# NOISE ADAPTIVE SPEECH RECOGNITION WITH ACOUSTIC MODELS TRAINED FROM NOISY SPEECH EVALUATED ON AURORA-2 DATABASE

*Kaisheng Yao*[⋆]*, Kuldip K. Paliwal*[†⋆] *and Satoshi Nakamura*[⋆]

[⋆]ATR Spoken Language Translation Research Laboratories, Kyoto, Japan
[†]School of Microelectronic Engineering, Griffith University, Australia
kaisheng.yao@atr.co.jp k.paliwal@me.gu.edu.au satoshi.nakamura@atr.co.jp

## ABSTRACT

In this paper, we apply the noise adaptive speech recognition for noisy speech recognition in non-stationary noise to the situation that acoustic models are trained from noisy speech. We justify it by that the noise adaptive speech recognition includes iterative processes between a noise parameter estimation step and a model adaptation step, which can possibly do non-linear mapping between the original training space and that for recognition. Experiments were performed on Aurora-2 task with multi-conditional training set which includes noisy utterances. Through experiments, we observed that the noise adaptive speech recognition can have better performance than the baseline system trained from multi-conditional training set without noise adaptive speech recognition.

## 1. INTRODUCTION

Speech recognition has to be carried out often in situations where there exists environment noise, which causes mismatches between pre-trained models and real testing data. Among many approaches for noisy speech recognition, the model-based approach assumes explicit model of noise effects on speech features. It has been shown promising for noisy speech recognition [1].

Associated with the explicitly parametric modeling of the noise effects on speech features, it normally requires that the speech models are trained from clean speech [2]. With the noise parameter estimated prior to speech recognition, the approach can transform the original clean speech HMM to models that have better modeling of the input noisy speech [2], or modify the input noisy observation features to their de-noised speech features which have statistics closer to the pre-trained speech models than the features without the modification [3].

However, in many situations, training data itself includes noisy utterances. There are research efforts to extend the model-based noise compensation to the above situation. For example, Jacobian adaptation [4] employs Taylor series expansion of the noise effects on speech features. The expansion divide the noise effects as two parts, with one part as the effect of the "reference" noise in the acoustic model training stage, and another part of the effect from the "target" noise in the real testing environments. The method includes a process of estimating the difference between the "reference" noise parameter and the "target" noise parameter. Acoustic models that are trained in "reference" noises are combined with the effects of the difference to generate new models that are adapted to the "target" noises.

As many methods in the model-based approach, this method also assumes stationary noise, so the parameters can be estimated given the explicit noise along segments.

In this paper, we apply our recent work on noise adaptive speech recognition [5], which was originally proposed to do noisy speech recognition in non-stationary noises, to the situation that speech models are trained from noisy speech. We justify the application in Section 2 with a novel view of the model-based noise compensation for noisy speech recognition. In particular, Section 2.2 shows that the parametric modeling of the noise effects on speech features can be seen as mapping between two spaces, where one space is considered as the original training data space and the second space is the testing environments. Based on this understanding, it is shown that the noise parameter for mapping between the two spaces can be learned from data. Accordingly, the speech models to be transformed need not be those trained from clean speech. The acoustic models can also be trained from noisy speech. In this situation, the "noise" parameter estimated does not have explicit meaning of noise, but works as the parameter for the mapping.

The noise adaptive speech recognition employs a time-recursive process to sequentially estimate "noise" parameter, which makes this approach can possibly handle non-stationary noises [5]. Another merit of the method is that it does not require explicit noise along segments. The "noise" parameter estimation is carried out during the recognition process. We describe the noise parameter estimation procedure in Section 3. Section 4 provides experimental results to show the efficacy of the method. Conclusions are in Section 5.

## 2. MODEL BASED NOISY SPEECH RECOGNITION

### 2.1. MAP Decision rule for automatic speech recognition

The speech recognition problem can be described as follows. Given a set of trained models $\Lambda_X = \{\lambda_{x_m}\}$ where $\lambda_{x_m}$ is the $m$th subword HMM unit trained from $X$, and an observation sequence $Y(T) = (y(1), y(2), \cdots, y(T))$, the aim is to recognize the word sequence $W = (W(1), W(2), \cdots, W(L))$ embedded in $Y(T)$. Each speech unit model $\lambda_{x_m}$ is a N-state CDHMM with state transition probability $a_{iq} (0 \leq a_{iq} \leq 1)$ and each state is modeled by a mixture of Gaussian probability density functions $\{b_{ik}(\cdot)\}$ with parameter $\{w_{ik}, \mu_{ik}, \mathbf{\Sigma}_{ik}\}_{k=1,2,\cdots,M}$, where $M$ denotes the number of Gaussian mixture components in each state. $\mu_{ik}$ and $\mathbf{\Sigma}_{ik}$ are the mean and variance vector of each Gaussian mixture component. $w_{ik}$ is the mixture weight.

In speech recognition, the model $\Lambda_X$ are used to decode $Y(T)$

using the maximum a posterior (MAP) decoder

$$\begin{aligned}\hat{W} &= \arg\max_{W} P(W|\Lambda_X, Y(T)) \\ &= \arg\max_{W} P(Y(T)|\Lambda_X, W)P_{\Gamma}(W) \end{aligned} \quad (1)$$

where the first term is the likelihood of observation sequence $Y(T)$ given that the word sequence is $W$, and the second term is denoted as the language model. However, in many situations, there exists mismatches due to environments, e.g., additive noise, and accordingly, there is a mismatch in the likelihood of $Y(T)$ given $\Lambda_X$ evaluated by (1).

### 2.2. Model-based noisy speech recognition

In the model-based approach to noisy speech recognition, models representing noise effects on speech features are used. In particular, the following function was proposed in [2][3] to represent additive noise effects on speech features.

$$Y^l = X^l + \log(1 + \exp(N^l - X^l)) \quad (2)$$

where $Y^l$, $X^l$ and $N^l$ each denote the noisy observation, speech, and additive noise. Superscript $l$ denotes that they are in log-spectral domain.

Training data of $X^l$ is used to train the acoustic model $\Lambda_X$. If data of $N^l$ is available, a model $\Lambda_N$ can be trained, so that, by explicit use of the function (2), (1) can be carried out as,

$$\hat{W} = \arg\max_{W} P(Y(T)|\Lambda_X, \Lambda_N, W)P_{\Gamma}(W) \quad (3)$$

In case that $N^l$ is stationary or available before recognition, $\Lambda_N$ can be estimated prior to speech recognition.

#### *2.2.1. Noise parameter estimation for the model-based noisy speech recognition*

Function (2) deserves to be further understood besides its representation as the noise effects on speech features. Figure 1 shows the function when $X^l = 1.0$ and $N^l$ ranges from -10.0 to 10.0. Through the figure, it is seen that the function is smooth and convex as a function of $N^l$ given $X^l$. The function approximates the masking effects of $N^l$ on $X^l$. Function (2) will output either $X^l$ or $N^l$ depends on whether $X^l$ is much larger than $N^l$ or $N^l$ is much larger than $X^l$. When $X^l \approx N^l$, the observation $Y^l$ is non-linearly related with $X^l$ and $N^l$.

Function (2) can be viewed as a parametric mapping between two spaces, $X^l$ and $Y^l$, given $N^l$. Inversely, $N^l$ is the parameter of the function that can be estimated given $X^l$ and $Y^l$.

As shown in the figure, the noise power $N^l$ is masked by speech power $X^l$ in the situation that the noise power is smaller than a certain value. This non-linearity of the function (2) may result in estimate of $N^l$ which is different from its true value. In this sense, it is better to view the estimate as parameter for the non-linear mapping by (2), instead of explicit meaning of noise parameter.

Accordingly, noise compensation in fact includes two steps, the noise parameter estimation step and an acoustic model ( or feature) adaptation step. In the noise parameter estimation step, $\Lambda_N$ ( the parameters of $N^l$) is estimated based on $Y^l$ and $X^l$. Note that, traditional model based methods, e.g., PMC [2] and CDCN [3], assume that $X^l$ is clean speech. In such a case, function (2) shows that $N^l$ is the contaminating noise. Furthermore, if it is stationary

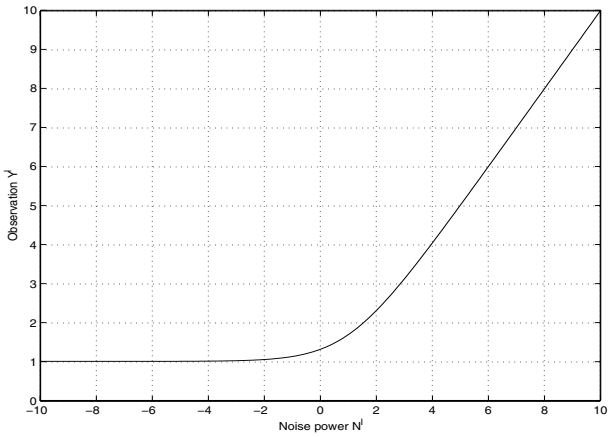


**Fig. 1**. Plot of function $Y^l = X^l + \log(1 + \exp(N^l - X^l))$. $X^l = 1.0$. $N^l$ ranges from -10.0 to 10.0.

noise, its parameter can be estimated given noise along segments. Thus $N^l$ is not a function of $X^l$ in this situation. In contrast to the traditional approach, in this work, parameter of $N^l$ needs to be estimated given $X^l$ and $Y^l$. With the estimated parameter of $N^l$, function (2) is applicable to the situation that noise segments is not explicitly available or the acoustic models of $X^l$ is trained from noisy speech. Note that, however, $N^l$ might not represent true noise in this situation. Normally, a direct observation of $X^l$ is not available, so the parameter of $N^l$ is estimated from $\Lambda_X$ ( the model of $X^l$ ), and $Y^l$ in either a supervised (with correct transcript) or unsupervised (correct transcript is not known) way.

In the acoustic model ( or feature) adaptation step, the estimated parameter of $N^l$ is used in function (2) to transform $\Lambda_X$ (which substitutes $X^l$ in function (2)) in the model space, so that the transformed model $\hat{\Lambda}_Y$ is close to $Y^l$. Similarly, the transformation can be carried out in the feature space to make $Y^l$ close to $\Lambda_X$. In the sequel, we still denote $\Lambda_N$ as the noise model, though it may not be true noise parameter.

## 3. NOISE ADAPTIVE SPEECH RECOGNITION

Furthermore, consider that the noise environment may change during the recognition process. $\Lambda_N$ (in (3)) thus have to be estimated sequentially, i.e., frame-by-frame. The noise adaptive speech recognition [5] is a time-recursive noise parameter estimation and noise compensation method for speech recognition in noises. Based on the discussion in Section 2.2, we apply the time-recursive noise adaptive speech recognition to the situation where acoustic models were trained from noisy speech.

The diagram of the noise adaptive speech recognition is shown in Figure 2. Sequential noise parameter estimation works in the log-spectral domain, and the acoustic model adaptation works in the model-space, i.e. modifying HMM parameters. We briefly review the noise adaptive speech recognition in Section 3.1. Readers please refer to [5] for detailed implementation of the approach.

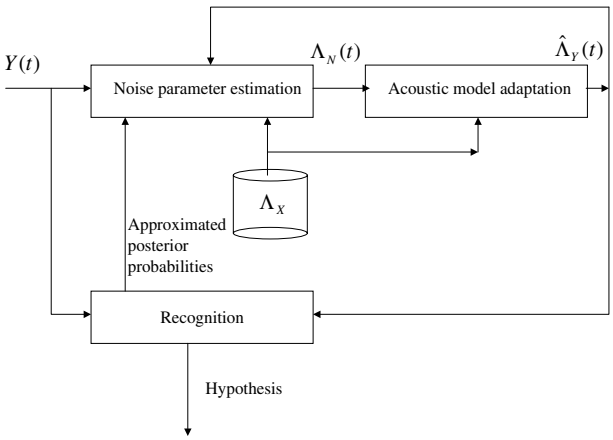


**Fig. 2**. Diagram of the noise adaptive speech recognition. $\Lambda_X$, $\Lambda_N(t)$ and $\hat{\Lambda}_Y(t)$ are the original acoustic model, noise model at frame $t$, and adapted acoustic model at frame $t$, respectively. $Y(t)$ is the input noisy speech observation sequence till frame $t$. Recognition module provides approximated posterior probabilities of state sequences given noisy observation sequences till frame $t$ to the noise parameter estimation module, which output $\Lambda_N(t)$ to adapt acoustic model $\Lambda_X$ to $\hat{\Lambda}_Y(t)$.

### 3.1. A brief review of the noise adaptive speech recognition

Denote the estimated noise parameter sequence till frame $t-1$ as $\Lambda_N(t-1) = (\lambda_N(1), \lambda_N(2), \cdots, \lambda_N(t-1))$. Given the current observation sequence $Y(t) = (y(1), y(2), \cdots, y(t))$ till frame $t$, the noise parameter estimation procedure in the noise adaptive speech recognition will find $\lambda_N(t)$ as the current noise parameter estimate, according to the following objective function.

$$F_t(\hat{\lambda}_N(t)) = Q_t(\lambda^\star_N(t); \hat{\lambda}_N(t)) - (\beta_t - 1) \sum_{S(t)} \log \frac{P(S(t)|Y(t), \Lambda_X, (\Lambda_N(t-1), \lambda_N(t-1)))}{P(S(t)|Y(t), \Lambda_X, (\Lambda_N(t-1), \hat{\lambda}_N(t)))} P(S(t)|Y(t), \Lambda_X, (\Lambda_N(t-1), \lambda_N(t-1))) \quad (4)$$

where the $Q_t(\lambda^\star_N(t); \hat{\lambda}_N(t))$ is the auxiliary function in sequential EM algorithm [6], which is given as,

$$\begin{aligned} & Q_t(\lambda^\star_N(t); \hat{\lambda}_N(t)) \\ &= \sum_{S(t)} P(S(t)|Y(t), \Lambda_X, (\Lambda_N(t-1), \lambda^\star_N(t))) \\ & \quad \log \frac{P(Y(t), S(t)|\Lambda_X, (\Lambda_N(t-1), \hat{\lambda}_N(t)))}{P(S(t)|Y(t), \Lambda_X, (\Lambda_N(t-1), \lambda^\star_N(t)))} \end{aligned} \quad (5)$$

$S(t) = (s(1), s(2), \cdots, s(t))$ is the state sequence till frame $t$. $\beta_t \in R^+$ works as a relaxation factor. $\lambda^\star_N(t)$ is initialized to be $\lambda_N(t-1)$.

At each iteration, the procedure will calculate the posterior probabilities $P(S(t)|Y(t), \Lambda_X, (\Lambda_N(t-1), \lambda^\star_N(t)))$, and then maximize the objective function to obtain $\hat{\lambda}_N(t)$. After one iteration, the estimated $\hat{\lambda}_N(t)$ will be set to $\lambda^\star_N(t)$, and a new iteration is carried out. Though, generally, several iterations are required to obtain the final $\hat{\lambda}_N(t)$ as the estimate of $\lambda_N(t)$, it can in fact be approximately estimated by only one iteration. The time recursive procedure is the sequential Kullback proximal algorithm [7], which is a generalization of the sequential EM algorithm. The sequential EM algorithm is a special case of this algorithm and corresponds to setting $\beta_t$ equal to 1.0 in the algorithm. The algorithm can achieve faster parameter estimation than that by sequential EM algorithm.

The joint likelihood of observation sequence $Y(t)$ and state sequence $S(t)$ can be approximately obtained from the Viterbi process, i.e.,

$$\begin{aligned} P(Y(t), S(t)|\Lambda_X, \Lambda_N(t)) \approx a_{s^\star(t-1)s(t)} b_{s(t)}(y(t)) \\ P(Y(t-1), S^\star(t-1)|\Lambda_X, \Lambda_N(t-1)) \end{aligned} \quad (6)$$

where

$$S^\star(t-1) = \arg \max_{S(t-1)} a_{s(t-1)s(t)} P(Y(t-1), S(t-1)|\Lambda_X, \Lambda_N(t-1)) \quad (7)$$

By normalizing the joint likelihood with respect to the sum of those from all active partial state sequences, an approximation of the posterior probability of state sequence can be obtained. This scheme of time-varying noise parameter estimation is denoted as noise adaptive speech recognition [5].

In particular, the noise adaptive speech recognition estimates time-varying noise parameter in the log-spectral domain. The noise model $\lambda_N(t)$ is a single Gaussian with time-varying mean vector $\mu^l_n(t) \in R^J$, which needs to be estimated, and constant variance $\mathbf{\Sigma}^l_N \in R^J$. $J$ is the number of filter banks. The model adaptation is carried out by the following function on mean vector $\mu^l_{ik} \in R^J$ in each mixture $k$ of state $i$ in speech models. That is,

$$\mu^l_{ik}(t) = \mu^l_{ik} + \log(1 + \exp(\mu^l_n(t) - \mu^l_{ik})) \quad (8)$$

Note that function (8) is an approximation to function (2) with the assumption that the "noise" $N^l$ has very small variance.

## 4. EXPERIMENTAL RESULTS

Normally, model-based noise compensation methods require that the acoustic models were trained from clean speech. In our discussion in Section 2.2, we have shown that by iterative process between estimation of "noise" parameter $\Lambda_N$ and transformation of acoustic model $\Lambda_X$ by function (2), the acoustic model can be trained from noisy speech.

We evaluated the noise adaptive speech recognition on Aurora-2 database [8], which is a noise-contaminated TI-Digits database down-sampled to 8kHz. The training set for noise adaptive speech recognition system (denoted as Adaptive) and the baseline without noise compensation (denoted as Baseline) was the multi-conditional training set with 8840 utterances containing Subway, Babble, Car and Exhibition hall noises in five different SNR conditions from 5dB to clean condition in 5dB step. The testing set contains 20020 noisy utterances with five SNR conditions from 0dB to 20dB, and with the same contaminating noise as the training set.

Digits and silence were respectively modeled by 10-state and 3-state whole word HMMs with 4 diagonal Gaussian mixtures in each state. The window size was 25.0ms with a 10.0ms shift. A filter-bank of twenty-six filters was used in the binning stage. Features were MFCC + C0 and their first order derivatives. The feature

**Table 1**. Word Accuracy (in %) in the Aurora-2 database, achieved by the noise adaptive speech recognition (denoted as Adaptive) with relaxation factor $\beta_t = 0.9$ and forgetting factor $\rho = 0.995$, in comparison with baseline without noise adaptive speech recognition (denoted as Baseline). Acoustic models were trained from multi-conditional training set. Averaged relative error rate reductions (ERR) over Baseline in each noise are in the last row. ERRs in each SNR are in the last column.

| SNR (dB) | Subway | | Babble | | Car | | Exhibit | | ERR |
|---|---|---|---|---|---|---|---|---|---|
| | Adaptive | Baseline | Adaptive | Baseline | Adaptive | Baseline | Adaptive | Baseline | ( in %) |
| 20.0 | 87.98 | 84.65 | 92.75 | 86.12 | 92.83 | 92.74 | 90.44 | 90.56 | 17.4 |
| 15.0 | 87.03 | 78.11 | 89.72 | 80.56 | 91.00 | 90.92 | 86.95 | 87.58 | 20.9 |
| 10.0 | 82.83 | 70.60 | 84.58 | 72.52 | 87.04 | 87.13 | 82.85 | 83.77 | 19.8 |
| 5.0 | 76.06 | 63.09 | 75.76 | 61.83 | 76.53 | 75.20 | 75.03 | 74.54 | 19.7 |
| 0.0 | 62.10 | 53.19 | 58.68 | 49.62 | 52.53 | 53.41 | 61.59 | 57.16 | 11.4 |
| ERR ( in % ) | 31.4 | | 38.7 | | 1.0 | | 0.0 | | |

dimension was 26. Though it was possible that we could improve performances by increasing feature dimension, or state and mixture numbers, our major objective was to show the validity of the noise adaptive speech recognition in multi-conditional training.

The noise adaptive speech recognition was set with relaxation factor $\beta_t = 0.9$ and forgetting factor $\rho = 0.995$. At the beginning, noise parameter $\mu_n^l(t)$ was initialized to be zero vector. Afterwards, the system did not make explicit initializations in each SNR and noise.

Recognition performances of the system and "Baseline" are shown in Table 1. We observed that the noise adaptive speech recognition has significantly better performance than "Baseline" in Subway and Babble noise. In the view of the averaged relative error rate reduction (ERR) in each noise, which is calculated as the average of the relative error rate reductions in the noise, system "Adaptive" achieved 31.4% and 38.7% ERR over "Baseline" in Subway and Babble noises, respectively. For other kinds of noise, it performed as good as "Baseline".

The average relative error rate reductions (ERR) in each SNR are shown in the last column in the table. Through the table, it is seen that the largest ERR was obtained in 15dB SNR, which achieved 20.9% ERR, while the smallest ERR was obtained in 0dB SNR.

## 5. CONCLUSIONS AND DISCUSSIONS

Model-based noise compensation [2] normally requires that acoustic models for speech are trained from clean speech. This limits its application when the training data in fact has noisy utterances. Though there are research efforts to extend the model-based approach to the above situation, for example, the Jacobian adaptation [4], they normally assume stationary noises and the availability of the explicit noise along segments.

In this paper, noise adaptive speech recognition [5], which was recently proposed to do speech recognition in non-stationary noises, is applied to the situation that training data includes noisy utterances. Our results confirm that the noise adaptive speech recognition [5] is applicable to the situation. In this situation, the "noise" parameter estimated may not have the explicit meaning of noise, but works as the parameter for the parametric mapping (2).

The above experimental results have shown that the noise adaptive speech recognition improves system performances in noises. Further improvement in this research can be achieved via incorporation of adaptation for the dynamic features and refinement of acoustic models. Since the "noise" parameter estimation process is iterative in the approach, initialization is important to get good performance. This is also a further research topic along this approach.

## 6. ACKNOWLEDGEMENT

This research was supported in part by the Telecommunications Advanced Organization of Japan.

## 7. REFERENCES

[1] S. V. Vaseghi and B. P. Milner, "Noise compensation methods for hidden markov model speech recognition in adverse environments," *IEEE. Trans. on Speech and Audio Processing*, vol. 5, no. 1, pp. 11–21, January 1997.

[2] M.J.F.Gales and S.J.Young, "Robust speech recognition in additive and convolutional noise using parallel model combination," *Computer Speech and Language*, vol. 9, pp. 289–307, 1995.

[3] A. Acero, *Acoustical and Environmental Robustness in Automatic Speech Recognition*, Ph.D. thesis, Carnegie Mellon University, September 1990.

[4] S. Sagayama, Y. Yamaguchi, S. Takahashi, and J. Takahashi, "Jacobian approach to fast acoustic model adaptation," in *ICASSP*, 1997, pp. 835–838.

[5] K. Yao, K. Paliwal, and S. Nakamura, "Noise adaptive speech recognition in time-varying noise based on sequential kullback proximal algorithm," in *ICASSP*, 2002, pp. 189–192.

[6] V. Krishnamurthy and J. B. Moore, "On-line estimation of hidden markov model parameters based on the kullback-leibler information measure," *IEEE Trans. on Signal Processing*, vol. 41, no. 8, pp. 2557–2573, August 1993.

[7] K. Yao, K. K. Paliwal, and S. Nakamura, "Sequential noise compensation by a sequential kullback proximal algorithm," in *EUROSPEECH*, 2001, pp. 1139–1142.

[8] D. Pearce, "Aurora project: Experimental framework for the performance evaluation of distributed speech recognition front-ends," in *ISCA ITRW ASR2000*, Sep. 2000.