CONTINUOUS SPEECH RECOGNITION UNDER NON-STATIONARY MUSICAL ENVIRONMENTS BASED ON SPEECH STATE TRANSITION MODEL

M.Fujimoto and Y.Ariki

Department of Electronics and Informatics
Ryukoku University, Seta, Otsu-shi, Shiga, 520-2194, JAPAN
masa@arikilab.elec.ryukoku.ac.jp

ABSTRACT

In this paper, we propose a non-stationary noise reduction method based on speech state transition model. Our proposed method estimates the speech signal under non-stationary noisy environments such as musical background by applying speech state transition model to Kalman filtering estimation. The speech state transition model represents the state transition of speech component in non-stationary noisy speech and is modeled by using Taylor expansion. In this model, the state transition of noise component is estimated by using linear predictive estimation. In order to evaluate the proposed method, we carried out large vocabulary continuous speech recognition experiments under 3 types of music and compared the results with conventionally used Parallel Model Combination(PMC) method in word accuracy rate. As a result, the proposed method obtained word accuracy rate superior to PMC.

1. INTRODUCTION

In recent years, many types of speech recognition systems have been proposed and developed toward the practical use in the real world. However, most of the works recognize clean speech collected in quiet environments. For practical use it is required for recognition systems to be robust for interfering noises, especially non-stationary noises.

Robust speech recognition systems are classified into two types. One adapts itself to any kinds of noises based on model adaptation techniques[1]-[3]. The other reduces the noise component from noisy speech based on noise reduction techniques[4]-[7].

Parallel Model Combination(PMC)[1, 2] has been proposed which adapts the speech recognition system to any kinds of noises. To improve the recognition accuracy under non-stationary noisy environments, time varying residual noise compensation method has been proposed[3]. But model adaptation method has a problem that it needs a huge quantity of computation, if it is applied to the acoustic model which has a large number of phonemes with mixture distributions like a triphone model HMM.

On the other hand, Spectral Subtraction(SS)[4] has been proposed as a conventional noise reduction method. However, SS has a problem that it deteriorates the recognition rate due to spectral distortion by over or under subtraction. In addition, subtracted noise spectral is mean spectral estimated from the time section assumed to be noise (beginning of utterance) and the SS does not consider the time varying of noise spectral.

In this paper, we propose a non-stationary noise reduction method based on speech state transition model. The method estimates the speech signal under non-stationary noisy environments by applying speech state transition model to Kalman filtering estimation. In order to evaluate the proposed method, we carried out Large Vocabulary Continuous Speech Recognition (LVCSR) experiments under non-stationary noisy environments and compared the results by our method with those by PMC in word accuracy rate.

2. SPEECH STATE TRANSITION MODEL

At the kth frame, power spectra of clean speech under noisy environments is represented as follows:

$$S(k) = \exp\left(X^l(k)\right) - \exp\left(N^l(k)\right) \qquad (1)$$

where $X(k)$, $S(k)$ and $N(k)$ denote the vectors of power spectra of noisy speech, clean speech and noise at the kth frame respectively, and superscript l denotes the log-spectral domain.

In Eq.(1), speech state transition from $S(k)$ to $S(k+1)$ is represented as follows:

$$S(k+1) \quad = \quad S(k) + \Delta S(k)$$

$$= \exp\left(X^l(k) + \Delta X^l(k)\right) - \exp\left(N^l(k) + \Delta N^l(k)\right) \quad (2)$$

where $\Delta X^l(k) = X^l(k+1) - X^l(k)$ and $\Delta N^l(k) = N^l(k+1) - N^l(k)$ respectively.
Here, by expanding Eq.(2) using first order Taylor series, speech state transition can be linearized as follows:

$$\begin{aligned} S(k+1) &\simeq S(k) + \frac{\partial S(k)}{\partial X^l(k)}\Delta X^l(k) + \frac{\partial S(k)}{\partial N^l(k)}\Delta N^l(k) \\ &= S(k) + X(k)\Delta X^l(k) - N(k)\Delta N^l(k) \\ &= S(k) + (S(k) + N(k))\,\Delta X^l(k) \\ &\quad -N(k)\Delta N^l(k) \\ &= \left(1 + \Delta X^l(k)\right) S(k) \\ &\quad +N(k)\left(\Delta X^l(k) - \Delta N^l(k)\right) \\ &= F_k S(k) + G_k W(k) \end{aligned} \quad (3)$$

$$\frac{\partial S(k)}{\partial X^l(k)} = \frac{\partial (X(k) - N(k))}{\partial X(k)} \cdot \frac{\partial X(k)}{\partial X^l(k)} = X(k) \quad (4)$$

$$\frac{\partial S(k)}{\partial N^l(k)} = \frac{\partial (X(k) - N(k))}{\partial N(k)} \cdot \frac{\partial N(k)}{\partial N^l(k)} = -N(k) \quad (5)$$

where $F_k = 1 + \Delta X^l(k)$, $G_k = N(k)$ and $W(k) = \Delta X^l(k) - \Delta N^l(k)$ respectively.
In Eq.(3)$\sim$(5), $\frac{\partial S(k)}{\partial X^l(k)}$ and $\frac{\partial S(k)}{\partial N^l(k)}$ mean that partial differentiation applies to each element independently, under the assumption that each element in the vector is uncorrelated.
In the above description, we defined Eq.(3) as speech state transition model and applyed Eq.(3) to Kalman filtering estimation to estimate speech power spectra $S(k)$ from noisy power spectra $X(k)$.

3. KALMAN FILTERING ESTIMATION

3.1. THE STATE SPACE MODEL

To estimate the $S(k)$ by using Kalman filtering estimation, we determined the state space model as follows:

$$S(k+1) = F_k S(k) + G_k W(k) \quad (6)$$

$$X(k) = S(k) + N(k) \quad (7)$$

In above equations, Eq.(6) corresponds to state equation, and Eq.(7) corresponds to observation equation.

3.2. KALMAN FILTERING ALGORITHM

By using the state space model described in 3.1, Kalman filtering algorithm is obtained as follows:

$$\hat{S}(k) = F_{k-1}\hat{S}(k-1) + K_k(X(k) - F_{k-1}\hat{S}(k-1)) \quad (8)$$

$$K_k = Q_k\left[Q_k + \Sigma_{N(k)}\right]^{-1} \quad (9)$$

$$\begin{aligned} Q_k &= F_{k-1}(I - K_{k-1})Q_{k-1}F_{k-1}^T \\ &\quad +G_{k-1}\Sigma_{W(k-1)}G_{k-1}^T \end{aligned} \quad (10)$$

where $\hat{S}(k)$ denotes the estimation of $S(k)$ and Q_k denotes diagonal co-variance matrix of the estimating error respectively.
The initial values for Eq.(8)$\sim$(10) are represented as follows:

$$\hat{S}(0) = \mathbf{0} \quad (11)$$

$$Q_0 = \mathbf{0} \quad (12)$$

In Eq.(10), $\Sigma_{W(k)}$ denotes diagonal co-variance matrix of $W(k)$. $\Sigma_{W(k)}$ is computed by the following equation under the assumption that $W(k)$ follows zero mean Gaussian process.

$$\Sigma_{W(k)} = W(k)W(k)^T \quad (13)$$

On the other hands, in Eq.(9), $\Sigma_{N(k)}$ denotes diagonal co-variance matrix of $N(k)$. $\Sigma_{N(k)}$ is computed by the following equation under the assumption that $N(k)$ follows zero mean Gaussian process as well as $W(k)$.

$$\Sigma_{N(k)} = N(k)N(k)^T \quad (14)$$

3.3. LINEAR PREDICTIVE ESTIMATION FOR $N(k)$

To compute the $\Sigma_{W(k)}$ and $\Sigma_{N(k)}$, the value of $N(k)$ is required. However, observable value is only $X(k)$. Therefore, we have to estimate the value of $N(k)$ by using pth order linear prediction expressed as follows:

$$N_j(k) = \begin{cases} X_j(k) & 0 \le k < p \\ \sum_{i=1}^{p} a_{ij} N_j(k-i) & k \ge p \end{cases} \quad (15)$$

where j denotes the channel number in FFT analysis and a_{ij} denotes the linear predictive coefficient at channel j.
In Eq.(15), when $0 \le k < p$, $N_j(k)$ is obtained by $N_j(k) = X_j(k)$ under the assumption that the time section $0 \le k < p$ exists where only the noise component is included as at the beginning of utterance and when $k \ge p$, $N_j(k)$ is estimated by the linear predictive estimation. In this paper, the number of linear predictive coefficient p was set to 12.

4. EXPERIMENTS

LVCSR experiments were carried out for the speech signals estimated by the proposed method. As a comparison, LVCSR by PMC was also carried out.

4.1. EXPERIMENTAL CONDITIONS

The experimental materials are 100 sentences spoken by 23 Japanese males. These materials are taken from the IPA(Information-technology Promotion Agency, Japan)-98-TestSet. The noises are non-stationary music of 3 piano solos(Piano 1, Piano 2 and Piano 3). They are added to clean speech signal by a computer as shown in Eq.(16), changing the SNR at 3 levels; 0dB, 10dB and 20dB.

$$x(t) = s(t) + \frac{Pow_s}{10^{SNR/20} Pow_n} \cdot n(t) \qquad (16)$$

where $x(t)$, $s(t)$ and $n(t)$ are noisy speech, clean speech and noise respectively. Pow_s and Pow_n are RMS power of clean speech and RMS power of noise respectively.
We carried out LVCSR using speaker independent monophone HMMs. Their structure is composed of 5 states with 3 loops and 12 mixtures for each state. They were trained using 21,782 sentences spoken by 137 Japanese males. These speech data was taken from the database of Acoustical Society of Japan. The feature parameters are 39 MFCCs with 12 MFCCs, log energy and their first and second order derivatives. Cepstral Mean Normalization(CMN) is applied to each sentence to remove the difference of input circumstances.
For noise HMM used in PMC, the HMM structure is 3 states with 1 loop and 1 mixture for each state. Table 1, 2 and 3 summarize the experimental conditions for acoustic analysis, phoneme HMM and noise HMM.
Here, MFCC as feature parameters for LVCSR was not computed from the wave form reconstructed from power spectra of the speech estimated by the proposed method, but it was computed directly from estimated power spectra by using Mel Filter Bank and DCT. Then CMN is applied to each sentence as well as the training data.

Table 1: Acoustic analysis conditions

Sampling frequency	16kHz
Pre-emphasis	$1 - 0.97z^{-1}$
Feature parameter (Noise reduction)	FFT spectra(512th order)
Feature parameter (Recognition)	MFCC(12th order) + Log-Power + Δ + $\Delta\Delta$
Analysis frame length	20ms
Analysis frame shift	10ms
Analysis window	Hamming window

Table 2: Structure of phoneme HMM

Number of states	5
Number of loops	3
Number of mixtures	12
Number of phonemes	41
Type	Left to right HMM

Table 3: Structure of noise HMM

Number of states	3
Number of loops	1
Number of mixtures	1
Type	Left to right HMM

A language model is bigram for the 1st-pass in the continuous speech decoder and trigram for the 2nd-pass. It was trained using the Mainichi newspaper articles of 75 months. The number of the words in the dictionary is 20,000.

4.2. EXPERIMENTAL RESULTS

Table 4, 5 and 6 show results of the LVCSR under the non-stationary 3 types of piano music. In each table, upper row shows word correct rate($Corr$) and lower row shows word accuracy(Acc). They are defined by Eq.(17) and Eq.(18).

$$Corr(\%) = \frac{N - S - D}{N} \times 100 \qquad (17)$$

$$Acc(\%) = \frac{N - S - D - I}{N} \times 100 \qquad (18)$$

S : The number of substituted words
D : The number of deleted words
I : The number of inserted words
N : Total number of words

Table 4: Recognition results(Piano 1)(%)
(Upper: $Corr$, Lower: Acc)

SNR	∞dB	20dB	10dB	0dB
No processing	88.78	83.26	69.88	35.64
	86.49	79.52	61.57	20.04
PMC	88.78	81.42	69.06	37.22
	86.49	78.50	63.47	30.94
Proposed	88.14	85.23	75.52	50.35
	85.16	81.17	67.91	36.33

Table 5: Recognition results(Piano 2)(%)
(Upper: *Corr*, Lower: *Acc*)

SNR	∞dB	20dB	10dB	0dB
No processing	88.78	82.69	66.01	34.43
	86.49	79.62	56.44	20.04
PMC	88.78	80.66	67.47	38.62
	86.49	77.62	61.64	31.90
Proposed	88.14	82.75	71.15	46.48
	85.16	78.19	62.08	31.96

Table 6: Recognition results(Piano 3)(%)
(Upper: *Corr*, Lower: *Acc*)

SNR	∞dB	20dB	10dB	0dB
No processing	88.78	83.45	70.13	37.54
	86.49	79.45	61.95	21.81
PMC	88.78	81.36	68.36	38.11
	86.49	78.38	63.35	32.53
Proposed	88.14	85.10	75.90	48.45
	85.16	81.23	67.53	33.61

In each table, comparing with PMC, the proposed method showed the significant improvement in $Corr$ under all the conditions. However, the improvement in *Acc* was small under all the conditions. From this fact, it can be assumed that the word substitution and deletion has decreased and the word insertion has increased.

From the fact that the word insertion has increased, it can be assumed that the estimation accuracy of noise power spectra $N(k)$ was not obtained because linear prediction error of $N_j(k)$ in Eq.(15) was large. In preliminary experiments, we confirmed that if $N(k)$ is true, the improvement of *Acc* is significantly large (for example, *Acc* is improved up to approximately 78% from 36.33% for Piano 1 under 0dB noisy environment.). Therefore, it is necessary to estimate $N(k)$ as accurate as possible.

In addition, in Eq.(3), we applied the first order Taylor expansion to Eq.(2). However, the first order Taylor expansion is rough approximation in accuracy. Therefore, approximation accuracy should be improved by using the high order Taylor expansion.

5. CONCLUSIONS

In this paper, we proposed the non-stationary noise reduction method based on speech state transition model and showed the significant improvement in word correct rate. In future, to improve the word accuracy under any types of non-stationary noisy environments, we will study accurate estimation method for state transition of the noise spectra.

6. REFERENCES

[1] M.J.F.Gales, S.J.Young: "An Improved Approach to the Hidden Markov Model Decomposition of Speech and Noise", *ICASSP*, I-233-236(1992)

[2] M.J.F.Gales, S.J.Young: "Robust Continuous Speech Recognition Using Parallel Model Combination", *IEEE Trans. Speech and Audio Processing*, Vol.4, No.5, pp.352-359, Sep.(1996)

[3] Kaisheng Yao, Bertram E. Shi, Pascale Fung, Zhigang Cao: "Residual Noise Compensation for Robust Speech Recognition in Nonstationary Noise", *ICASSP*, II-1125-1128(2000).

[4] S.F.Boll: "Suppression of Acoustic Noise in Speech Using Spectral Subtraction", *IEEE Trans. Acoustic Speech Signal Processing*, Vol.27, No.2, pp.113-120, (1979)

[5] D.C.Popescu, I.Zejiković: "Kalman Filtering of Colored Noise for Speech Enhancement", *ICASSP*, II-997-1000(1998)

[6] Z.Goh, K.Tan, B.T.G.Tan: "Kalman-Filtering Speech Enhancement Method Based on Voiced-Unvoiced Speech Model", *IEEE Trans. Speech and Audio Processing*, Vol.7, No.5, pp.510-524, Sep.(1999)

[7] M.Fujimoto, Y.Ariki: "Noisy Speech Recognition Using Noise Reduction Method Based on Kalman Filter", *ICASSP*, III-1723-1726(2000)