An Efficient Block-based Dynamic Range Adjustment Method in Noise-robust Continuous Speech Recognition

Yiming SUN* and Constantin SIRITEANU* and Yoshikazu MIYANAGA*
* Information and Communication Network Laboratory Hokkaido University, Sapporo, 060-0814, Japan
E-mail: sunny@icn.ist.hokudai.ac.jp Tel: +81-117-06-6493

***Abstract*—This paper proposes a new technique for speech feature estimation under noise circumstances. This new approach yields noise-robust continuous speech recognition (CSR). Noise-robust techniques for isolated word speech recognition typically employ the running spectrum analysis (RSA), the running spectrum filtering (RSF) and the dynamic range adjustment (DRA) methods. Among them, only RSA has been applied into a CSR system. However, we propose an enhanced DRA for a noise-robust CSR system. Thus, in the speech recognition stage, the continuous speech waveform is automatically divided into short blocks and DRA is applied to these blocks. We find that the proposed method improves recognition performance under several different noise and SNR conditions.**

I. Introduction

There has been much effort devoted to improving recognition rates for continuous speech recognition (CSR) systems [1]. In recent years, CSR has made great progress to yield high recognition rates in clean conditions. However, current technology has not matured enough to yield high performance in noisy environments.

There are several approaches to obtain continuous speech feature vectors, namely mel-frequency cepstral coefficients (MFCC), linear predictive cepstral coefficients (LPCC) [2], and perceptually-based linear predictive coefficients (PLPs) [3]. For this paper, we have selected the MFCC to extract the speech feature vectors. Then, CSR is defined as the process and related technology for converting signals into a sequence of phones or words. Herein, we use cepstrum mean substraction (CMS) and running spectrum analysis (RSA) [4] to optimize the MFCC speech feature vectors. Finally, we use dynamic range adjustment (DRA) to adjust the MFCC speech feature vectors.

Although RASTA is a well known method focusing on modulation spectrum domain (MSD), a primary RASTA employs IIR filtering and it may cause a problem such as phase distortion [5]. RSF is based on a FIR filter. RSA is directly used in the MSD. Compared with RASTA and RSF, RSA can realize an ideal processing [6]. In this paper, we select RSA to reduce any noise effects on the MSD.

Among the noise robust methods used in a CSR system, the method of RSA and CMS has been developed in [7] and it can show a little higher performance than others. The RSA and CMS are used for the reduction of distortion embedded into a training data set and the CMS is also used for the time invariant noise reduction to an observed speech waveform in a recognition stage. By using the above noise robust techniques, the recognition accuracy can be improved. However, compared with the results of isolated speech recognition accuracy, its performance is insufficient for almost any actual application.

In this paper, we propose a new algorithm to adjust the speech feature vectors for recognition. If we divide the continuous speech feature vectors into blocks according to some conditions, we can use different maxima for normalization in different blocks. Until now, little attention has been paid to block-based adjustment of the speech feature vectors. Therefore, this paper proposes a noise robust speech recognition approach using RSA and DRA for modeling as well as block-based DRA for recognition.

II. Methods

A. CMS

CMS is a channel normalization approach to compensate for the acoustic channel [8]. Time-invariant channel parameters in a recording system and convolutional disturbance noise are evaluated by CMS and reduced from the observed speech waveform. CMS improves the distortion between training speech data and observed speech data for recognition.

B. RSA

RSA is applied for both low and high frequency components in the MSD. These components in MSD are then processed using RSA [9]. Reduction of low frequency components has the same effect as the CMS technique. In addition, reduction of high frequency components results in the elimination on time-varying noises which are not be created by human speech production models.

C. DRA

The DRA strategy tends to decrease the variability of noise feature vectors. DRA adjusts the dynamic range by normalizing the amplitude of speech feature vectors. In DRA, each MFCC is adjusted in proportion to its maximum amplitude after all RSA speech frames. If we define the i-th component of the MFCC vector $\mathbf{c}_k$ as $c_{k,i}$ after CMS, the DRA algorithm calculates the following new value:

$$c'_{k,i} = \frac{c_{k,i}}{\max_{j=1,\dots,M} |c_{j,i}|} \tag{1}$$

where $c'_{k,i}$ denotes the i-th element of the post-DRA MFCC feature vector.

III. Noise Corruption and Removal Algorithm

A. Noise Corruption

Although only a clean continuous speech can be observed, the selection of each word and the dynamic rage adjustment for the selected word are not difficult. However, under noisy conditions, the selection of words may be difficult issue. In this paper, the following two step processing is considered.

(1) From an observed noisy continuous speech waveform, all short sentences are selected.

(2) A short sentence is divided into several blocks and then each block is independently applied by DRA.

The above processing is applied to an observed unknown continuous speech in recognition.

In the first step (1), Non-speech parts are eliminated. A continuous speech has many Non-speech parts and only noises. These parts effects DRA inappropriately. In the second step (2), the unbalance of several dynamic ranges existed in a continuous speech can be compensated.

B. First Step: Separating Blocks

In this paper, the proposed algorithm identifies a block between the zero-crossings of $c_{j,i}$ in a short sentence. The definition of the block is a part between the zero-crossing points in the trajectory of $c_{j,i}$. In a different block, we use different maximum value to calculate the $c'_{j,i}$ from $c_{j,i}$ by DRA.

In order to adjust $c_{j,i}$, the trajectory of $c_{j,i}$ given in a sentence is divided into some blocks according to the the zero-crossing points. we search the zero-crossing points in $c_{j,i}$ by the equation:

$$f_{j,i} = c_{j,i-1}/c_{j,i}. \tag{2}$$

If $f_{j,i} < 0$, we consider there must be a zero-crossing point between $c_{j,i-1}$ and $c_{j,i}$.

We define P_0^j as $\max|c_{j,i}|$ in a short sentence. Then we record P_0^j location as $L_j(P_0)$. We subtract and add L_m frames from $L_j(P_0)$ backward and frontward, and record the location as $\hat{L}_j(P_{-1}) = L_j(P_0) - L_m$ and $\hat{L}_j(P_1) = L_j(P_0) + L_m$. Next we search the zero-crossing points nearest to $\hat{L}_j(P_{-1})$ and $\hat{L}_j(P_1)$ by Eq. (2). Once we find the zero-crossing points by Eq. (2), we define them as the $L_j(P_{-1})$ and $L_j(P_1)$ as the locations of the zero-crossing points in this block.

From $L_j(P_{-1})$ to $L_j(P_1)$, we can get a block. The block divides the trajectory of $c_{j,i}$ into three segments. The middle segment is called the main block. The range of the main block is from a start-point as $L_j(P_{-1})$ to the end-point as $L_j(P_1)$. There is no changes of relations among elements in the main block whether we use block-based DRA or original DRA. We focus on adjust the relations among elements on both left and right sides of the main block.

There are numerous zero-crossing points in a short sentence due to noise. Furthermore, the noise caused some abrupt changes between zero-crossing points. We consider some limitations to select the zero-crossing points of blocks. The limitations focus on preserving the continuity of the $c_{j,i}$ in zero-crossing points.

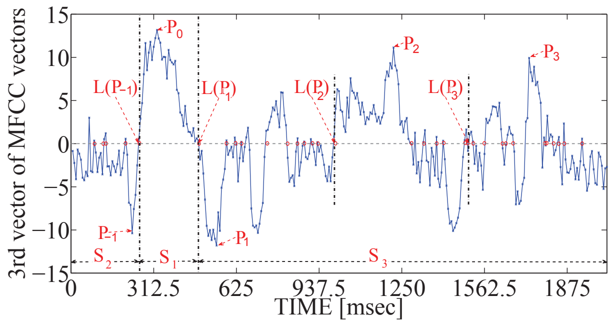

Fig. 1: An example of $c_{j,i} (j = 3)$ for separating blocks and determining maximum

If $|c_{j,i-1}| < 2$ or $|c_{j,i}| < 2$, it means a smooth variation near the zero-crossing points. Otherwise, there is a discontinuity between $|c_{j,i-1}|$ and $|c_{j,i}|$. In a word, the zero-crossing points used in a short sentence are selected under the limitations: $|c_{j,i-1}| < 2$ or $|c_{j,i}| < 2$.

We continue to divide the other two segments into blocks. The block shortest length is defined as $L_j(P_i) - L_j(P_{i+1}) > L_w$. Nearest to $L_j(P_{i+1})$, we use Eq. (2) to search zero-crossing points which satisfy the limitations. Then, we set $i = \pm i \pm 1$ to search next block boundary. Symbol $\pm i$ is the $\pm i$th block whose boundary satisfies the above limitations.

From the above selection, we can get all zero-crossing points which give the block boundaries. They are given as $L_j(P_{-N})$, $L_j(P_{1-N})$, $L_j(P_{2-N})$, ... , $L_j(P_{-1})$, $L_j(P_1)$, ... , $L_j(P_M)$. The main block is given from $L_j(P_{-1})$ to $L_j(P_1)$. In the left hand side, the $-i$th block is given from $L_j(P_{-i-1})$ to $L_j(P_{-i})$. In the right hand side, the ith block is given from $L_j(P_i)$ to $L_j(P_{i+1})$.

C. Second Step: Determining Maximum

There are many peaks in the trajectory of $c_{j,i}$ within a short sentence. From the first step, we have found that each block includes one or more peaks. All the peak parts of $c_{j,i}$ exhibit stronger speech features than those around zero parts of $c_{j,i}$ in noisy conditions.

In the above step, P_0^j is defined as the maximum of main block. The main block divides $c_{j,i}$ into right-hand side and left-hand side. Then, we define $P_{\pm i}^j$ as the $\max|c_{j,i}|$ within $\pm i$th block in each right-hand side block and each left-hand side block, respectively. By using different $P_{\pm i}^j$, we can normalize $\pm i$th block. Thus, all the peaks $(P_{\pm i}^j)$ are enhanced in each block. These enhanced peaks can improve the recognition rate.

Fig. 1 shows an example of block diagram, for $j = 3$, for separating blocks and determining the maxima. Therein, two longer vertical dot-lines show the boundary of the main block. Further, S_1, S_2 and S_3 indicate the main block, left-hand side block and right-hand side blocks, respectively. The shorter vertical dot-lines show the boundary of a block in S_3. The maxima in different blocks are given by P_{-1}, P_0, P_1, P_2 and P_3. We can use these peak values to improve the recognition rate.

D. Third Step: Noise Coefficient Addition

Different noise level cause different corruption degree for $c_{j,i}$. We define S_{SNR} as the noise level coefficient. S_{10}, S_{15} and S_{20} correspond to 10 dB, 15 dB and 20 dB respectively. By simulations, we set $S_{10} = 0.1$, $S_{15} = -0.1$ and $S_{20} = -0.8$.

E. Final Step: Using Block-based DRA

We have obtained blocks and determined maxima and noise level coefficients. In each identified block, we substitute is corresponding maximum and noise level coefficient in (1), which can be represented by Eq. (3).

$$c'_{k,i} = \frac{c_{k,i}}{P^j_{\pm i} + S_{SNR}}, \quad (3)$$

IV. Experiments

In the training of HMM [10] [11], all sentences are assumed to be recorded under clean or low noise situation. In other words, any time varying noises and high level noises are not considered in this training stage. From these reasons, conventional CMS, RSA and DRA are applied to all given training speech data set.

A. Model Building

Even when the speech data sets for the training are recorded under low noise circumstances, the effect of convolution disturbance, i.e., microphone, may influence speech features. During the training stage for HMMs, CMS, RSA and DRA should be used where conventional systems have employed only CMS and CMS/RSA.

As the merit of RSA, the un-speech feature over 15 Hz on MSD can be more accurately reduced than RSF method. In addition, using RSA with CMS, the high accurate noise and disturbance components can be eliminated effectively.

The effects of CMS and RSF are not small for the dynamic range of speech feature trajectory mentioned in the previous section. The conventional DRA is applied to them for the dynamic rage normalization of their estimated and processed speech features.

The benefits of normalizing the feature vectors by DRA in the entire sentence are two-fold: it significantly reduces noise corruption and it preserves important speech characteristics. The speech sound condition for model building is shown in Table I.

B. Blocks in Recognition

In training, in order to get noise-robust model, we use CMS, RSA and DRA to process the speech feature vectors. For recognition, we use only block-based DRA algorithm without the RSA method. The condition is showed as Table II.

In the speech recognition stage, the block-based DRA is applied. In the speech recognition, we do not know any time range for observed speech and thus it is impossible to known the length of speech waveform as a prior information. In addition, during the speech recording, some different noises and disturbances may happen. For the accurate noise and disturbance reduction, the proposed block-based DRA is applied.

In addition, CMS is also applied to the estimate of speech features. However, RSA by which the speech features over 15 Hz in MSD are reduced is not applied to the speech recognition stage. In order to improve the speech recognition performance more, the detail speech features of observed waveform are used in its stage.

In Table III, the mean lengths of all long vowels are less than 15. We can calculate out the main block width more than $2L_m$ from Section III-B. If we set $L_m = 15$, the main block width includes at least one long vowel.

Block width is determined by L_w, which also changes relations among elements. Small L_w causes substantial changes of relations among elements and leads to an abrupt decrease is recognition rate. On the other hand, large L_w causes no changes of relations among elements and leads to the recognition rates near those of conventional DRA. We set $L_w = 80$ in simulations.

The cepstral variance normalization (CVN) technique normalizes the feature variance to a same scale. The cepstral mean normalization (CMN) and CVN are usually used in cascade to form the mean and variance normalization (MVN) to normalize the features. We show all the results for comparison in the Table V and Table VI.

TABLE I: Acoustic analysis conditions

Sampling frequency	16 kHz
Frame shift	10.0 ms
Frame length	25.0 ms
Window type	Hanning
Training data	23651 sentences from 153 people
Emphasizing of High Frequency	$1 - 0.97z^{-1}$
HMM state number	5 states (include start and end states)
Number of Gaussian Mixtures	16
Clustering	about 2000 states

TABLE II: Recognition conditions

Known data for testing	50 sentences from 12 people
Unknown data for testing	180 sentences from 6 people
Sampling and frame conditions	the same with Table I

TABLE III: Long vowel phoneme frame average length [%]

Phoneme	Means	Variance	Appear Times
a:	13.35	13.50	2054
e:	14.46	15.59	12688
i:	14.93	20.97	1724
o:	13.83	19.01	37657
u:	10.64	17.50	4831

TABLE IV: Noise Kinds

Noise Name	Noise Name	Noise Name
babble	buccaneer1	buccaneer2
destroyerenginer	destroyerops	f16
factory1	factory2	hfchannel
leopard	m109	machinegun
pink	volvo	white

TABLE V: Recognition rates for clean conditions [%]

	Proposed		RSA		MVN	
	Corr	Acc	Corr	Acc	Corr	Acc
known data	92.55	91.49	91.89	90.69	91.26	90.49
unknown data	82.99	81.56	80.58	79.00	82.88	81.60

V. Results

In this experiments, all HMM have been trained by using JNAS database [12]. It is produced by 153 males' native Japanese speakers.

We use two measures for the performance of speech recognition:

$$R_C = \frac{N - S - D}{N} \times 100 \quad [\%], \tag{4}$$

$$R_A = \frac{N - S - D - I}{N} \times 100 \quad [\%], \tag{5}$$

where N is the total number of words in the set of speech sentences, S is the number of misrecognized words, D denotes the number of words which are not selected as words by the system, I denotes the number of words which are misrecognized as words, i.e., noise components and non-speech sounds. Above, R_C shows the correct word recognition rate for the entire set of speech words, and R_A shows the accuracy of the total CSR performance.

We simulated all data not only in clean conditions but also for various noise types for several SNR values. We have added 15 kinds of noise for testing shown by Table IV. In all result tables, the 'Proposed' column denotes the method that used CMS, RSA and DRA for modeling, and both CMS and block-based DRA algorithm for testing. The 'RSA' column denotes the method that used CMS and RSA for modeling, and CMS for testing. The 'MVN' column denotes the case when RSA and MVN was used for modeling, and MVN for testing. Table V shows the results in the clean conditions. Table VI shows the average results in different SNR conditions.

VI. Conclusion

A new block-based DRA algorithm has been implemented for unspecific speaker recognition. The proposed method has enhanced the recognition rate under lower SNR noise environments. The DRA normalizes the maximum amplitudes of MFCC in each selected block. The proposed CSR system yields higher accuracy than conventional systems under 20, 15 and 10 dB noise environments.

TABLE VI: Average recognition rates in noisy conditions [%]

	SNR	Proposed		RSA		MVN	
		Corr	Acc	Corr	Acc	Corr	Acc
known data for recognition	20 dB	80.08	77.72	77.80	75.82	78.25	76.05
	15 dB	68.06	64.81	61.10	58.40	63.32	59.82
	10 dB	49.85	46.27	39.23	36.98	46.15	41.86
unknown data for recognition	20 dB	73.76	71.31	72.46	70.23	73.08	70.63
	15 dB	63.01	60.14	58.18	55.95	59.60	56.64
	10 dB	47.85	44.75	37.06	35.19	42.87	40.18

VII. Acknowledgement

The authors would like to thank the global COE program, Graduate School of Information Science and Technology, Hokkaido University for fruitful discussions. This study is supported in parts by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Scientific Research (B2) (20300014).

References

[1] I. Koji, S. Takahiro, and F. Sadaoki, "Noise robust speech recognition using F0 contour information," in *Institute of Electronics, Information, and communication Engineers. IEICE*, no. 5, May 2004, pp. 1102–1109.

[2] M. Islam, M. Rahman, and M. Khan, "Improvement of speech enhancement techniques for robust speaker identification in noise," in *International Conference on Computers and Information Technology, ICCIT.*, Dec. 2009, pp. 255–260.

[3] M. Goyani, N. Dave, and N. Patel, "Performance analysis of lip synchronization using LPC, MFCC and PLP speech parameters," in *International Conference on Computational Intelligence and Communication Networks (CICN)*, Nov. 2010, pp. 582–587.

[4] K. Ohnuki, "The research about robust acoustic modeling and continuous speech recognition system," Ph.D. dissertation, Hokkaido University, Japan, 2009.

[5] S. Yoshizawa and Y. Miyanaga, "Robust recognition of noisy speech and its hardware design for real time processing," *ECTI TRANSACTIONS ON ELECTRICAL ENG., ELECTRONICS, AND COMMUNICATIONS*, vol. 3, no. 1, pp. 36–43, Feb. 2005.

[6] N. OHTSUKI, Y. UCHIKAWA, and Y. MIYANAGA, "Speech noise reduction using high-precision RSA," in *IEICE Technical Report*, Jun. 2008, pp. 1–4.

[7] K. Ohnuki, W. Takahashi, S. Yoshizawa, and Y. Miyanaga, "New acoustic modeling for robust recognition and its speech recognition system," in *International Conference on Embedded Systems and Intelligent Technology, ICESIT.*, Jan. 2009, pp. 1–4.

[8] S. V. and J. W, Eds., *Advanced Digital Signal Processing and Noise Reduction, Second Edition*. SONS, LTD, 2000.

[9] K. Ohnuki, W. Takahashi, and Y. Miyanaga, "Noise robust speech features for automatic continuous speech recognition using running spectrum analysis," in *International Symposium on Communications and Information Technologies, ISCIT.*, 2008, pp. 150–153.

[10] L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989.

[11] P. Banerjee, G. Garg, P. Mitra, and A. Basu, "Application of triphone clustering in acoustic modeling for continuous speech recognition in Bengali," in *International Conference on Pattern Recognition, ICPR.*, Dec. 2008, pp. 1–4.

[12] K. Itou, M. Yamamoto, K. Takeda, Matsuoka, K. Shikano, and S. Tahashi, "Japanese speech corpus for large vocabulary continuous speech recognition research," Journal of the Acoustical Society of Japan, Tokyo, Japan, Tech. Rep., 1999.