

# UNSUPERVISED SPEAKER ADAPTATION BASED ON HMM SUFFICIENT STATISTICS IN VARIOUS NOISY ENVIRONMENTS

Shingo Yamade<sup>\*1</sup>, Akinobu Lee<sup>\*1</sup>, Hiroshi Saruwatari<sup>\*1</sup>, Kiyohiro Shikano<sup>\*1</sup>

<sup>\*1</sup>Graduate School of Information Science, Nara Institute of Science and Technology

8916-5 Takayama-cho, Ikoma-shi, Nara, 630-0101, JAPAN

E-mail: shing-y, ri, sawatari, shikano@is.aist-nara.ac.jp

## ABSTRACT

Noise and speaker adaptation techniques are essential to realize robust speech recognition in noisy environments.

In this paper, first, a noise robust speech recognition algorithm is implemented by superimposing a small quantity of noise data on spectral subtracted input speech. According to the recognition experiments, 30dB SNR noise superimposition on input speech after spectral subtraction increases the robustness against different noises significantly.

Next, we apply this noise robust speech recognition to the unsupervised speaker adaptation algorithm based on HMM sufficient statistics in different noise environments. The HMM sufficient statistics for each speaker are calculated from 25dB SNR office noise added speech database beforehand.

We evaluate successfully our proposed unsupervised speaker adaptation algorithm in noisy environments with 20k dictation task using 11 kinds of different noises, including office, car, exhibition, and crowd noises.

## 1. INTRODUCTION

Unsupervised speaker adaptation is required in real noisy environments to attain better accuracy. Existence of various noises in the real world makes unsupervised speaker adaptation difficult. We have been researched the unsupervised speaker adaptation based on HMM sufficient statistics [7,8,9]. Our proposed unsupervised speaker adaptation algorithm [9] requires noise matched acoustic models to prepare HMM sufficient statistics for each speaker in training speech database. The noise matched acoustic models require a large amount of computation for training from noise added speech database. Speech spectra after the spectral subtraction [1,9] still retain the property of the original noise, although SNR is much improved.

We propose to superimpose a small quantity of noise on input speech to cancel the residual noise after spectral subtraction, in order to improve the robustness against different noises. The proposed noise robust speech recognition by spectral subtraction and noise superimposition is successfully evaluated with 20k dictation task [3,4] using different types of noises, office, car, exhibition, and crowd noises [2].

Our proposed unsupervised speaker adaptation algorithm [7] is also combined with the noise robust

speech recognition, and successfully applied to the 20k dictation task in different noise environments. This unsupervised speaker adaptation can work very quickly without calculating noise matched acoustic models.

## 2. NOISE ROBUST SPEECH RECOGNITION

Large vocabulary continuous speech recognition in various noisy environments requires noise adaptation or noise robust speech recognition. There exist huge numbers of different noises. It is almost impossible to collect all kinds of environment noise data beforehand. We propose a noise robust speech recognition algorithm based on spectral subtraction and noise superimposition, and evaluate its recognition accuracy in different noise environments.

### 2.1. Spectral subtraction

Spectral subtraction is a technique to reduce noise from noisy speech by subtracting noise spectrum [1]. Spectral subtraction offers a computationally efficient technique for reducing noise. Assume that a speech signal  $s(n)$  has been degraded by 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 beginning. To avoid negative speech spectrum, flooring operation is introduced as

$$|\hat{S}(k)| = |X(k)| A, \\ \text{when } |\hat{S}(k)|^2 = |X(k)|^2 - a E|V(k)|^2 < 0.$$

$A$  is a flooring parameter.  $a = 2.0$  and  $A = 0.5$  are adopted according to preliminary experiments [9].

### 2.2. Noise superimposition

Spectral subtraction improves SNR of input speech about 10dB. There still exists the spectral feature of the original noise spectrum usually. To recognize input speech after spectral subtraction accurately, we need the noise matched acoustic models [9]. Instead of the use of noise matched acoustic models, we superimpose a small quantity of noise on the spectral subtracted input speech. We carry out

recognition experiments with 20k dictation task to check the quantity of noise. Table 1 shows the word accuracy by PTM, phonetic tied mixture model [5], for clean input, and 30dB SNR office noise superimposition using the noise SNR matched PTM acoustic models, 30dB, 25dB and 20dB. The noise superimposition of 30dB SNR affects the word accuracy very little comparing with the clean input by the clean speaker independent PTM. The 25dB SNR matched PTM shows almost the same word accuracy as the 30dB matched PTM, and seems to be more noise robust. The noise superimposition of 25dB SNR shows 86.3% word accuracy, which is clearly worse than one by 30dB SNR noise superimposition. Hereafter, we adopt 30dB office noise superimposition after spectral subtraction to implement noise robust speech recognition, and the 25dB SNR matched PTM of office noise is adopted. This noise robust speech recognition procedure is shown in Fig. 1.

Table 1: Word accuracy for office noise superimposition.

Noise superimposition	clean	30dB SNR office noise		
PTM acoustic model	clean	30dB	25dB	20dB
Speaker independent	91.1	89.9	89.3	86.8

### 2.3. Noise robust speech recognition experiment in large vocabulary continuous speech recognition

In this section, we evaluate our proposed noise robust speech recognition algorithm using four different kinds of noises, office, car, exhibition, and crowd noises from JEIDA noise database [2]. Evaluation task is the 20k dictation task [3,4,5].

#### 2.3.1. Evaluation task and conditions

The evaluation task is the JNAS newspaper dictation task with 20k vocabulary size [3,4,5]. The baseline speaker-independent acoustic models are trained from 260 training speakers' data in the JNAS speech database [3]. PTM, phonetic tied mixture models, [5] are used. The PTM training speech database includes 260 speakers (39,000 sentence utterances in total). The test set consists of another 46 speakers from JNAS. Each test speaker utters 4 or 5 newspaper article sentences (200 test sentence utterances in total), according to the IPA'99 test set [3,4]. We adopt the decoder JULIUS and the language model from the IPA dictation project [4].

The experiment conditions are summarized in Table 2. Noisy speech data are prepared by adding four different types of noises on the JNAS clean speech database according to SNR levels.

#### 2.3.2. Noise robust speech recognition evaluation

The proposed noise robust speech recognition algorithm is evaluated with the 20k dictation task in three different types of 20dB SNR noises, automobile cabin (car),

exhibition hall and crowded street. We compare our proposed noise robust speech recognition (4) with (1) noise matched model, (2) noise matched model with spectral subtraction, and (3) office noise matched model.

Table 2: Experiment conditions

Task	20k newspaper dictation task
Number of training speakers in JNAS	260 speakers (130 male speakers and 130 female speakers)
Number of testing speakers in JNAS	46 speakers ( 23 male speakers and 23 female speakers)
Speech analysis and feature extraction	25 ms hamming window, 10 ms frame shift, CMN, 12 MFCC, 12 delta-MFCC, and delta-power
Acoustic model	25dB SNR matched PTM (phonetic tied mixture model)
Noises ( four different types)	office room, automobile cabin, exhibition hall, and crowded street
Spectral subtraction	$a = 2.0$ and $A=0.5$
Noise superimposition	30dB office noise

These evaluation results are summarized in Fig.2. The experimental conditions are described in Table 3. As for office noise, the experimental condition 3 is noise matched one. The difference between condition 4 and 2 in office noise shows the small degradation by noise superimposition. As for car, exhibition and crowd noises, condition 4 with 30dB office noise superimposition shows the robustness against different noise, comparing with condition 3 without noise superimposition. The average word accuracy of car, exhibition and crown noise is also included in Fig.2. The noise superimposition (4) improves the word accuracy from 81.6% (3) to 84.6% on the average of three different noises. The average difference between the noise matched model (2) and the noise robust recognition (4) is only 2.0%. The comparison with PMC (parallel model combination) was studied in our paper [6]. The robust speech recognition works much better than PMC.

Word accuracy for 15dB SNR noises is also investigated to confirm the robustness against lower SNR. The average word accuracy rates for car, exhibition and

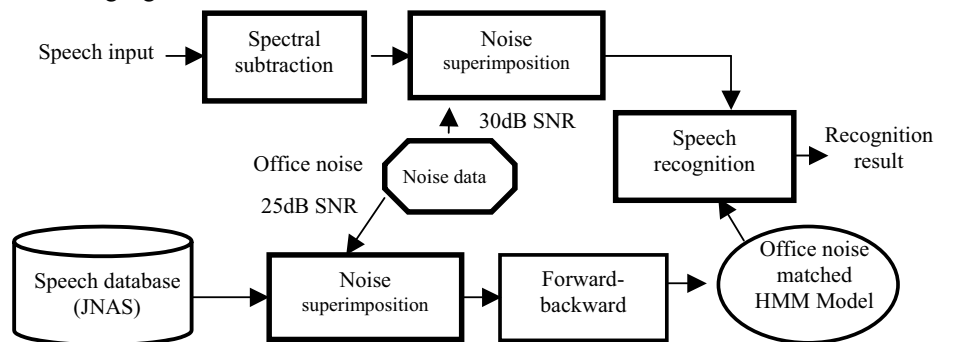


Figure 1: Noise robust speech recognition by spectral subtraction and noise superimposition

crowd noises are 81.4% for noise matched models, 76.4% for office noise models, and 79.1% for the noise robust speech recognition, respectively.

Table 3: Experimental conditions for noise robust speech recognition and speaker adaptation.

Condition	1	2	3	4	5	6
Noise matched acoustic model	o	o			o	
Office noise acoustic model			o	o		o
Spectral subtraction		o	o	o	o	o
<b>Office noise superimposition</b>				o		o
Speaker independent	o	o	o	o		
Speaker adaptation					o	o

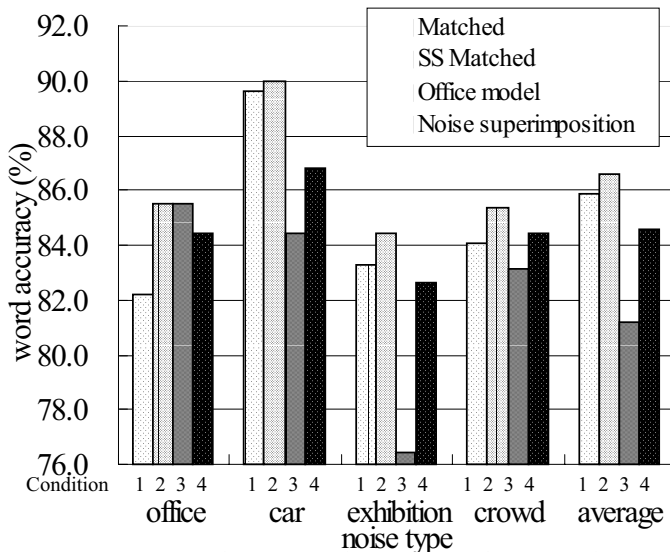


Figure 2: Evaluation result for noise robust speech recognition algorithm

### 3. UNSUPERVISED SPEAKER ADAPTATION IN DIFFERENT NOISE ENVIRONMENTS

The procedure of our unsupervised speaker adaptation algorithm based on noise robust speech recognition is shown in Fig. 3. This algorithm requires only one arbitrary utterance. The previous algorithm [9] requires noise matched acoustic models to calculate HMM sufficient statistics for each speaker. In the use of the noise robust speech recognition, the HMM sufficient statistics for each speaker are calculated from the 25dB office noise matched model beforehand. Online quick adaptation based on one arbitrary utterance is possible. JNAS speech database [3] from 306 speakers are adopted as the algorithm implementation and evaluation, as well as in Section 2.3.

#### 3.1. Speaker adaptation algorithm

Unsupervised speaker adaptation procedure [7] is modified based on the robust speech recognition, which consists of the following five steps, as shown in Fig.3.

(Step 1) 25dB SNR office noise matched acoustic model is trained from noise added speech database.

(Step 2) HMM sufficient statistics for each speaker, which include average, variance and E-M count of each Gaussian distribution, are calculated from the office noise matched acoustic model, and stored.

(Step 3) According to one arbitrary utterance, speakers close to a test speaker are selected using speaker GMMs.

(Step 4) Speaker adapted acoustic models are constructed from HMM sufficient statistics of the selected speakers.

(Step 5) Spectral subtraction and 30dB SNR office noise superimposition on input speech is carried out for the robust speech recognition with the speaker adapted acoustic model.

#### 3.2. Speaker adaptation experiment in different noise environments

The proposed speaker adaptation algorithm in different noise environments is evaluated with the 20k dictation task.

First, HMM sufficient statistics for each training speaker are calculated from the 25dB SNR office noise matched speaker-independent PTM using the 25dB noise added training JNAS database. This HMM sufficient statistics preparation is carried out off-line.

Second, number of selected speakers from the JNAS training set is 40 for PTM, according to the previous report [7]. Speaker adapted HMM acoustic models are quickly constructed from the HMM sufficient statistics of the selected speakers online.

Third, spectral subtraction and 30dB office noise superimposition are carried out on input noisy speech with different three types of noises, car, exhibition and crowd.

The recognition results are shown in Fig.4, in comparison with adaptation ones by noise matched PTM and speaker-independent results in Section 2.3. The experimental conditions are already described in Table 2 and 3. In Fig.4, condition 5 indicates speaker adaptation results by the noise matched PTM, and condition 6 indicates speaker adaptation results with the noise robust speech recognition.

The average word accuracy of car, exhibition and crowd noises is also included in Fig.4. The speaker adaptation attains 2.9% improvement from 84.6% to 87.5% in the average of three different noises. The degradation from the noise matched PTM is only 0.7%.

We further evaluate our proposed unsupervised speaker adaptation algorithm with the noise robust speech recognition for another eight kinds of noises, which include a railway station, a public park, crossing road, a shopping mall, ticket vending machine, an airport, airplane inside, and a poster hall. The word accuracy for 11 kinds of noises including above eight kinds of noises, car, exhibition and crowd were investigated. As for our noise robust speech recognition, the speaker independent word accuracy is 81.9%, and the speaker adapted word accuracy is 85.1%. As for the noise matched PTM, the

speaker independent word accuracy is 84.0%, and the speaker adapted word accuracy is 86.4%. These small degradation rates from noise matched models show the effectiveness of our proposed online unsupervised speaker adaptation algorithm in noisy environments.

### 3.3 Comparison with supervised MLLR

The comparison with supervised MLLR using three kinds of noises, car, exhibition and crowd, is carried out using clean speaker independent PTM as an initial model. In MLLR, averages and variances are adapted, and the number of iterations is 3. The numbers of training sentence utterances are 10 and 50. The average word accuracy rates of three noises are 87.2% for 10 utterances and 89.6% for 50 utterances. Our proposed unsupervised speaker adaptation by the robust speech recognition attains 88.2%, which is better than supervised 10-utterance MLLR.

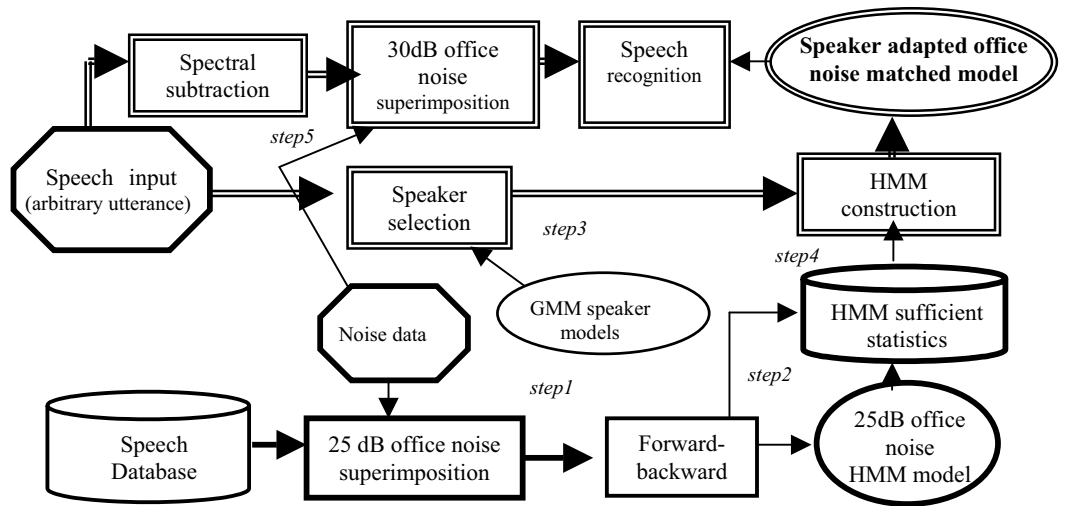


Figure 3: Unsupervised speaker adaptation based on noise robust speech recognition

### 4. CONCLUSION

A noise robust speech recognition algorithm based on spectral subtraction and noise superimposition was proposed and evaluated. We also applied the noise robust speech recognition to unsupervised speaker adaptation, successfully.

### References

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

[2] JIIDA Noise database, <http://it.jeita.or.jp/jhistory/committee/humanmed/speech/noisedbj.html>

[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", *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", *ICASSP*, pp.1269-1272, 2000

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

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

[8] S.Yoshizawa, A.Baba, K.Matsunami, Y.Mera, M.Yamada, A.Lee, K.Shikano, "Evaluation on Unsupervised Speaker Adaptation Based on Sufficient HMM Statistics of Selected Speakers", *EuroSpeech*, pp.1219-1222, September 2001

[9] Shingo Yamade, Kanako Matsunami, Akira Baba, Akinobu Lee, Hiroshi Saruwatari, Kiyohiro Shikano, "Spectral Subtraction in Noisy Environments Applied to Speaker Adaptation Based on HMM Sufficient Statistics", *ICSLP2002*, pp.1045-1048, September 2002

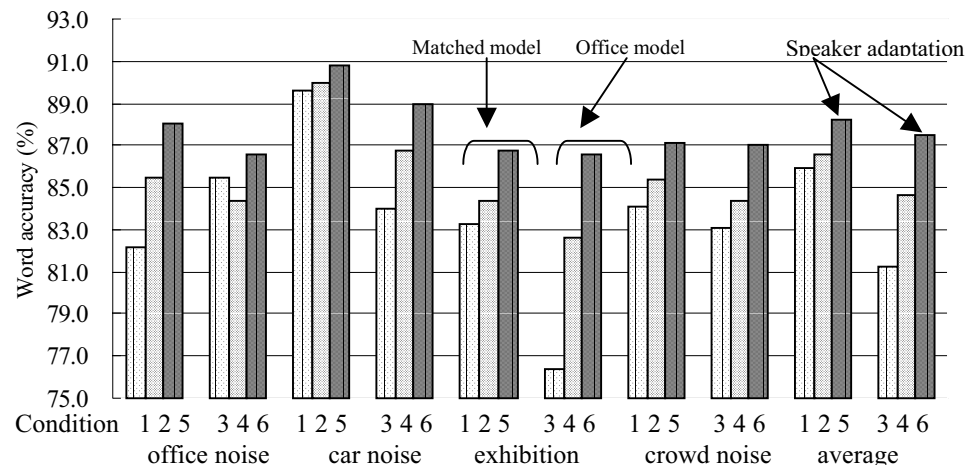


Figure 4: Word accuracy improvement by unsupervised speaker adaptation based on noise robust speech recognition in 20k dictation task