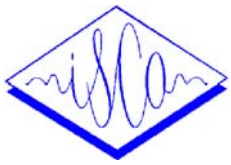

INTERSPEECH 2004 - ICSLP
8th International Conference on Spoken Language Processing
ICC Jeju, Jeju Island, Korea
October 4-8, 2004

# Robust Speech Recognition based on HMM Composition and Modified Wiener Filter

*KOBASHIKAWA Satoshi, SAKAUCHI Sumitaka, YAMAGUCHI Yoshikazu, and TAKAHASHI Satoshi*

NTT Cyber Space Laboratories,
NTT Corporation
{kobashikawa.satoshi, sakauchi.sumitaka, yamaguchi.yoshikazu, takahashi.satoshi}@lab.ntt.co.jp

## Abstract

This paper combines the HMM composition method with a highly efficient noise reduction method to create a robust speech recognition technique for additive noise environments. Speech recorded by hands-free microphones in the real world suffer from 1) low Speech/Noise [S/N] and 2) changes in S/N. In particular, S/N varies with the speaker and from utterance to utterance even in a same noise environment. To deal with the low S/N, the proposed technique uses the modified Wiener filter (WF) method for noise reduction and so keeps S/N higher than is possible with spectral subtraction (SS), as well as minimizing speech distortion. To compensate the remaining additive noise, the proposed technique uses the HMM composition method with clean speech models and a noise model trained by the remaining noise. To offset the rapid changes in S/N where S/N may not be known, HMMs composed under various S/N conditions are run in parallel to obtain better recognition results; rapid response is achieved since adaptation to handle speech distortion is not necessary. The new technique shows a reduction in average recognition error of 21.6% under various noise conditions compared to using the basic HMM composition method.

## 1. Introduction

Research into practical automatic speech recognition systems is growing quickly, due to improvements in computer performance. Hands-free speech recognition techniques need to be developed to lessen the user's burden. Hands-free environments have more problems than close-talking environments. One is multiplicative distortion introduced by the microphone specifications and space transfer characteristics. The other is the additive noise from the ambient environment. In this paper, we will discuss the latter issue and describe robust speech recognition techniques for environments with additive noise.

Two significant problems arise in hands-free environments regarding additive noise: 1) low S/N and 2) changes in S/N. As the distance between the speaker's mouth and the microphone increases, ambient noise increases and the S/N decreases. In addition, even if the ambient noise level fixed, the speech level picked up at the microphone depends on the loudness of the speaker's voice, the words uttered, and the position of the speaker in relation to the microphone; thus S/N changes often and widely.

The HMM composition method, i.e. noise and voice composition (NOVO) [1] and parallel model combination (PMC) [2], is well-known as an effective noise adaptation method that can improve speech recognition performance in noisy environments. However, recognition performance of noise adaptation methods like NOVO, is not sufficient if the S/N is low, because the speech features are buried in the noise. Methods to raise the S/N of the observed signals by using noise reduction such as the spectral subtraction (SS) method [3] and the Wiener filter (WF) method [4] can be used. These methods, however, are not able to remove noise completely, and they create new problems in that insufficient or excessive reduction processing leads to remaining noise or speech distortion, respectively.

This paper counters problem 1) by investigating a technique based on a combination of the HMM composition method and the highly effective modified Wiener filter as the noise reduction method. Techniques that combine the HMM composition method with noise reduction methods such as SS have been proposed, for example SS-PMC [5] or continuous spectral subtraction PMC (CSS-PMC) [6]. These conventional techniques include adaptation processing to handle the inevitable speech distortion caused by noise reduction, and have difficulty in handling rapid changes in the input signal. Here, we propose noise reduction by modified Wiener filter NOVO (NRWF-NOVO), a combination of the NOVO method [1] and the modified Wiener filter [7]. The latter realizes strong noise reduction with lower speech distortion, and the remaining noise is handled by the NOVO method.

To overcome problem 2), we propose the use of several acoustic models formed under various S/N conditions; speech recognition processing is carried out in parallel using these models and the output of the best performing model with the highest likelihood is selected.

The rest of paper is organized as follows. NRWF-NOVO is described in Section 2. Section 3 introduces experiments conducted to compare NRWF-NOVO to conventional techniques. Section 4 discusses the results. Our conclusion is drawn in Section 5.

## 2. NRWF-NOVO

### 2.1 Known S/N case

Figure 1 shows the framework of NRWF-NOVO for the known S/N case; this represents the core of NRWF-NOVO and so provides an assessment of its basic performance. NRWF-NOVO starts with a noise reduction section. To optimize the noise reduction process, noise reduction parameter was changed through our preliminary experiments. The remaining noise signals in the non-utterance regions are extracted and used in the noise adaptation section to yield noise adapted models with clean acoustic model. To ensure that the noise adapted acoustic model has the best-possible S/N condition, we control the composition processing from clean speech power level in the training data of a clean

acoustic model and the noise power level in the observed noise data.

This technique assumes that the remaining noise signals in non-utterance regions and the remaining noise in noise-superposed utterance regions are equal.

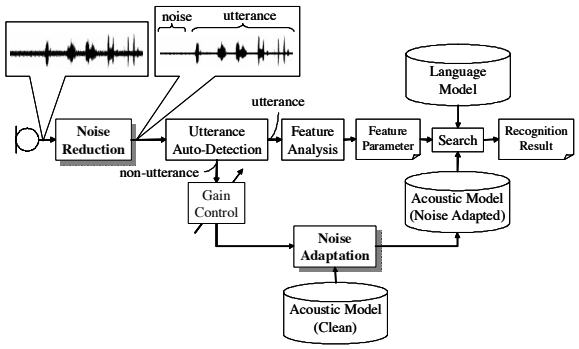


*Figure 1:* Framework of NRWF-NOVO in known S/N case.

### 2.1.1 *Noise Reduction section*

This section explains the modified noise reduction process which is based on an extension of the WF (Wiener Filter) method. We adopt the noise reduction algorithm of the audio teleconferencing system proposed by Sakauchi et al. [7]. In [7], $\hat{S}$ denotes the output of noise reduced signal following the WF method and is given by

$$\hat{S} = G \cdot |Z| e^{j\theta}, \tag{1}$$

where $Z$ denotes the untreated input signal.

The gain function, based on Wiener filter, $G$, is given by

$$G = \frac{E\left[|S|^2\right]}{E\left[|S|^2\right] + E\left[|N|^2\right]}, \tag{2}$$

where $E\left[|S|^2\right]$ and $E\left[|N|^2\right]$ denote the ensemble average of speech signals and ambient noise signals, respectively. Furthermore, the gain function, $G$, is smoothed in both the frequency-domain and the time-domain for minimizing speech distortion.

In [7], the noise reduction algorithm also minimizes speech distortion by adding a part of the untreated input signal, $Z$, to the output of noise reduced signal, $\hat{S}$ following the WF method. The signal after noise reduction, $\tilde{S}$, is given by

$$\tilde{S} = (1-\alpha)Z + \alpha\hat{S}. \tag{3}$$

Preliminary experiments targeting speech recognition indicated that the optimum value of $(1-\alpha)$ was 0.3 which is different from the optimum value for the audio teleconferencing system [7], and this value was used in subsequent experiments with NOVO method.

To assess the effectiveness of the modified WF method, we compared it to the SS method [3]. In the SS method, $|Z|^2$ is the untreated input power spectrum. $|\hat{N}|^2$ is estimated noise power spectrum and is held constant. The power spectrum, $|\tilde{S}|^2$, after noise reduction is given by

$$|\tilde{S}|^2 = \max\left\{|Z|^2 - \beta\left|\hat{N}\right|^2, f|Z|^2\right\}. \tag{4}$$

Preliminary experiments showed that the optimum overestimation factor, $\beta$, was 1.0, while the spectral flooring parameter, $f$, was set to 0.5; these values were used in subsequent experiments with NOVO method.

### 2.1.2 *Noise Adaptation section*

We explain here the noise adaptation section of the proposed technique. We use the NOVO method as described [1] for noise adaptation. In this work, the main parameter of the acoustic model is based on the cepstrum. Accordingly, $c_S$ is the cepstrum of clean speech, $c_N$ is the cepstrum of noise, and $\tilde{c}$ is given by

$$\tilde{c} = F^{-1}\left(\log\left[\exp\left\{F\left(c_S\right)\right\}\right] + k\log\left[\exp\left\{F\left(c_N\right)\right\}\right]\right). \tag{5}$$

$F$, $F^{-1}$ denote Fourier Transform and Inverse Fourier Transform, respectively. Gain $k$ is a coefficient dependent upon S/N; it is calculated from the speech power level in training data of clean acoustic model and the noise power level of observed noise data.

## 2.2 Unknown S/N case

Figure 2 shows the framework of NRWF-NOVO for the unknown S/N case. In a real environment, it is impossible to know the S/N before finishing an utterance. We offset this problem by using several acoustic models created under various S/N conditions; speech recognition processing is performed in parallel using these models. At first, we create sets of S/N condition acoustic models by changing gain $k$. Next, we carry out speech recognition processing in parallel using these noise adapted acoustic models. Finally, the recognition result is taken as the output of the models with the highest likelihood. In this paper, we use acoustic models with five values of S/N: 0 dB, 5 dB, 10 dB, 15 dB, and 20 dB.

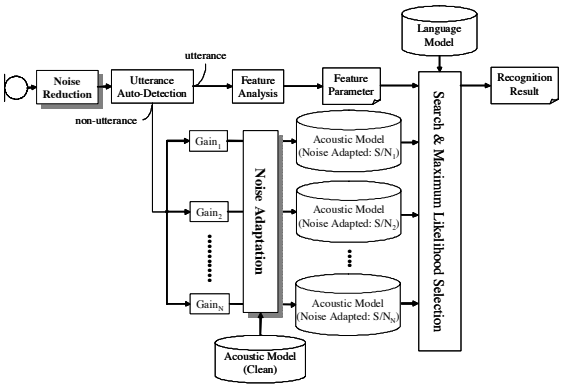


*Figure 2:* Framework of NRWF-NOVO in unknown S/N case.

## 3. Experiments

### 3.1. Parameters

Table 1 shows the speech analysis conditions, Table 2 shows the acoustic model (HMM) conditions used in the experiments, and Table 3 shows the evaluation task.

As shown in Table 3, we used artificial noisy speech data created by adding noise to the ATR 216 word set uttered by one male speaker. The PC fan noise was recorded using a PC's internal-microphone. The other noises were acquired from a domestic sound database [8]. When adding noise, S/N was set using the speech level defined as the average power level including pauses; the noise level was the average power level. The noise level was fixed, and we controlled the S/N by changing the gain level of speech. Default noise was PC fan; the additional noises all had the same power level as the default noise.

*Table 1:* Speech analysis conditions

| | |
|---|---|
| Sampling Rate | 16 kHz |
| Frame Width | 30 msec Hamming Window |
| Frame Shift | 10 msec |
| Feature Parameter | MFCC (12), ΔMFCC (12), Δ POW |

*Table 2:* Acoustic model conditions

| | |
|---|---|
| Training Data | ATR 503 sentences spoken by 1 male speaker |
| HMM | monophone continous mixture distribution |
| # of states | 3 |
| # of mixtures | 4 |
| # of phonemes | 30 |

*Table 3:* Evaluation Task

| | |
|---|---|
| test data | ATR 216 words spoken by 1 male speaker |
| task | speaker-dependent 1000-words grammar |
| noise type | PC fan / +sink, +cleaner, +ventilation fan |
| S/N | 5 dB, 10 dB, 15 dB |

### 3.2. Conventional method

To show effectiveness of NRWF-NOVO, we compared it to two conventional techniques. The first one was the NOVO method without the noise reduction section. The second one was the SS-NOVO combination; the noise reduction section implemented the SS method.

### 3.3. Experiments: known S/N case

Figure 3 shows recognition results of PC fan noise + speech in the known S/N case. In this figure, the values of S/N are the an average S/N values calculated using all evaluation data. That is, some words have different S/N from the average S/N. Ideally, we should use a compounded S/N matched-condition acoustic model for evaluating a word. Here, we prepared five different S/N level acoustic models (0 dB, 5 dB, 10 dB, 15 dB, and 20 dB) beforehand, and we used the closest matching S/N condition acoustic model for speech recognition.

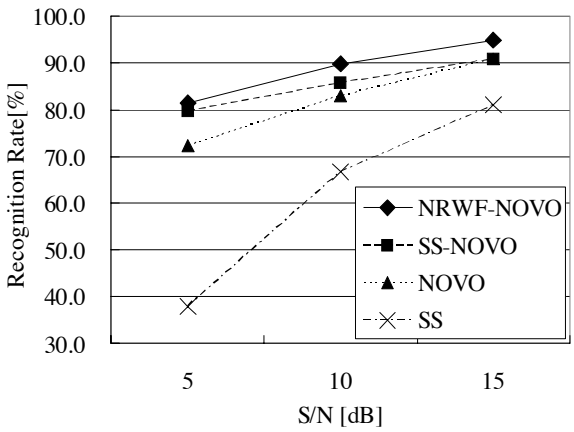


*Figure 3:* Recognition result: PC fan noise in known S/N case.

Figure 3 shows that while that recognition performance of all methods falls with S/N, and NOVO minimizes the drop off in performance. A conventional NOVO method the speech recognition performance is lower as S/N lowered. Of the 4 methods examined, NRWF-NOVO offers the best performance, regardless of the S/N.

### 3.4. Experiments: unknown S/N case

Figure 4 shows recognition result for PC fan noise + speech in the unknown S/N case. In the technique described the previous section, S/N was known beforehand, but in practice this is not possible and the accuracy of S/N estimation is problematic. Our solution is to create several acoustic models under various S/N conditions, and to carry out speech recognition processing in parallel using these models. Here we prepared acoustic models using five S/N levels: 0 dB, 5 dB, 10 dB, 15 dB, and 20 dB.

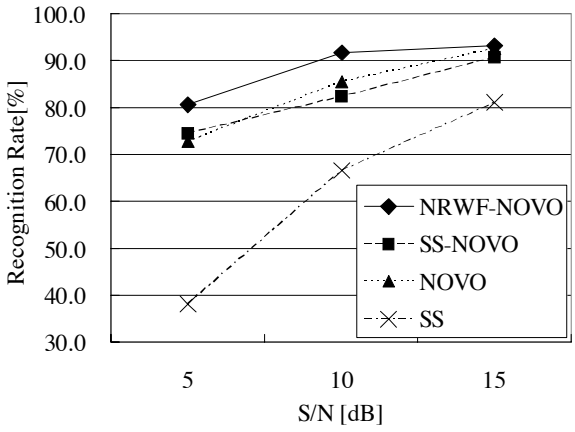


*Figure 4:* Recognition result: PC fan noise in unknown S/N case.

Figure 4 shows basically the same trends shown in Fig. 3. The main difference is the relatively weak performance of the SS-NOVO technique. This is because its noise reduction is imperfect and leads to speech distortion which causes errors in acoustic model selection. Once again, NRWF-NOVO offers the best speech recognition performance. It is especially good at low S/N values.

Table 4 shows the recognition rates of noisy speech with three different kinds of additional noise; i.e. PC fan noise with sink, cleaner, and ventilation fan (vfan). For example noise added sample, the S/N drops about 3 dB. Table 4 shows that NRWF-NOVO is equally superior in most of noisy situations examined.

*Table 4:* Recognition results with different noise combinations.

| Noise | NOVO | SS-NOVO | **NRWF-NOVO** |
|---|---|---|---|
| PC fan 15dB + sink | 87.5 | 88.4 | **92.6** |
| PC fan 10 dB + sink | 78.7 | 77.3 | **84.7** |
| PC fan 5 dB + sink | 56.9 | 58.3 | **63.4** |
| PC fan 15 dB + cleaner | 88.9 | 88.0 | **90.7** |
| PC fan 10 dB + cleaner | 80.6 | 75.9 | **80.1** |
| PC fan 5 dB + cleaner | 57.4 | 56.9 | **63.4** |
| PC fan 15 dB + vfan | 91.7 | 89.4 | **94.0** |
| PC fan 10 dB + vfan | 81.9 | 80.1 | **88.9** |
| PC fan 5 dB + vfan | 66.2 | 67.1 | **73.6** |

## 4. Discussion

Conventionally, the NOVO method is effective for noisy environment speech recognition. However, it involves the adaptation of an acoustic model of noise + speech, and this is a significant problem when the S/N is low. Addressing this problem, our proposed technique improves the S/N by using noise reduction methods for the front-end of speech recognition processing and offsets the remaining noise by using the NOVO method, which improves speech recognition performance at low S/N values.

In known S/N cases, the speech recognition performance of NOVO-based methods becomes falls as the S/N falls. However, since the proposed technique includes a noise reduction method like SS or WF as a front-end to speech recognition, it holds speech recognition performance relatively constant; this was confirmed by experiments. The results show the effectiveness of improving S/N by noise reduction processing.

In unknown S/N cases, the proposed method uses various S/N condition models in parallel for speech recognition processing and the output of the best model is selected. Combining SS with NOVO creates a performance penalty at high S/N values (compared to just NOVO). We believe this is because the error of acoustic model selection based on likelihood increases under the influence of speech distortion in the low power level speech region. Since SS offers only fixed noise reduction processing based on the average power level in the non-utterance regions, it tends to suffer from excessive signal distortion. To counter this, the proposed technique uses the modified WF method for noise reduction; it keeps the speech quality high since noise reduction processing is proportional to S/N frame by frame and untreated input signals are utilized.

## 5. Conclusions

The NOVO method is well known to improve the performance of speech recognition in noisy environments. However, its performance falls strongly as the S/N value falls. To counter this problem, we proposed NRWF-NOVO; it adds a noise reduction stage based on the modified Wiener filter (WF) in the front-end to speech recognition processing. Tests done in the known S/N case confirms the effectiveness of the proposed noise reduction method. To handle the case of unknown S/N, we proposed that speech recognition be carried using, in parallel, acoustic models created under different S/N conditions. The use of the modified WF technique keeps recognition performance high even though the S/N is unknown, unlike spectral subtraction. As a result, the proposed technique achieves a reduction in average recognition error of 21.6% compared to the NOVO method.

## 6. Acknowledgements

We are grateful to our project manager Mr. Hisashi Ohara of NTT Cyber Space Laboratories for giving us the opportunity to pursue this work. We also wish to thank the members of the Speech, Acoustics and Language Laboratory for their useful advice.

## 7. References

[1] F. Martin, K. Shikano, and Y. Minami, "Recognition of Noisy Speech by Composition of Hidden Markov Models," Proc. EUROSPEECH, pp.1031-1034, Sep. 1993.

[2] M. J. Gales, and S. J. Young, "Robust Continuous Speech Recognition using Parallel Model Combination", IEEE trans. on Speech and Audio Processing, Vol. 4, pp.352-359, 1996.

[3] S. F. Boll, "Suppression of Acoustic Noise in Speech using Spectral Subtraction," IEEE trans. on ASSP, vol. 27, no. 2, pp. 113-120, April, 1979.

[4] J. S. Lim and A. V. Oppenheim, "Enhancement and Bandwidth compression of noisy speech," Proc. IEEE, vol.67, no. 12, pp. 1586-1604, Dec. 1979.

[5] J. A. Nolazco Flores and S.J. Young, "Adapting a HMM-based Recogniser for Noisy Speech Enhanced by Spectral Subtraction," CUED/F-INFENG/TR.123, Cambridge University Engineering Department, England, April 1993.

[6] J.A. Nolazco Flores and S. J. Young, "CSS-PMC: a Combined Enhancement/Compensation Scheme for Continuous Speech Recognition in Noise," CLUED/F-INFENG/TR.128 June 1993.

[7] S. Sakauchi, A. Nakagawa, Y. Haneda, A. Kataoka, "Implementing and Evaluating of an Audio Teleconferencing Terminal with Noise and Echo Reduction," Proc. International Workshop on Acoustic Echo and Noise Control (IWAENC2003), pp. 191-194, Kyoto, Sep. 2003.

[8] JIS TR S 0001:2002, "A guideline for determining the acoustic properties of auditory signals used in consumer products -- A database of domestic sounds," Jan. 2002.