Enhanced Running Spectrum Analysis for Robust Speech Recognition Under Adverse Conditions: Case Study on Japanese Speech

George Mufungulwa[1], **Hiroshi Tsutsui**[2],
Yoshikazu Miyanaga[3], and **Shin-ichi Abe**[4], Non-members

ABSTRACT

In real environment, many noises degrade the performance of Automatic Speech Recognition (ASR) systems. In addition, in case of similar pronunciations, it is not easy to realize high accuracy of recognition rate. From this point of view, our work envisaged an enhanced processing algorithm into speech modulation spectrum as Running Spectrum Analysis (RSA). It is also adequately applied to observed speech data. In the envisaged method, a modulation spectrum filtering (MSF) method directly modifies the observed cepstral modulation spectrum by Fourier transform of the cepstral time frequency. The method and experiments carried out for various passbands had favorable results that showed the improvement of about 1-4 % recognition accuracy as compared with current conventional methods.

Keywords: MFCC, HMM, ASR, RSF, RSA

1. INTRODUCTION

The fundamental stages in speech recognition are speech feature extraction and feature matching. Various speech features, including ones from linear prediction coding (LPC) [1-4], time-varying linear prediction coding (TVLPC) [5], mel frequency cepstral coefficients (MFCC) [6-9] among others, have been used to model speech recognition either singularly or collectively in improving speech recognition accuracies. MFCC, which is based on spectral content of the signal and can be considered as one of the standard method for feature extraction [10] is opted for use in our study.

Speech recognition systems often suffer from multiple sources of variability due to corrupted speech signal features [11]. In compensating for distortions, most speech recognizers use normalization methods and noise filtering techniques in conjunction with voice activity detection (VAD) techniques. Improved accuracy in noise robust speech recognition can be realized by processing speech using running spectrum filtering (RSF)[12, 13], for example. The downside, is high computation costs and high demand on memory.

In recent past, several typical methods relating to the use of modulation spectrum features for noisy speech recognition have been developed [14–16]. Running spectrum analysis (RSA) is not only an effective technique for reduction of noise on the modulation spectrum domain (MSD)[17] but it can also be deployed to realize ideal processing [18].

Although running spectrum analysis (RSA) is a well known method focusing on modulation spectrum, it has mostly been applied for automatic continuous speech recognition [19]. Furthermore, in speech communication, its application has been mainly focused on frequency components in the range of 2-8 Hz because this range contains the dominant components of the amplitude envelope of speech [20][21]. Modulation frequency band higher than 8 Hz can be regarded as miscellaneous noise components or such unnecessary speech components for recognition as speaker's characteristics such as tone, pronunciation, etc [22].

However, this work presented a novel noise-robust feature extraction framework that leveraged the technique of RSA on isolated phrase recognition. This work was envisaged with the goal to enhance RSA for the purpose of achieving higher recognition accuracy for both male and female, similar and non-similar pronunciation Japanese speech phrases under noisy conditions. Robust speech features realized using this method can be required in many applications, including modelling for analysis/synthesis and recognition of isolated utterances with "Listen/Not-Listen" states. Situations in which this method can be applied include tasks that require human machine interface such as automatic call processing in telephone networks and query based information systems such as voice dictation, stock price quotations, [23] among others. Authors assume that the proposed method performance relates with gender just as recognition accuracy can be influenced by the signal-to-noise ra-

Manuscript received on April 3, 2017 ; revised on May 12, 2017.

Final manuscript received on June 6, 2017.

[1,2,3] The authors are with Graduate School of Information Science and Technology, Hokkaido University, Kita 14, Nishi 9, Kita-ku, Sapporo 060-0814, Japan., E-mail: mufungulwac@gmail.com, hiroshi.tsutsui@ist.hokudai.ac.jp and miya@ist.hokudai.ac.jp

[4] The author is with Vehicle Information and Communication System Center (VICS Center), Nittochi Kyobashi Bldg., 8F, 2-5-7 Kyobashi, Chuo-ku, Tokyo, 104-0031, Japan., E-mail: s-abe@vics.or.jp

tio (SNR) which the authors aim to ascertain.

In this study, the work applied running spectrum analysis (RSA) on modulation spectrum for noise robust speech recognition of adequately selected frequency components. The noise effect was dealt with filtering the range of frequency components, 1-7 Hz, 1-15 Hz, 1-35 Hz and 1-40 Hz in the modulation spectrum domain. Further, it is argued that the expected speech recognition accuracy can be improved when modulation spectrum filtering (MSF) directly modify the cepstral modulation spectrum (CMS) [16] which is specifically referred to as the Fourier transform of the cepstral time sequence.

Although hidden Markov modelling (HMM) based approaches require training in automatic speech recognition (ASR) systems, the HMM method has been widely used. Since there are several noise reduction methods and speech enhancement methods against any noises, almost all of ASR systems using HMM and noise reduction can show higher accuracy of speech recognition rate than that given by a conventional standard HMM based ASR.

The rest of the paper is organized as follows. In Section 2, the proposed system is explained. In Section 3, performance of proposed method is evaluated. In the same section, experimental conditions are explained and the results stated. Section 4 discusses the results and in Section 5 which is the conclusion compares the enhanced RSA over the RSF.

2. PROPOSED SYSTEM

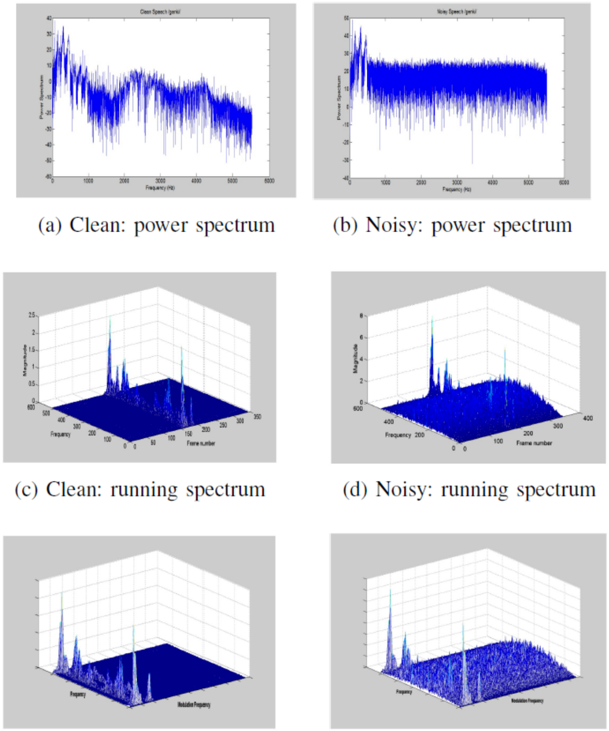
The motivation of this study is to evaluate the effectiveness of the enhanced running spectrum analysis (RSA), which is explained later, as it compares with running spectrum filtering (RSF). RSA is the processing of speech over modulation spectrum domain. Linguistically dominant factors of the speech signal may occupy different parts of the modulation spectrum than do some non-linguistics factors such as steady additive noise [24]. A proper processing of modulation spectrum of speech may improve quality of noisy speech. Investigations on possibilities of the modulation spectrum domain for enhancement of noisy speech [25][26] support the dominance of modulation spectrum components in the vicinity of 2-8 Hz in speech communication.

We now explain the effect of noise in running and modulation spectrum domains.

For standard speech information processing, the frame concept has been applied. The 256 sample point length frame is first defined and using this frame, a short time speech waveform is extracted. For the short time speech waveform, a speech power spectrum is calculated as a typical speech analysis. The frame is shifted with 128 points and then many short time speech waveforms can be obtained. Running spectrum is defined as the time trajectory in frequency domain. It consists of many speech power spectra given from short time frames. The modulation spectrum is defined as the spectrum in time varying of short-time running spectrum.

Figures 1(a) and 1(b) show the power spectra of clean speech and speech with additive white noise at 10 dB SNR for a Japanese phrase /genki/. Both spectra are calculated from short time speech waveforms. These figures indicate that the dynamic range on a power spectrum of a noisy speech is smaller than that of a clean spectrum. In addition, some of the power spectrum characteristics are unobservable under noisy conditions. Figure 1(c) shows the running spectrum of clean speech while Figure 1(d) shows the running spectrum of noisy speech of the same phrase /genki/. There are three axes, i.e., frequency axis, frame number axis and power amplitude axis.

When we observe the data on the frame number axis, the frequency is fixed to a specific value, its data can be recognized in the time domain. They can be applied by using fast Fourier transform (FFT). After such FFT is applied to all frequencies, we can get new 3-d data in the modulation spectrum domain. Modulation spectrum of the noisy signal is shown in Figure 1(f) and the modulation spectrum of the clean speech is shown in Figure 1(e).



(a) Clean: power spectrum

(b) Noisy: power spectrum

(c) Clean: running spectrum

(d) Noisy: running spectrum

(e) Clean: modulation spectrum with RSF

(f) Noisy: modulation spectrum with RSF

Fig.1: *Power spectra of (a) clean speech, and (b) noisy speech phrase /genki/ with white noise at 10 dB SNR. Running spectrum of (c) clean speech and (d) noisy speech. Modulation spectrum of (e) clean speech and (f) noisy speech with RSF*

Figure 2 shows the proposed system for which results and analysis are presented in Section 3. The left side of the figure shows the processes for male speakers while the right side of the same figure shows processes for female speakers. For each gender case, two output models for similar pronunciation (SP) and non similar pronunciation (NSP) respectively are realized. In the proposed system, there are four different kinds of filtering in RSA. The optimal filtering of RSA is applied for male and female speakers, SP and NSP.

In Figure 2, noisy speech at different signal-to-noise ratio (SNR) is input into a short-term energy (STE) based VAD for the purpose of retaining speech segments with sufficient energy while eliminating segments classified as noisy as well as silent. As in the case of training, the speech features are extracted using the standard MFCC as spectral analysis. A HMM based automatic speech recognition (ASR) system is utilized for testing. The gender of speaker (male or female) as well as the speech type, SP or NSP for each gender case are decided. This process results in four outputs; male SP, male NSP, female SP and female NSP, respectively. For each gender and speech type combination, the speech signal is passed through a voice activity detection (VAD) process in order to retain segments with speech activity or segments with high energy while eliminating segments with background noise or the ones with less energy prior to feature extraction.

Figure 3 shows the feature extraction process using fast Fourier transform (FFT) based MFCC with running spectrum filtering (RSF) for log spectra as a noise reduction technique.

In figure 3, it is shown that in order to obtain mel cepstrum, speech data is initially pre-emphasized and the pre-emphasized speech waveform in time domain is frame-blocked and windowed with a predefined analysis window. Later, fast Fourier transform (FFT) is computed. The magnitude of the output is then weighted by a series of mel filter frequency responses whose center frequencies and bandwidth roughly match those of auditory critical band filters [27]. The FFT bins are later combined so that each filter has unit weight. From the weighted sums of all amplitudes of signals, a vector is obtained by logarithmic amplitude compression computation. RSF is then applied before transforming the result to MFCC parameter by discrete cosine transform (DCT).

The performance of most if not all speech/audio processing methods is crucially dependent on the robustness of the extracted speech features. The accuracy of automatic speech recognition remains one of the important research challenges [23]. Most current feature extraction methods are still vulnerable against certain noises such as car noise [28].

Figure 4 shows the MFCC feature extraction process with running spectrum analysis (RSA). After spectral analysis, RSA is applied to realize the modulation spectrum. After which stage the process is as explained under feature extraction with RSF. In both cases, the features are trained into HMM, respectively.

In this paper, different types of enhanced RSA were selected for male and female speakers under noisy conditions.

During our preliminary study, among the RSA type (c) and type (f) were found to be better performers for male NSP and for SP respectively. Our study have also shown that, for example, in the case of female NSP, RSA with type (h) is better performer at high noise while type (c) and type (d) perform better at low noise. Similarly, for female SP, RSA with type (c) and type (h) were found to be better performers at high noise while type (d) performed better at less noise, respectively. The candidate of results with male or female speech are selected based on the maximum likelihood of HMM. Under noisy conditions different types of RSA show different performance for male and female speakers.

The proposed RSA differs from the one discussed in [19], for example. The former focuses on modulation frequency range of 2-8 Hz. However, in this study we evaluate the performance of several RSA types shown in Table 1. Table 1 shows 8 RSA passband specifications whose different sets of values are given as examples of filtering. In the modulation spectrum, it is possible to see the frequency range of the power concentration for each phrase and thereby help to decide which RSA type is most suitable. Each passband has a low cut-off frequency (LCF), and a high cut-off frequency (HCF). The difference between the two frequencies represents the number of frequency components over the modulation spectrum domain that are to be processed. In this way, we aim to determine the performance of new RSA over that of RSF by changing parameters such as; i) the number of frequency components (7, 15, 30, or 40 components), ii) the type of speaker (male or female), and iii) the signal-to-noise ratio (SNR) (10 dB, 15 dB, or 20 dB).

Table 1: *RSA passband specifications*

RSA Type	LCF (Hz)	HCF (Hz)
(a)	1	7
(b)	1	15
(c)	1	35
(d)	1	40
(e)	0.5	7
(f)	0.5	35
(g)	0.1	7
(h)	0.1	35

3. EXPERIMENTAL RESULTS

3.1 Objectives of the Experiments

The first objective of the experiments is to compare the performance of the proposed enhanced RSA

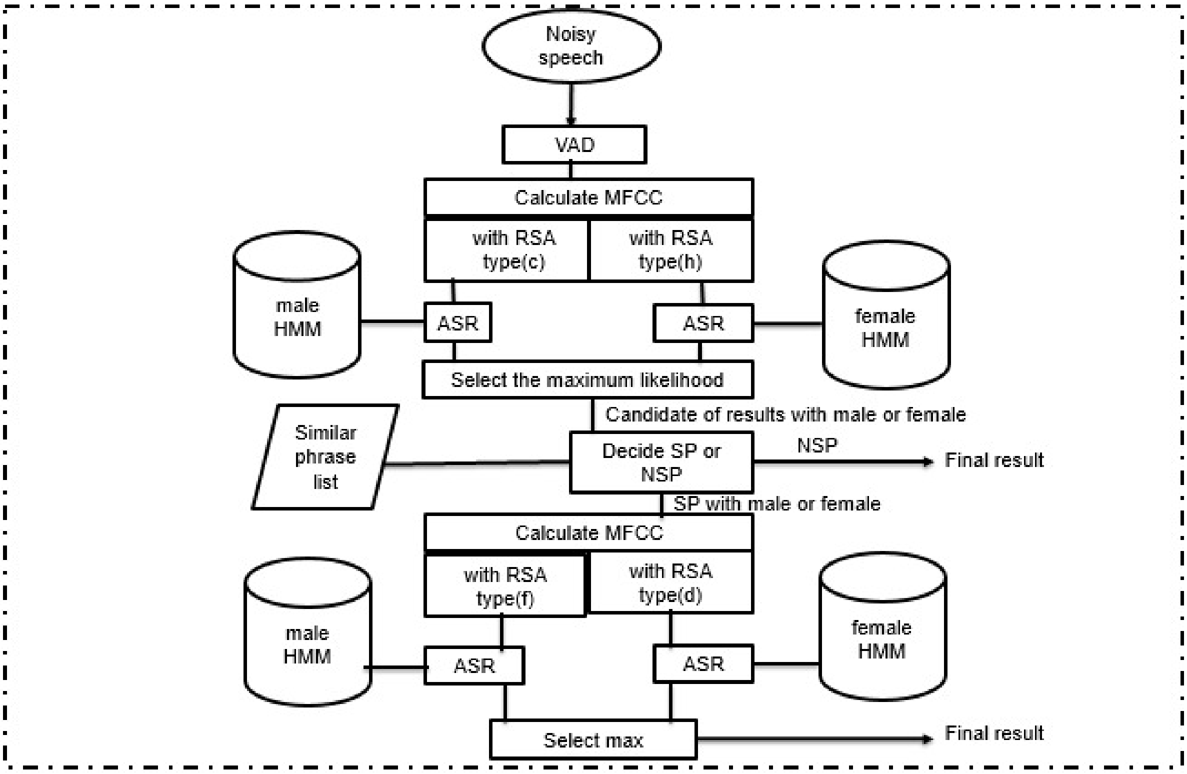


Fig.2: *Proposed system.*

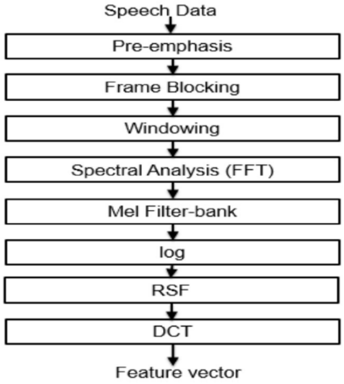


Fig.3: *Feature extraction with RSF.*

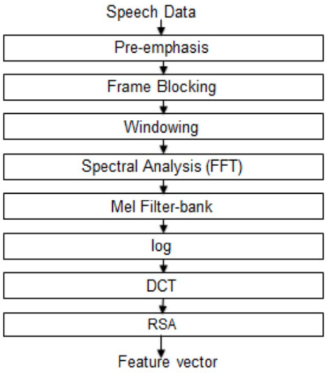


Fig.4: *Feature extraction with RSA.*

to that of RSF on similar and non-similar Japanese pronunciation phrases. The second objective is to evaluate how the performance relates to gender. The main method used for speech enhancement is filtering. We have evaluated the adaptability of our proposed RSA over modulation spectrum and compared its results to those of RSF. In this study, RSF is employed to act as the basis for comparing the tendency and to determine better performing RSA types at the given SNR for both gender.

3.2 Simulation parameters and conditions of experiments

Table 2 shows the simulation parameters.

Training sets of 30 male speakers and 30 female speakers, each speaker uttering 6 similar phrases and 100 Japanese common phrases, respectively, and each phrase repeated 3 times, are used for the front-end feature extraction and 32-states isolated phrase hidden Markov modeling (HMM) in training. Testing sets consisting of 10 male speakers and 10 female speakers (not used in training), with each speaker uttering 6 similar phrases and 100 Japanese common phrases and each phrase repeated 3 times respectively are utilized.

The speech sample is 11.025 KHz and 16-bit quantization. Frame-by-frame, 38-dimensional FFT based MFCC feature vectors are extracted after pre-emphasis and Hanning windowing. In the testing

Table 2: *The condition of speech recognition experiments*

Parameter name	Parameter value/type
Sampling	11.025 kHz (16-bit)
Frame length	23.2 ms (256 samples)
Shift length	11.6 ms (128 samples)
Pre emphasis	$1{-}0.97z^{-1}$
Windowing	Hanning window
Speech Feature vectors	$b_i(i=1,\ldots,12)$ $\Delta b_i(i=0,\ldots,12)$, $\Delta^2 b_i(i=0,\ldots,12)$,
Training Set	30 male , 30 female 3 utterances each
Tested Set	10 male, 10 female, 3 utterance each
Acoustic Model	32-states isolated phrase HMMs
Noise varieties	4 types from NOISEX-92 (white,pink, HF radio channel, babble)
SNR	10 dB, 15 dB, 20 dB
Filtering methods	RSF, RSA,

stage, 10 dB, 15 dB, and 20 dB of the 4 types of noises are artificially added to the original speech. We compare the performance of proposed enhanced RSA of specified passbands to those by RSF under 4 types of noises; white, pink, HF channel and babble noises in MATLAB (R2014a) software. Under the stated conditions, we measure the average recognition rates for 10 speakers on RSF and 8 enhanced RSA passband specifications given as Types (a) to Type (h) at 10 dB 15 dB, and 20 dB SNR.

Table 3 shows the average recognition accuracy for 100 Japanese common male speech phrases. Table 4 shows the average recognition accuracy for Japanese similar pronunciation male speech phrases. Table 5 shows the average recognition accuracy for 100 Japanese common female speech phrases. Table 6 shows the average recognition accuracy for Japanese similar pronunciation female speech phrases.

3.3 Simulation results and analysis

Analysis is carried out for the Japanese common and similar phrases databases. We use gender (male and female) and 4 SNR (at 10 dB, 15 dB, and 20 dB) as variables. Results analysis focuses on the performance of the enhanced RSA types on the various acoustic measures. The 4 kinds of noises used in the experiments are based on Signal Processing Information Base (SPIB) noise data measured in field by Speech Research Unit (SRU) at Institute for Perception-TNO, Netherlands, United Kingdom, under the project number 2589-SAM (Feb. 1990) In this paper the model formulation is as follows: the model uses FFT based MFCC coefficients consisting of 38-dimensional feature vectors. The 38-parameter feature vector consisting of 12 cepstral coefficients (without the zero-order coefficient) plus the corresponding 13 delta and 13 acceleration coefficients is given by $[b_1 b_2 \ldots b_{12} \Delta b_0 \Delta b_1 \ldots \Delta b_{12} \Delta^2 b_0 \Delta^2 b_1 \ldots \Delta^2 b_{12}]$ where b_i, Δb_i and $\Delta^2 b_i$, are MFCC, delta MFCC and delta-delta MFCC, respectively.

3.4 Results Explanations

In Table 3 at 10 dB SNR, RSA with type (c) performs better (76.6 %) compared to RSF (72.5 %). At 15 dB SNR, RSA with type (c) performs better (90.1 %) compared to RSF (87.6 %). RSA with type (c) performs better (94.9 %) than RSF (92.8 %) at 20 dB SNR.

RSA with type (c) (1 35) performs better than RSA with type (a). For RSA with type (c), the recognition accuracy results decline (from 76.6 % to 72.6 % for type (c) and type (f) and (h), respectively) with increase in bandwidth (for (c)(1 35), (f) (0.5 35), and (h) (0.1 35)).

Overall, RSA with type (c) (1 35) performs better at the given SNR.

In Table 4 RSA with type (f) performs better (69 %) than RSF (58 %) at 10 dB SNR. RSA with types (f) and (h) perform better (67 %) than RSF (60 %) at 15 dB SNR. RSA with types (f) and (h) perform much better (73 %) than RSF (66 %) at 20 dB SNR.

At 10 dB, increase in bandwidth from RSA with type (f)(0.5 35) to RSA with type (h)(0.1 35) there is a slight decline in recognition accuracy of 1 % (from 69 % to 68 %). On the other hand, at 15 dB and 20 dB SNR similar increase in bandwidth of RSA with type (f)(0.5 35) to that of RSA with type (h) (0.1 35) shows no change in results, both at 67 % and 73 % respectively.

Overall, RSA with type (f) (0.5 35) performs better.

In Table 5 at 10 dB SNR, RSA with type (h) performs better (58.7 %) than RSF (56.3 %). RSA with type (h) is a better performer (82.7 %) among the new RSA and is better than RSF (79.9 %) at 15 dB SNR. RSA with types (c) and (d) are better performers (91.1 %) among the new RSA and their performance is better compared to RSF (89.1 %) at 20 dB SNR.

Generally, RSA with a 35 frequency component range shows a better performance than RSA with a 7 frequency component range.

For RSA with a 35 frequency component range, the recognition accuracy results increases from 55.8 % to 57.6 % and later to 58.7 % at 10 dB SNR and from 80.8 % to 82.3 % and later to 82.7 % at 15 dB SNR for RSA with type (c) (1 35), RSA with type (f) (0.5 35) and RSA with type (h) (0.1 35),respectively. At 20 dB SNR, there is a slight decline in accuracy from 91.1 % to 90.5 % for RSA with type (c) (1 35) and both RSA with types (f) (0.5 35) and (h) (0.1 35) respectively.

RSA with type (h) (0.1 35) performs better at $<$

Table 3: *Average recognition accuracy(%) for 100 Japanese common male speech phrases*

	Avg(%) for 4 Noises		
	10 dB	15 dB	20 dB
RSF	72.5	87.6	92.8
RSA:Type(a)	69.3	83.5	88.5
RSA:Type(b)	74.0	87.0	91.3
RSA:Type(c)	76.6	90.1	94.9
RSA:Type(d)	76.5	89.9	94.8
RSA:Type(e)	66.4	81.2	86.5
RSA:Type(f)	72.6	87.2	92.7
RSA:Type(g)	66.9	81.2	86.4
RSA:Type(h)	72.6	87.2	92.7

Table 4: *Average recognition accuracy(%) for Japanese similar pronunciation male speech phrases*

	Avg(%) for 4 Noises		
	10 dB	15 dB	20 dB
RSF	58	60	66
RSA:Type(a)	57	61	61
RSA:Type(b)	63	65	71
RSA:Type(c)	65	66	68
RSA:Type(d)	65	66	70
RSA:Type(e)	62	63	67
RSA:Type(f)	69	67	73
RSA:Type(g)	55	56	61
RSA:Type(h)	68	67	73

Table 5: *Average recognition accuracy(%) for 100 Japanese common female speech phrases*

	Avg(%) for 4 Noises		
	10 dB	15 dB	20 dB
RSF	56.3	79.9	89.1
RSA:Type(a)	51.5	75.9	84.4
RSA:Type(b)	56.3	80.3	89.4
RSA:Type(c)	55.8	80.8	91.1
RSA:Type(d)	55.3	80.5	91.1
RSA:Type(e)	55.0	80.2	88.2
RSA:Type(f)	57.6	82.3	90.5
RSA:Type(g)	55.5	80.3	88.2
RSA:Type(h)	58.7	82.7	90.5

Table 6: *Average recognition accuracy(%) for Japanese similar pronunciation female speech phrases*

	Avg(%) for 4 Noises		
	10 dB	15 dB	20 dB
RSF	55	62	71
RSA:Type(a)	60	67	70
RSA:Type(b)	60	67	70
RSA:Type(c)	62	63	73
RSA:Type(d)	58	66	75
RSA:Type(e)	60	62	69
RSA:Type(f)	57	64	69
RSA:Type(g)	62	62	69
RSA:Type(h)	59	64	68

20 dB SNR while RSA with types (c) (1 35) and (d)(1 40) perform better at > 15 dB SNR.

In Table 6 RSA with types (c) and (h) show better performance (64 %) among RSA schemes and are better than RSF (57 %) at 10 dB SNR. At 15 dB SNR, RSA with type (d) performs better (72 %) than other RSA schemes and better than RSF (68 %). RSA with type (d) is a better performer (77 %) among the RSA schemes and equally performs better than RSF (75 %) at 20 dB SNR. Generally, RSA with a 35 frequency component range shows a better performance than RSA with a 7 frequency component range.

For RSA with a 35 frequency component range, the recognition accuracy shows a tendency of decline from 64 % to 62 % at 10 dB SNR and a decline from 71 % to 69 % at 15 dB SNR and from 78 % to 76 % at 20 dB SNR for RSA with type (c) (1 35) and RSA with type (f) (0.5 35),respectively.

3.5 Analysis

Conventionally, RSF is a bandpass filter in a system that reduces the amplitudes of signal components that lie outside a given frequency range. It only lets through components within a band of frequencies. Bandpass filters are particularly useful for analysing the spectral content of signals. The proposed RSA simulates bandpass filtering by processing selected frequency components in modulation spectrum domain.

Experimental results show that the proposed RSA performs better than conventional RSF. In the case of Japanese common speech phrases for male speaker in Table 3, new RSA with type (c) (1 35) produce better results while for Japanese similar pronunciation male speech phrases in Table 4, new RSA with type (f) (0.5 35) show better performance among the evaluated specifications.

In the case of Japanese common female speech phrases in Table 5, the proposed RSA with type (h) (0.1 35) show better results while for Japanese similar pronunciation female speech phrases in Table 6, the proposed RSA with type (c) (1 35) and RSA with type (g) (0.1 7) at 10 dB, the new RSA with type (a) (1 7) and with type (b) (1 15) at 15 dB, and the RSA with type (d) at $>$ 15 dB SNR perform better, respectively.

Based on the experimental results, for male NSP we found the most effective method to be RSA with type (c) (1 35) at all SNR under consideration while for male SP RSA with type (f) (0.5 35) was better at $>$ 10 dB SNR. In the case of female speaker, the results indicate that for NSP the most effective method is RSA with type (h) (0.1 35) at $<$ 20 dB SNR, while at $>$ 15 dB SNR, RSA with type (d) (1 40) show better performance. For SP, RSA with type (h) (0.1 35) is better at $<$ 15 dB SNR while RSA with type (d) (1 40) performs better at $>$ 10 dB SNR.

4. DISCUSSION

In this section, we discuss the findings of our experiments. We show the positive contributions in applying the proposed enhanced RSA types with high frequency components on isolated speech recognition. By using a different number of frequency components, we mimic bandpass filtering to isolate each frequency region of the signal in turn so that we can measure the energy in a selected region. The same process is applied both on male and female speech recognition. Table 7 shows the average improvement on recognition accuracy for the better performers at each SNR.

Table 7: *Average recognition improvement(%)*

	Avg improvement(%)		
	10 dB	15 dB	20 dB
Male, NSP	4.1	2.5	2.1
Male, SP	11	7	7
Female, NSP	2.4	2.8	2.0
Female, SP	7	4	2

Both, the speech type (NSP and SP) and SNR (at 10 dB, 15 dB, and 20 dB) tend to have an influence on performance of proposed method hence the difference in results. The results indicate that proposed enhanced RSA depends on the input signal. Although in each speaker and speech categories there is a enhanced RSA type that shows a superior performance. Both the wide band and narrow band perform differently on male and female speech phrases. For instance, male SP has 11 % improvement at 10 dB compared to 7 % for female SP. Our proposed method shows improved performance on male SP compared to female SP (11 %, 7 %, 7 %, versus 7 %, 4 %, 2 %,) at 10 dB, 15 dB, and 20 dB, respectively. On the other hand, results for male NSP versus female NSP are given as (4.1 %, 2.5 % 2.1 % versus 2.4 %, 2.8 %, and 2.0 %), respectively. It has been observed that under the experimental conditions, male NSP is better than female NSP at 10 dB , while female NSP is slightly better than male NSP at 15 dB.

The accuracy of a speech recognition system can be defined as the percentage of time that the recognizer correctly identifies an input utterance. Recognition errors can be generally classified as misrecognitions or as nonrecognition errors. The tendency of differences in recognition accuracy between male and female can be attributed to many factors including user characteristics(age, sex), the language (vocabulary size), and the channel and environment (noise), for example, among many others [29]. The more varied the group of speakers using the system, the more challenging the recognition process. It is more difficult for a speaker-independent system to recognize accurately both male and female speakers.

The most limiting problem of larger vocabulary sizes is the corresponding decrease in recognizer accuracy. This refers to the total number of different phrases the speech recognizer is able to identify. Therefore, the tendency of differences in recognition accuracy between the 100 Japanese phrases and the Japanese similar pronunciation phrases is due to the differences in sizes of databases. A smaller database (of similar pronunciation phrases) has an increased chance of better recognition accuracy compared to a much larger database (of 100 Japanese phrases), in this case. In the latter, increased number of misrecognitions and false recognitions are often recorded as a result compared to in the former.

5. CONCLUSION

The paper proposes to use running spectrum analysis (RSA) with certain passbands for noisy speech recognition. Performances of speech recognition for Japanese short phrases are compared with those by running spectrum filtering (RSF). Experiments are conducted for various passbands, and the results show an advantage over RSF method.Filtering is optimized as in the case of RSA.

Theoretical analysis indicates the proposed RSA bandpass schemes are less complex to realize and experimental results demonstrate the effectiveness of the proposed approach in improving the robustness of automatic isolated phrase recognition.

From the experimental results it has been demonstrated that the use of RSA with high frequency components, particularly the ones in the range of (0.5 35), for example can be useful in ASR. In this study, RSA on a 35 frequency component range shows a better performance than RSA on a 7 frequency component range used in other related research study. Under noisy conditions different types of RSA show different performance for male and female speakers. It has also been discovered that in the case of male speakers system performance is influenced mostly by the RSA type while that of female speakers, the performance relies mostly on SNR. In future we plan to evaluate our proposed method on recognizing children's speech and develop a recognition system that can distinguish between a child voice and that of an elderly person.

ACKNOWLEDGEMENT

The authors would like to thank Raytron, inc, Japan for fruitful discussions. This study is supported in parts by the Japan Science and Technology Agency for A-Step Program (AS2416901H).

References

[1] M. Watanabe, H. Tsutsui and Y. Miyanaga,"Robust speech recognition for similar pronunciation phrases using MMSE under noise environments," *Proc. 13th International Symposium on Communications and Information Technologies (ISCIT)*, Surat Thani, pp.802-807, 2013.

[2] J. Tierney, "A study of LPC analysis of speech in additive noise," *IEEE Trans. on Acoustic., Speech, and Signal Process.*, vol. ASSP-28, no. 4, pp. 389-397, Aug. 1980.

[3] S. Kay, "Noise compensation for autoregressive spectral estimation," *IEEE Trans. on Acoust., Speech, and Signal Process.*, vol. ASSP-28, no. 3, pp. 292-303, Jun 1980.

[4] P. B. Patil, "Multilayered network for LPC based speech recognition," *IEEE Transactions on Consumer Electronics*, vol. 44, no. 2, pp. 435-438, May 1998.

[5] Mark G. Hall, Alan V. Oppenheim, and Alan S. Willsky,"Time-varying parametric modeling of speech," *Signal Processing*, vol. 5, pp. 267-285, 1983.

[6] S. Tanweer, A. Mobin and A. Alam,"Analysis of Combined Use of NN and MFCC for Speech Recognition," *International Journal of Computer, Electrical, Automation, Control and Information Engineering*, vol. 8, no. 9, 2014.

[7] L. Muda, M. Begam and I. Elamvazuthi, "Voice Recopgnition Algorithm using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques," *Journal of Computing*, vol. 2, no. 3, pp. 138-143, 2010.

[8] C. Ittichaichareon, S. Suksri, and T. Yingthawornsuk, "Speech Recognition Using MFCC," *International Conference on Computer Graphics, Simulation and Modelling (ICGSM2012)*, pp. 135-138, 2012.

[9] Anjali Bala, Abhijeet Kumar, Niddhika Birla,"Voice command recognition system based on MFCC and DTW," *International Journal of Engineering Science and Technology*, vol. 2, no. 12, pp. 7335-7342, 2010.

[10] Petr Motliček, "Feature Extraction in speech coding and recognition," *Report of PhD research internship in ASP Group*, OGI-OHSU, 2002,

[11] K. Yao, K. K. Paliwal and S. Nakamura, "Model-based noisy speech Recognition with Environment Parameters Estimated by noise adaptive speech Recognition with prior," *EUROSPEECH 2003-GENEVA*, Switzerland, Tech. Rep., 2003.

[12] Q. Zhu, N. Ohtsuki, Y. Miyanaga, and N. Yoshida,"Robust speech analysis in noisy environment using running spectrum filtering," *International Symposium on Communications and Information Technologies*, vol. 2, pp. 995-1000, Oct. 2004.

[13] N. Ohtsuki, Qi Zhu and Y. Miyanaga, "The effect of the musical noise suppression in speech noise reduction using RSF," *International Symposium on Communications and Information Technologies*, vol. 2, pp. 663-667, Oct. 2004.

[14] V Tyagi, I. McCowan, H. Misra, and H. Boulard, "Mel-Cepstrum modulation spectrum (MCMS) features for Robust ASR," *in Proc. 2003 IEEE Workshop on Automatic Speech Recognition and Understanding*,St. Thomas, pp. 399-404, 2003.

[15] Dimitrios Dimitriadis,Petros Maragos, and Alexandros Potamianos, "Modulation features for Speech Recognition," *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, May 2002.

[16] Jeih-Weih Hung, Hsin-Ju Hsieh, and Berlin Chen, "Robust Speech Recognition via Enhancing the Complex-Valued Acoustic Spectrum in Modulation Domain," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 24, Issue 2, pp. 236-251, Feb. 2016.

[17] K. Ohnuki, W. Takahashi, S. Yoshizawa, and Y. Miyanaga, "New acoustic modeling for robust recognition and its speech recognition system," *International Conference on Embedded Systems and Intelligent Technology*, 2009.

[18] S. Yoshizawa and Y. Miyanaga, "Robust recognition of noisy speech and its hardware design for real time processing.," *ECTI Trans. Elect., Eng., Electron., and Commun.*, vol.3, no.1, pp. 36-43, Feb. 2005.

[19] K. Ohnuki, W. Takahashi, S. Yoshizawa, and Y. Miyanaga, "Noise Robust speech features for Automatic Continuous Speech Recognition using Running Spectrum Analysis," *in: Proc. of 2008 International Symposium on Communications and Information Technologies (ISCIT)*, pp. 150- 153 (October 2008).

[20] Yiming Sun and Yoshikazu Miyanaga, "A Noise-Robust Continuous Speech Recognition System Using Block-Based Dynamic Range Adjustment," *IEICE Trans. INF. & SYST*, vol.95-D, no.3, March 2012.

[21] T. Chi, Y. Gao, M. C. Guyton, P. Ru and S. Shamma, "Spectro-temporal modulation transfer functions and speech intelligibility," *J. Acoust. Soc. Am.*, 106(5), pp. 2719–2732, 1999.

[22] Naoya Wada and Yoshikazu Miyanaga, "Robust Speech Recognition with MSC/DRA Feature Extraction on Modulation Spectrum Domain," *in Proc. Second International Symposium on Communications, Control and Signal Processing (ISCCSP)*, Marakech, Morocco, Mar. 2006.

[23] M.A Anusuya and S.K. Katti, "Speech Recognitionb by Machine: A Review," *International Journal of Computer Science and Information Security*, (IJCSIS), vol. 6. no. 3, pp. 181-205, 2009

[24] Noboru Kanedera, Takayuki Arai, Hynek Hermansky and Misha Pavel, "On the importance of various modulation frequencies for speech recognition," *Proceedings of EUROSPEECH 97*, Rhodos, Greece, Sep. 1997.

[25] Hynek Hermansky, Eric Wan, and Carlos Avendano, "Speech enhancement based on temporal processing," *IEEE International Conference on Acoustic, Speech and Signal Processing*, Detroit, Michigan, Apr.1995.

[26] Carlos Avendano, Sarel van Vuuren and Hynek Hermansky, "On the properties of temporal processing for speech in adverse environments," *Proceedings of 1997 Workshop on Applications of Signal Processing to Audio and Acoustics*, Mohonk Mountain House, New Paltz, New York, October 18-22, 1997.

[27] Eslam Mansour mohammed, Mohammed Shraf Sayed, Abdalla Mohammed Moselhy and Abdelaziz Alsayed Abdelnaiem, "LPC and MFCC Performance Evaluation with Artificial Neural Network for Spoken Language Identification," *International Journal of Signal Processing, Image Processing and Pattern Recognition*, vol. 6, no. 3, Jun. 2013.

[28] M. H. Moatta and M. M. Homayounpour, "A Simple but Efficient Real-Time Voice Activity Detection Algorith," *17th European Signal Processing Conference (EUSIPCO)*, August 24-28, 2009.

[29] Sherry P. Casall and Robert D. Dryden, "The Effects of Recognition Accuracy and Vocabulary Size Of A Speech Recognition System on Task Performance and User Acceptance," *Industrial Engineering and Operations Research*, 1988.

George Mufungulwa received his B.S. degree in Computer Science from The Copperbelt University, Kitwe, Zambia, in 1999 and his M.Sc. in Distributed Interactive Systems from University of Lancaster, UK, in 2002. He is currently a Ph.D. student at Graduate School of Information Science and Technology, Division of Media and Network Technologies, Hokkaido University, Sapporo, Japan. His research interests include digital signal processing and multimedia systems. He is a member of IEEE and IEICE.

Hiroshi Tsutsui received his B.E. degree in Electrical and Electronic Engineering and his master and Ph.D. degrees in Communications and Computer Engineering from Kyoto University in 2000, 2002, and 2005, respectively. He is currently an associate professor in Division of Media and Network Technologies, Hokkaido University. His research interests include circuits and systems for image processing and VLSI design methodology. He is a member of IEEE, ACM, IPSJ, IEEJ, and IIEEJ.

Yoshikazu Miyanaga received the B.S., M.S., and Dr. Eng. degrees from Hokkaido University, Sapporo, Japan, in 1979, 1981, and 1986, respectively. Since 1983 he has been with Hokkaido University. He is now Professor at Division of Information Communication Systems in Graduate School of Information Science and Technology, Hokkaido University. From 1984 to 1985, he was a visiting researcher at Department of Computer Science, University of Illinois, USA. His research interests are in the areas of speech signal processing, wireless communication signal processing and low-power VLSI system design.

Shinichi Abe developed an automotive electronics products at Pioneer Co. and he is currently a stuff of Vehicle Information and Communication System Center (VICS Center), Business Research division, Tokyo, Japan.