Speech Recognition using MFCC with Time Varying LPC for Similar Pronunciation Phrases

George Mufungulwa, Alia Asheralieva, Hiroshi Tsutsui, and Yoshikazu Miyanaga
Graduate School of Information Science and Technology, Hokkaido University
Kita 14, Nishi 9, Kita-ku, Sapporo 060-0814, Japan

Abstract: Noises pause a serious challenge in speech recognition systems. Both multiplicative and additive noises affect the performance of automatic speech recognition (ASR) systems. In this paper, a method which makes use of fast Fourier transform (FFT) based mel-frequency cepstral coefficients (MFCCs) with time-varying linear predictive coding (TVLPC) is proposed. Evaluation results of similar pronunciation phrases and 142 words recognition experiments demonstrate that the proposed approach achieves better results at 10 dB SNR than the conventional approach.

Keywords— **MFCC, time-varying LPC, HMM, Speech Recognition, Similar Pronunciation**

1. Introduction

It is a fundamental requirement that speech recognition systems are noise robust and that their performance are resilient in a variety of noisy environments. In this regard, a number of speech recognition technologies that give a high recognition accuracy in noise environments have been developed before [1], for example.

The difference in additive noise and difference in multiplicative noise in the spectral domain often degrades the recognition accuracy in speech as discussed in [2]. Speech enhancement is, therefore, aimed at improving the performance of speech recognition systems in such noisy environments.

Some of the most popular methods for reducing additive noise are spectral subtraction (SS) method, dynamic range adjustment (DRA) and the Wiener filter approach. The SS method, as detailed in [3], is used for restoration of the power spectrum or the magnitude spectrum of an observable signal in additive noise as applied in [4]. The DRA is a technique that corrects the difference between dynamic ranges by normalizing amplitude of speech features for each dimension of cepstrum as discussed in [5]. The Wiener filter approach [6] obtains a least squares estimate of the clean signal assuming that speech and noise are stationary and therefore, can help enhance speech as explained in [7].

Channel and speaker adaptation are fundamental requirements in most conversational speech recognition systems as presented in [8]. As such, cepstrum mean subtraction (CMS) is often used in order to compensate for the acoustic channel in a channel normalization approach. After Fourier transform in the running spectrum the influence of multiplicative noise is concentrated on the modulation frequency lower than 1 Hz. An important component in speech recognition is present in the range of about 1 to 10 Hz modulation frequency. The running spectrum analysis (RSA) method is used in order to remove un-speech components over 15 Hz in modulation spectrum domain thereby emphasizing speech features as detailed in [9].

This work is an extension of our baseline study we conducted [10]. The difference is that our baseline study focused on model variations and similar pronunciation phrases with white noise while this one aims at similar pronunciation phrases as well as 142 isolated word recognition using 15 types of noise. We seek to determine the noise level under which our proposed approach performs better than the conventional approach.

The rest of the paper is organized as follows. In Section 2, we present the steps involved in feature extraction process. In Section 3, we present simulation parameters, experimental conditions and feature vector representation. In Section 4, we present simulation results and their analysis. Finally, in Section 5, we conclude this paper.

2. Proposed Speech Features Based on Time Varying LPC

The FFT based MFCCs are computed as discussed in [11], [12] for examples. For all-pole signal modeling, the output signal $s[n]$ at time n is modelled as a linear combination of the past p samples and the input $u[n]$, with G as a gain constant i.e.,

$$s[n] = -\sum_{i=1}^{p} a_i s[n-i] + Gu[n]. \qquad (1)$$

For the method of time-varying linear prediction, the prediction coefficients are allowed to change with time progress [13]. The time-varying model can be represented by

$$s[n] = -\sum_{i=1}^{p} a_i[n] s[n-i] + Gu[n]. \qquad (2)$$

The assumption is that the signal is not stationary in an observed frame. Therefore, the time-varying nature of the coefficient $a_i[n]$ must be specified. We have chosen to model these coefficients as the linear combinations of some known functions of time $u_k[n]$:

$$a_i[n] = \sum_{k=0}^{q} a_{ik} u_k[n]. \qquad (3)$$

In this model, the coefficients a_{ik} are to be estimated from the speech signal, where the subscript i is a reference to the time-varying coefficient $a_i[n]$, the subscript k is a reference to the set of time functions $u_k[n]$ and q is the basis function order. From (2) and (3), the predictor equation is given by

$$\hat{s}[n] = -\sum_{i=1}^{p} \Big(\sum_{k=0}^{q} \hat{a}_{ik} u_k[n] \Big) s[n-i], \qquad (4)$$

and based on (2) and (4), the predictor error $e[n]$ is defined by

$$e[n] = s[n] - \hat{s}[n]. \tag{5}$$

The predictor error must be minimized with respect to each coefficient. A model is usually optimized for the data it was trained for. Therefore, the accurate measure of predictor error is significant in model assessment. It minimizes chances of choosing a model that may produce misleading results on testing data. We assume a constant optimism such that the model that minimizes training error will also be the model that will minimize the true predictor error for our testing data.

If we assume $q = 1$, then from (4) the following equation can be obtained

$$\hat{s}[n] = -\sum_{i=1}^{p}(\hat{a}_{i,0}u_0[n] + \hat{a}_{i,1}u_1[n])s[n-i]. \tag{6}$$

If we allowed arbitrary variations in the coefficients, we would have as many degrees of freedom in the parametric model as in the original data, thereby achieving no data compression or insight into the structure of the signal. By judicious choice of the basis functions $u_k[n]$ we can accurately approximate a wide variety of coefficient time variations. Possible sets of functions that could be used include powers of time

$$u_k[n] = n^k, k = 0, 1, ..., \infty. \tag{7}$$

Using (7) we have $u_0[n] = 1$, and $u_1[n] = n$, then we can obtain the following equation

$$\hat{s}[n] = -\sum_{i=1}^{p}(\hat{a}_{i,0} + \hat{a}_{i,1}n)s[n-i]. \tag{8}$$

Although (8) can not represent the various types of time varying models, it is applied for the limited structure of a linearly and slowly time-variations on an autoregressive model in an observed frame $s[n]$. Note that n equals zero at the beginning of each observed frame. We assume in this paper the above model can be employed for the representation of speech features in a frame.

Using this model, the following speech model is given from (8):

$$\hat{s}[n] = -\sum_{i=1}^{p}\hat{a}_{i,0}s[n-i] - n\sum_{i=1}^{p}\hat{a}_{i,1}s[n-i]. \tag{9}$$

In (9), the first part of the right-hand side represents time-invariant factor and thus the second part represents time varying factor. Accordingly, if we can assume the observed speech signal can be represented as $s[n] \to s_0[n] + s_1[n]$ where $s_0[n]$ shows a time-invariant factor and $s_1[n]$ shows a time varying factor respectively, then we get

$$H_0^p(z^{-1}) = \frac{1}{1 + \hat{a}_{1,0}z^{-1} + \hat{a}_{2,0}z^{-2} + ... + \hat{a}_{p,0}z^{-p}}, \tag{10}$$

$$H_1^p(z^{-1}) = \frac{1}{1 + \hat{a}_{1,1}z^{-1} + \hat{a}_{2,1}z^{-2} + ... + \hat{a}_{p,1}z^{-p}}, \tag{11}$$

Table 1. Parameters for 3 similar pronunciation phrases using conventional FFT based MFCC analysis -[Set A] and proposed FFT based MFCC with TVLPC analysis -[Set B]

Parameter	**Set[A]**	**Set[B]**
Sampling	11.025 kHz, (16-bit)	11.025 kHz, (16-bit)
Frame length	23.2 ms (256 samples)	23.2 ms (256 samples)
Shift length	11.6 ms (128 samples)	11.6 ms (128 sampples)
Pre emphasis	$1-0.97z^{-1}$	$1-0.97z^{-1}$
Windowing	Hanning window	none
Feature vectors	b_i, Δb_i, $\Delta^2 b_i$	$\Delta b_i(i = 1, \ldots, 12)$, $\Delta b_i(i = 0, \ldots, 12)$, $\Delta^2 b_i(i = 0, \ldots, 12)$, $c_{i,1}(i = 1, \ldots, 12)$
TVLPC order		p = 14
Training set	30 male speakers, 3 utterances each	30 male speakers, 3 utterances each
Recognition set	10 male speakers, 1 utterance each	10 male speakers, 1 utterance each
HMM states	32	32
Noise type	white noise and 15 types from NOISEX-92	white noise and 15 types from NOISEX-92
SNR	10 dB, 20 dB	10 dB, 20 dB
Noise reduction methods	DRA, CMS/DRA, RSA, RSA/DRA	DRA, CMS/DRA, RSA, RSA/DRA

where $H_0^p(z^{-1})$ indicates a time invariant transfer function and $H_1^p(z^{-1})$ indicates a time varying transfer function.

The intra-frame cepstrum $c_{i,0}$ for time invariant coefficients $a_{i,0}$ is given by

$$c_{i,0} = -a_{i,0} - \frac{1}{i}\sum_{m=1}^{i-1} mc_{m,0}a_{i-m,0} \tag{12}$$

and the intra-frame cepstrum $c_{i,1}$ for time varying coefficients $a_{i,1}$ is given by

$$c_{i,1} = -a_{i,1} - \frac{1}{i}\sum_{m=1}^{i-1} mc_{m,1}a_{i-m,1}. \tag{13}$$

3. Simulation Parameters and Conditions of Experiments

The simulation parameters are as shown in Table 1, where set [A] was used for the conventional approach while set [B] was used for the proposed approach. The noise robustness of our independent speaker speech recognition was evaluated based on an isolated word recognition task using the Noisex-92 database [14]. Two main experiments were conducted. In the first experiment, 30 male speakers each uttering 3 similar pronunciation Japanese phrases (“denki”, “genki”, “tenki”) with a utterance frequence of 3 were utilised. In the second experiment, the same number of speakers each uttering 142 isolated words with an utterance frequence of 3 were utilised. In both experiments speech sample was 11.025 kHz and 16-bit quantization. In our experiment, FFT based MFCC features were extracted after pre-emphasis and Hanning windowing. The features were then converted to 38-dimensional feature vectors. Frame length and shift length were 23.2 ms (256 samples) and 11.6 ms (128 samples) respectively. On the other hand, TVLPC features were extracted separately after pre-emphasis but without windowing. The features were converted to model number dimensional feature vectors respectively.

In the proposed approach, the time invariant MFCC features are appended with intra-frame cepstrum estimated using TVLPC according to the model number. In the recognition stage, 10 dB and 20 dB of 15 types of noise respectively are artificially added to the original speech. In the first phase of our two major experiments we measure the average recognition rates of 10 male independent speakers, uttering 3 similar Japanese phrases while in the second phase of the experiment we measure for 142 isolated Japanese words.

3.1 Feature Representation

We formulate thirteen models, models 0 to 12 of feature vectors. The model 0 is considered as a conventional approach and its 38-parameter feature vector consisting of 12 cepstral coefficients (without the zero-order coefficient) plus the corresponding delta and acceleration coefficients is given by $[b_1 b_2 \dots b_{12} \Delta b_0 \Delta b_1 \dots \Delta b_{12} \Delta^2 b_0 \Delta^2 b_1 \dots \Delta^2 b_{12}]$ where b_i, Δb_i and $\Delta^2 b_i$, are MFCC, delta MFCC and delta-delta MFCC, respectively. As for the proposed models, j ($j = 1, \dots, 12$), we append the intra-frame cepstrum (without the zero-order coefficient) for time-varying coefficients $[c_{1,1}, c_{2,1}, \dots, c_{j,1}]$ to the model 0 feature vector depending on the model number.

4. Simulation Results

In our experiments we evaluate the performance of both the conventional approach and proposed approach on DRA, CMS/DRA, RSA and RSA/DRA using MATLAB (R2014a) software. Table 2 shows the average recognition accuracy on clean speech for models 0 to 12. Shown in Table 3 are comparative recognition results for similar pronunciation phrases on white noise at 10 dB and at 20 dB SNR.

Table 2. Average recognition (%) accuracy of similar pronunciation phrases on clean speech

Clean	DRA	CMS/DRA	RSA	RSA/DRA
Model 0	80.00	80.00	73.33	70.00
Model 1	80.00	80.00	73.33	70.00
Model 2	**83.33**	80.00	73.33	66.67
Model 3	**83.33**	**86.67**	**80.00**	**73.33**
Model 4	80.00	83.33	73.33	70.00
Model 5	**83.33**	80.00	73.33	66.67
Model 6	**83.33**	76.67	73.33	66.67
Model 7	**83.33**	76.67	73.33	70.00
Model 8	80.00	73.33	73.33	70.00
Model 9	80.00	76.67	70.00	70.00
Model 10	76.67	80.00	70.00	**73.33**
Model 11	80.00	76.67	66.67	66.67
Model 12	76.67	76.67	66.67	63.33

Table 3. Average recognition accuracy (%) of similar pronunciation phrases on white noise at 10 dB and 20 dB SNR

Models	DRA		CMS/DRA		RSA		RSA/DRA	
	10dB	20dB	10dB	20dB	10dB	20dB	10dB	20dB
Model 0	40.00	56.67	46.67	**73.33**	**63.33**	**66.67**	46.67	60.00
Model 1	40.00	56.67	50.00	**73.33**	60.00	63.33	**53.33**	**66.67**
Model 2	43.33	63.33	53.33	**73.33**	60.00	63.33	50.00	56.67
Model 3	**50.00**	63.33	53.33	**73.33**	60.00	60.00	43.33	60.00
Model 4	36.67	**80.00**	50.00	63.33	46.67	60.00	43.33	60.00
Model 5	**50.00**	60.00	56.67	60.00	46.67	63.33	50.00	63.33
Model 6	43.33	76.67	**60.00**	60.00	40.00	63.33	50.00	60.00
Model 7	43.33	73.33	**60.00**	66.67	43.33	63.33	**53.33**	63.33
Model 8	40.00	**80.00**	56.67	70.00	43.33	63.33	**53.33**	60.00
Model 9	40.00	73.33	53.33	66.67	43.33	63.33	**53.33**	63.33
Model 10	33.33	**80.00**	**60.00**	63.33	40.00	63.33	**53.33**	63.33
Model 11	40.00	**80.00**	**60.00**	66.67	43.33	63.33	50.00	63.33
Model 12	36.67	**80.00**	**60.00**	70.00	46.67	60.00	43.33	60.00

Table 4. Average recognition accuracy (%) of similar pronunciation phrases for 15 types of noise at 10 dB and 20 dB SNR

Models	DRA		CMS/DRA		RSA		RSA/DRA	
	10dB	20dB	10dB	20dB	10dB	20dB	10dB	20dB
Model 0	55.56	67.78	58.00	**74.00**	61.33	**71.56**	52.89	66.00
Model 1	56.45	70.67	58.00	73.33	61.78	**71.56**	**54.67**	**67.11**
Model 2	56.22	70.45	56.67	71.55	**62.64**	70.67	52.89	66.44
Model 3	**58.45**	68.67	**58.44**	73.56	60.89	70.67	53.33	66.44
Model 4	56.00	**72.00**	55.78	71.78	58.67	67.11	51.55	64.12
Model 5	57.78	69.56	55.56	71.78	58.00	67.11	51.56	65.33

Table 5. Average recognition accuracy (%) of 142 isolated words for 15 types of noise at 10 dB and 20 dB SNR

Models	DRA		CMS/DRA		RSA		RSA/DRA	
	10dB	20dB	10dB	20dB	10dB	20dB	10dB	20dB
Model 0	56.57	**80.09**	**72.70**	**87.62**	71.59	85.30	69.79	82.66
Model 1	**56.66**	80.08	72.53	87.52	71.36	85.35	**70.72**	**83.37**
Model 2	55.58	79.17	71.35	87.11	**72.57**	**87.43**	69.48	83.13

Based on results from Table 3, models 0 to 5 are chosen for further analysis using 15 types of noise from NOISEX-92 database. Experiments are conducted on 3 similar pronunciation phrases for models 0 to 5 and on 142 isolated words for model 0 to 2 respectively. Table 4 shows results for models 0 to 5 on 3 similar pronunciation phrases while results in Table 5 show the average recognition accuracy for models 0 to 2 on 142 isolated words. On clean speech, model 3 performs better at 83.33 % with DRA, 86.67 % with CMS/DRA, 80.00 % with RSA and 73.33 % with RSA/DRA compared with 80.00 %, 80.00 %, 73.33 % and 70.00 % obtained using model 0 under similar experimental conditions.

On 10 dB SNR white noise, models 2, 5, 6 and 7 perform well at 43.33 % 50.00 % 43.33 % and 43.33 % with DRA compared to 40.00 % obtained with model 0. In the same order, 53.33 %, 56.67 %, 60.00 % and 60.00 % recognition accuracy are obtained with CMS/DRA respectively, compared to 46.67 % of model 0. Further, recognition accuracy of 50.00 %, 50.00 %, 50.00 % and 53.33 % are obtained with RSA/DRA for the same models in the stated sequence order. The results are in comparison with 46.67 % obtained with model 0. As shown in Table 3, at 20 dB SNR white noise, models 2 to 12 performs well above average with DRA. However, all models under-performs with RSA.

Results from the experiments of similar pronunciation phrases on 15 types of noise, as shown in Table 4, indicate that model 3 performs better giving 58.45 % at 10 dB DRA compared with with model 0 at 55.56 %, while model 4 performs better at 20 dB DRA giving 72.00 % compared with model 0 at 67.78 %. From the same table, model 3 yield 58.44 % with CMS/DRA at 10 dB compared to 58.00 % of model 0. All models of the proposed approach perform below 74.00 % obtained using model 0 with CMS/DRA at 20 dB. Models 1 and 2 show recognition accuracy of 61.78 % and 62.64 % at 10 dB with RSA respectively compared to model 0 at 61.33 %. On the other hand, only model 1 yield similar results to model 0, other models performed below the baseline model results of 71.56 % at 20 dB RSA. Model 1 and model 3 yield 54.67 % and 53.33 % recognition accuracy, at 10 dB RSA/DRA respectively compared with model 0 at 52.89 %, while model 1 yield 67.11 %, and models 2 and 3 yield 66.44 % at 20 dB RSA/DRA compared with 66.00 % of model 0. Shown in Table 5 are results for 142 isolated words on 15 types of noise. Results show that model 1 yield the average recognition accuracy of 56.66 % at 10 dB DRA, 85.35 % at 20 dB RSA and 70.72 % and 83.37 % at 10 dB and 20 dB RSA/DRA respectively compared to 56.57 %, 85.30 %, 69.79 % and 82.66 % of model 0 respectively.

5. Conclusions

This paper presents the analysis of the results obtained for 3 similar pronunciation phrases and results for 142 isolated word recognition of independent male speakers. Results from the conventional and proposed approaches have been compared using different feature vector dimensions. Based on experimental results, it can be concluded that the performance of proposed approach depends on the influence of a particular noise and the effectiveness of the kind of noise reduction technique applied on such noise. The proposed approach is more effective at 10 dB SNR. The proposed approach equally demonstrates the effective use of inter-frame with intra-frame variations in speech recognition at high signal-to-noise ratio.

Acknowledgement

This study is supported in part by the Vehicle Information and Communication System Center (VICS Center).

References

[1] S. Yoshizawa, N. Hayasaka, N. Wada and Y. Miyanaga, "Central Amplitude Range Normalization for Noise Robust Speech Recognition," *IEICE Information and Systems*, pp.2130-2137, 2004.

[2] K. Takagi, H. Hattori and T. Watanabe, "Rapid environment adaptation for speech recognition," *Journal of the Acoustical Society of Japan (E)*, vol. 16, no. 5, pp.273-281, 1995.

[3] S. V. Vaseghi, Advanced Digital Signal Processing and Noise Reduction, John Wiley Sons Ltd, 2000.

[4] S. F. Boll, "Suppression of acoustic noise in speech using Spectral Subtraction," *IEEE Trans. Acoust. Speech Signal Process*, vol. ASSP-33, no. 27, pp.113-120.

[5] Y. Sun and Y. Miyanga, "A Noise-Robust Continuous Speech Recognition System Using Block-Based Dynamic Range Adjustment," *IEICE Transactions on Information and Systems*, vol. E95.D(2012), no. 3, pp.844-852, 2012.

[6] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proceedings of the IEEE*, vol. 12, pp.197-210, 1979.

[7] M. A. Abd El-Fattah, M. I. Dessouky, A. M. Abbas, S. M. Diab, S. M. El-Rabaie and F. E. Abd El-samie, "Speech enhancement with an adaptive Wiener filter," *International Journal Speech Technology*, 2014.

[8] Martin Westphal, "The use of Cepstral Means in Conversational Speech Recogntion," *Interactive Systems Laboratories*, University of Karlsruhe, 76128 Karlsruhe, Germany.

[9] K. Ohnuki, W. Takahashi, S. Yoshizawa and Y. Miyanaga, "Noise Robust Speech Features for Automatic Continuous Speech Recognition using Running Spectrum Analysis," in Proc. *International Symposium on Communications and Information Technologies*, 2008.

[10] G. Mufungulwa, A. Asheralieva, H. Tsutsui and Y. Miyanaga, "New Speech Features Based on time-varying LPC for Robust Automatic Speech Recognition," *IEICE Technical Report, SIS*, June, 2016.

[11] D. Sanjib, "Speech Recognition Technique: A Review," *International Journal of Enginering Research and Applications*, vol. 2, no. 3, pp.2071-2087, 2012.

[12] B. Anjali, K. Abhijeet, and B. Nidhika, "Voice Command Recognition System Based on MFCC and DTW," *International Journal of Engineering Science and Technology*, vol. 2 no. 12, pp.7335-7342, 2010.

[13] M. G. Hall, A. V. Oppenheim, and A. S. Willsky, "Timevaring Parametric Modeling of Speech," *Signal Processing*, vol. 5, pp.267-285, 1983.

[14] A. Varga, and H.J.M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech communication*, vol. 12, no. 3, pp.247-252, 1993.