Word Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monophone</td>
<td>Supervised MLLR</td>
<td>10</td>
<td>72.0%</td>
</tr>
<tr>
<td></td>
<td>Unsupervised Proposed</td>
<td>1</td>
<td>68.9%</td>
</tr>
<tr>
<td></td>
<td>(Noise only) HMM Composition</td>
<td>0</td>
<td>56.2%</td>
</tr>
<tr>
<td>PTM</td>
<td>Supervised MLLR</td>
<td>10</td>
<td>80.3%</td>
</tr>
<tr>
<td></td>
<td>Unsupervised Proposed</td>
<td>1</td>
<td>79.2%</td>
</tr>
<tr>
<td></td>
<td>(Noise only) HMM Composition</td>
<td>0</td>
<td>68.9%</td>
</tr>
</tbody>
</table>

3.4 Improvement Effects of Initial HMM Models by HMM Noise Composition Algorithm

MLLR adaptation algorithms are highly dependent on the initial models [2][8]. We use the clean speaker adapted HMM models for noise adaptation. We try to improve the clean initial HMM models by introducing the HMM composition algorithm. To generate better initial HMM models, the speaker adapted model and the one-state noise HMM model are composed by the HMM composition algorithm [1][7]. To reduce the mismatch of CMN operation, average CMN coefficients are used.

These initial model effects are evaluated in the same dictation task, as shown in Table 4. We attain slightly better word recognition rates especially at the low 20dB and 15dB SNR levels.

Table 4: Initial model effect by HMM noise composition algorithm

<table>
<thead>
<tr>
<th>SNR</th>
<th>25dB</th>
<th>20dB</th>
<th>15dB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monophone</td>
<td>78.4%</td>
<td>68.8%</td>
<td>56.3%</td>
</tr>
<tr>
<td>+ HMM noise composition</td>
<td>78.2%</td>
<td>70.5%</td>
<td>58.0%</td>
</tr>
<tr>
<td>PTM</td>
<td>86.6%</td>
<td>78.5%</td>
<td>68.4%</td>
</tr>
<tr>
<td>+ HMM noise composition</td>
<td>86.7%</td>
<td>79.3%</td>
<td>69.4%</td>
</tr>
</tbody>
</table>
4. Conclusion

We proposed an unsupervised acoustic model adaptation algorithm using MLLR and speaker selection for noisy environments. The proposed noise adaptation algorithm was evaluated in the 20k JNAS dictation task in noisy environment. The experiment results for PTM under 25dB SNR are summarized in Figure 2. In 25dB SNR level, we attained 85.7% word correct rate, which is slightly better than the matched model by the E-M training with noise superimposed JNAS whole speech data. The proposed algorithm attained much better results than the HMM composition algorithm.

Acknowledgements

This research is partially supported by NEDO (New Energy and Industrial Technology Department Organization) and CREST (Core Research for Evolutional Science and Technology), JST.

References