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

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

ABSTRACT

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

Keywords: MFCC, HMM, ASR, RSF, RSA

1. INTRODUCTION

The fundamental functions in speech recognition are speech feature extraction and feature matching. Various speech features, including ones from linear prediction coding (LPC) [14], time-varying linear prediction coding (TVLPC) [5], and mel frequency cepstral coefficients (MFCC)[69] among others, have been used to model speech recognition either singularly or collectively in improving the accuracy of speech recognition. MFCC, which is based on the spectral content of a signal and can be considered as one of the standard methods for feature extraction [10] and it was used 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 its computational costs and high demand on memory.

In the recent past, several common methods related to the use of modulation spectrum features for noisy speech recognition have been developed [1416]. Running spectrum analysis (RSA) is not only an effective technique for reduction of noise in 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, its application has been mainly focused on the frequency components in the range of 2-8 Hz because this range contains the dominant elements of the amplitude envelope of speech [20-21]. The modulation frequency band at higher than 8 Hz can be regarded as miscellaneous noise or unnecessary speech components related to the speakers characteristics, such as tone and pronunciation, among other factors [22].

However, this work presents a novel noise-robust feature extraction framework that leverages the technique of RSA on isolated phrase recognition. This work was envisioned with the goal to enhance RSA for the purpose of achieving higher recognition accuracy for both males and females, as well as for similar and dissimilar pronunciation of spoken Japanese 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 a human machine interface such as automatic call processing in telephone networks and query based information systems such as voice dictation, and stock price quotations, [23] among others. The authors assumed that the proposed method performance relates to gender just as recognition accuracy can be influ-

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

enced by the signal-to-noise ratio (SNR), which the authors aim to ascertain.

In this study, running spectrum analysis (RSA) was applied to the modulation spectrum for noise robust speech recognition of adequately selected frequency components. The noise effect was dealt with by filtering the range of frequency components, 1-7 Hz, 1-15 Hz, 1-35 Hz and 1-40 Hz in the modulation spectrum domain. Furthermore, it is argued that the expected speech recognition accuracy can be improved when modulation spectrum filtering (MSF) directly modifies 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 noise, almost all ASR systems using HMM and noise reduction can show higher accuracy of speech recognition than 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, the 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.

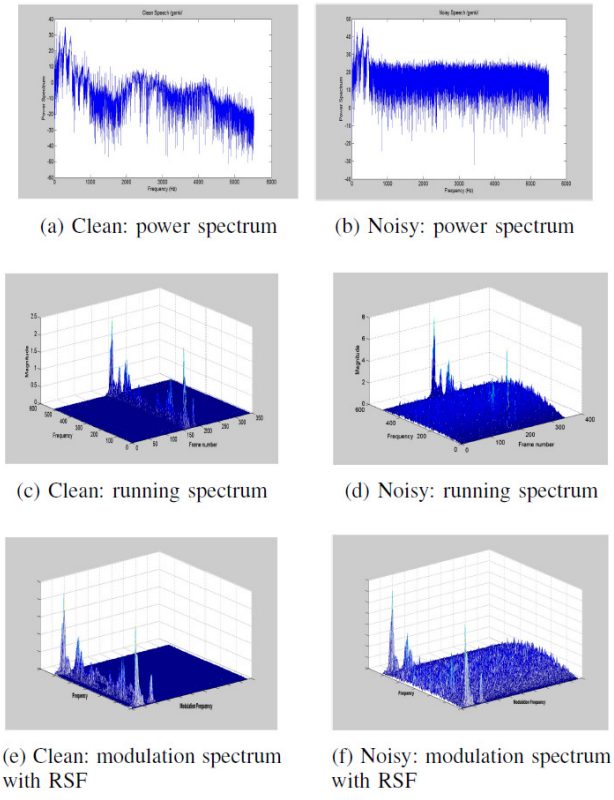
2. PROPOSED SYSTEM

The motivation of this study is to evaluate the effectiveness of enhanced running spectrum analysis (RSA), which is explained later, as it compares with running spectrum filtering (RSF). RSA is a processing of speech over modulation of the 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 the modulation spectrum of speech may improve the quality of noisy speech. Investigations on the possibilities of the modulation spectrum domain for enhancement of noisy speech [25-26] support the dominance of the modulation spectrum components in the vicinity of 2-8 Hz. We now explain the effect of noise in running and modulation spectrum domains.

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

Figures 1(a) and 1(b) show the power spectra of cleanspeech and speech with additive white noise at a 10 dB SNR for a Japanese phrase /genki/. Both spectra were calculated from short time speech waveforms. These figures indicate that the dynamic range of a power spectrum of a noisy speech is smaller than that of a clean spectrum. Additionally, 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, frame number and power amplitude axes.

When we observed the data on the frame number axis, the frequency was fixed to a specific value and its data can be recognized in the time domain. They can be applied using a fast Fourier transform (FFT). After such a FFT is applied to all frequencies, we can obtain new 3-D data in the modulation spectrum domain. The 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 the processes for female speakers. For each gender case, two output models for similar pronunciation (SP) and dissimilar pronunciation (NSP) are, respectively, realized. In the proposed system, there are four different kinds of filtering in RSA. The optimal filtering of RSA was applied for male and female speakers, SP and NSP.

In Figure 2, noisy speech at different signal-to-noise ratios (SNRs) was 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 or silent. As in the case of training, the speech features were extracted using the standard MFCC for spectral analysis. A HMM based automatic speech recognition (ASR) system was utilized for testing. The gender of speaker (male or female) as well as the speech type, SP or NSP for each gender were decided. This process results in four outputs: male SP, male NSP, female SP and female NSP. For each gender and speech type combination, the speech signal was passed through a voice activity detection (VAD) process to retain segments with speech activity or segments of 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 a 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 to obtain mel cepstrum, speech data was initially pre-emphasized and the pre-emphasized speech waveform in the time domain is frame-blocked and windowed with a predefined analysis window. Later, a fast Fourier transform (FFT) was computed. The magnitude of the output was then weighted using a series of mel filter frequency responses whose center frequencies and bandwidth roughly matched those of the auditory critical band filters [27]. The FFT bins were later combined so that each filter had unit weight. From the weighted sums of all amplitudes of signals, a vector was obtained by logarithmic amplitude compression. RSF was then applied beforetransforming the result to a MFCC parameter using a 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 an important research challenge [23]. Most current feature extraction methods are still vulnerable to certain noises such as car noise [28].

Figure 4 shows the MFCC feature extraction process with running spectrum analysis (RSA). After spectral analysis, RSA was applied to realize the modulation spectrum. After that stage, the process was done as explained under feature extraction with RSF. In both cases, the features were trained using HMM.

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 shows that, for example, in the case of a female NSP, RSA with type (h) performed better at high noise levels, while types (c) and (d) performed better at low noise levels. Similarly, for a female SP, RSA with types (c) and (h) were found to perform better at high noise levels, while type (d) performed better with less noise. The candidate for use with results examining male or female speech were selected based on the maximum likelihood of HMM. Under noisy conditions, various types of RSA showed different performance for male and female speakers.

The proposed RSA differs from the one discussed in [19]. The former focuses on a modulation frequency range of 2-8 Hz. However, in this study we evaluate the performance of several RSA types shown in Table 1. This table 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 had 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 the new RSA over the 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 was to compare the performance of the proposed enhanced RSA

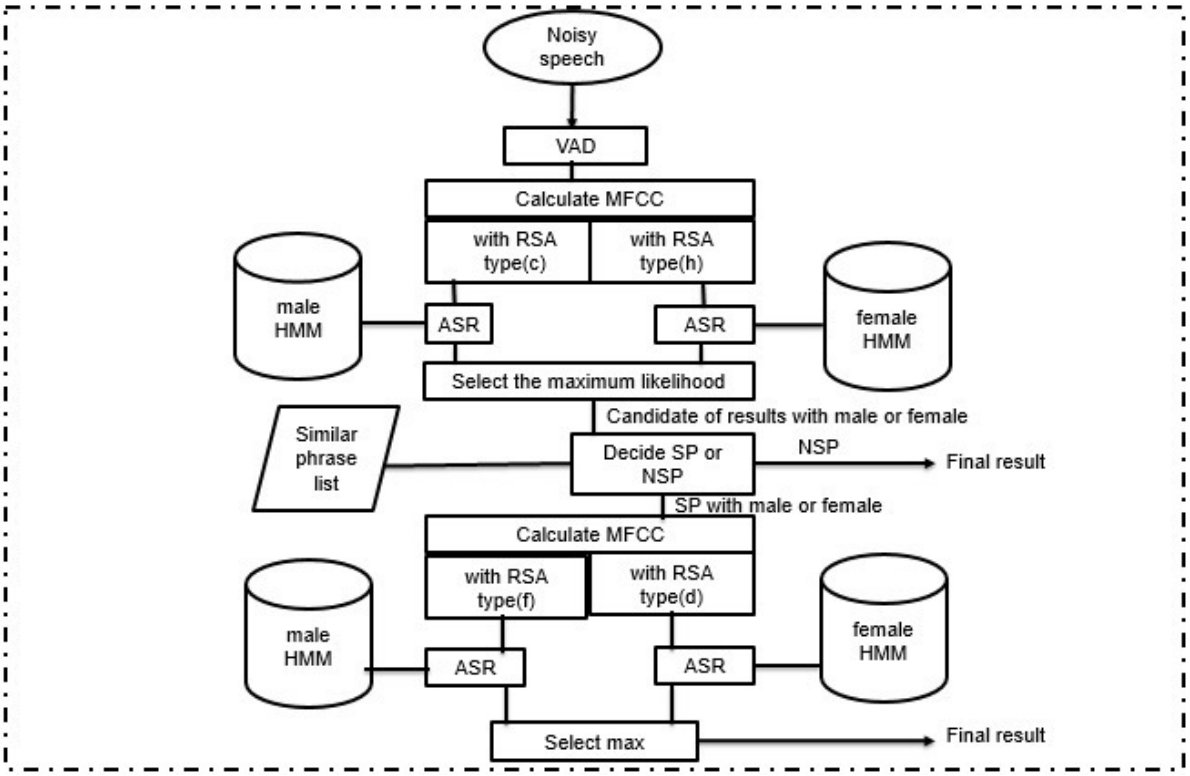


Fig.2: *Proposed system.*

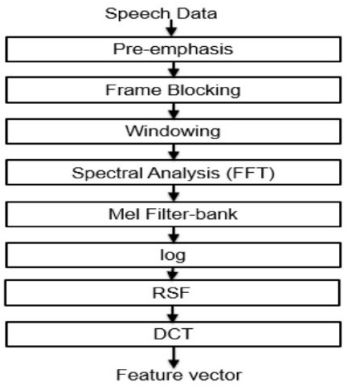


Fig.3: *Feature extraction with RSF.*

to the RSF on similar and dissimilar pronunciation of Japanese phrases. The second objective was to evaluate how the performance related to gender. The main method used for speech enhancement was filtering. We evaluated the adaptability of our proposed RSA over the modulation spectrum and compared its results to those of RSF. In this study, RSF was employed to act as the basis for comparison and to determine better performing RSA types at given SNRs for both genders.

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 six similar phrases and 100 common Japanese phrases, and each phrase repeated three times, were used for the front-end feature extraction and for 32-state isolated (HMM) in training. Testing sets were utilized consisting of ten male speakers and ten female speakers (not used in training), with each speaker uttering six similar phrases and 100 common Japanese phrases and each phrase repeated three times.

Table 2 shows the simulation parameters.

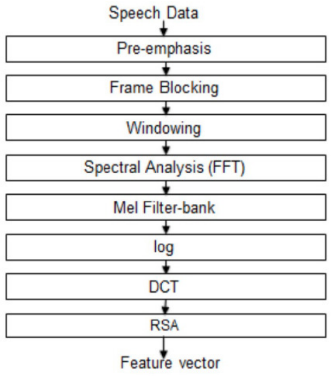


Fig.4: *Feature extraction with RSA.*

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,

The speech sample was at 11.025 KHz with 16-bit quantization. Frame-by-frame, 38-dimensional FFT based MFCC feature vectors were extracted after pre-emphasis and Hanning windowing. In the testing stage, 10, 15, and 20 dB of four types of noises were artificially added to the original speech. We compare the performance of proposed enhanced RSA of specified passbands to those by RSF with four types of noises; white, pink, HF channel and babble noises using MATLAB (R2014a) software. Under the stated conditions, we measure the average recognition rates for ten speakers on RSF and eight enhanced RSA passband specifications given as Types (a) to Type (h) at 10, 15, and 20 dB SNR.

Table 3 shows the average recognition accuracy for 100 common Japanese phrases spoken by males. Table 4 shows the average recognition accuracy for similar spoken Japanese phrases by male speakers. Table 5 shows the average recognition accuracy for 100 common Japanese phrases spoken by female speakers. Table 6 shows the average recognition accuracy for similar Japanese phrases spoken by female speakers.

3.3 Simulation results and analysis

Analysis is carried out for the Japanese common and similar phrase databases. We used gender (male and female) and four SNRs (10, 15, and 20 dB) as variables. Analysis of results analysis focused on the performance of the enhanced RSA types on the various acoustic measures. The four kinds of noises used in the experiments were based on Signal Processing Information Base (SPIB) noise data measured in the field by the Speech Research Unit (SRU) at the Institute for Perception-TNO, Netherlands, United Kingdom, under 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) performed better (76.6 %) compared to RSF (72.5 %). At 15 dB SNR, RSA with type (c) performed better (90.1 %) compared to RSF (87.6 %). RSA with type (c) performed better (94.9 %) than RSF (92.8 %) at 20 dB SNR.

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

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

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

At 10 dB, an increase was observed in a bandwidth from RSA with type (f)(0.5 35) to RSA with type (h)(0.1 35). There was a slight decline in recognition accuracy of 1 % (from 69 % to 68 %). Alternatively, at a 15 and 20 dB SNR, a similar increase in bandwidth of the RSA with type (f)(0.5 35) to that of RSA with type (h) (0.1 35) showed no change in the results, both at 67 % and 73 %.

Overall, RSA with type (f) (0.5 35) performed better. In Table 5, at a 10 dB SNR, RSA with type (h) performed better (58.7 %) than RSF (56.3 %). RSA with type (h) was a better performer (82.7 %) among the new RSAs and was better than RSF (79.9 %) at a 15 dB SNR. RSA with types (c) and (d) were better performers (91.1 %) among the new RSAs and their performance was improved compared to RSF (89.1 %) at a 20 dB SNR.

Generally, RSA with a 35 frequency component range showed better performance than RSA with a 7 frequency component range. For RSA with a 35 frequency component range, the recog- nition accuracy increased from 55.8% to 57.6% and later to 58.7% at a 10 dB SNR and from 80.8% to 82.3% and later to 82.7% at a 15 dB SNR for RSA with type (c) (1 35), RSA with type (f) (0.5 35) and RSA with type (h)

Table 3: *Average recognition accuracy(%) for 100 common Japanese phrases spoken by males*

	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 similar Japanese phrases spoken by males*

	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 common Japanese phrases spoken by females*

	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 similar Japanese phrases spoken by females*

	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

(0.1 35), respectively. At a 20 dB SNR, there was 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).

RSA with type (h) (0.1 35) performed better at a < 20 dB SNR while RSA with types (c) (1 35) and (d) (1 40) performed better at a > 15 dB SNR. In Table 6 RSA with types (c) and (h) showed better performance (64%) among RSA schemes and were better than RSF(57%) at a 10 dB SNR. At a 15 dB SNR, RSA with type (d) performed better (72%) than other RSA schemes and better than RSF (68%). RSA with type (d) was a better performer (77%) among the RSA schemes and also performed better than RSF (75%) at a 20 dB SNR. Generally, RSA with a 35 frequency component range showed better performance than RSA with a 7 frequency component range.

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

with type (f) (0.5 35).

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 simulated bandpass filtering by processing selected frequency components in the modulation spectrum domain.

Experimental results showed that the proposed RSA performed better than a conventional RSF. In the case of common Japanese phrases spoken by males in Table 3, the new RSA with type (c) (1 35) produced better results while for similar Japanese phrases spoken by males in Table 4, the new RSA with type (f) (0.5 35) showed better performance with reference to the evaluated specifications.

In the case of common Japanese phrases spoken by females in Table 5, the proposed RSA with type (h) (0.1 35) showed better results for similar Japanese phrases in Table 6. The proposed RSA with type (c) (1 35), the 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 a $>$ 15 dB SNR performed better.

Based on the experimental results, for male NSP we found the most effective method was RSA with type (c) (1 35) at all SNRs under consideration, while for male SP, RSA with type (f) (0.5 35) was better at $>$ 10 dB SNR. In the case of female speakers, the results indicated that for NSP, the most effective method was RSA with type (h) (0.1 35) at a $<$ 20 dB SNR,while at a $>$ 15 dB SNR, RSA with type (d) (1 40) showed better performance. For SP, RSA with type (h) (0.1 35) was better at a $<$ 15 dB SNR, while RSA with type (d) (1 40) performed 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. Using a different number of frequency components, we mimicked bandpass filter to isolate each frequency region of the signal in turn so that we could measure the energy in a selected region. The same process was applied for both male and female speech recognition. Table 7 shows the average improvement of recognition accuracy for the better performers at each SNR.

Both, the speech type (NSP and SP) and SNR (at 10, 15, and 20 dB) tended to influence performance of the proposed method hence the variation in results. The results indicate that proposed enhanced RSA depended on the input signal. For each speaker and speech categories, there was an enhanced RSA type that showed a superior performance. Both the wide and narrow bands performed differently on phrases spoken by males and females. For instance, a male SP had an 11% improvement at 10 dB compared to 7% for female SPs. Our proposed method showed improved performance on male SPs compared to female SPs (11%, 7%, 7%, versus 7%, 4%, 2%,) at 10, 15, and 20 dB, respectively. Alternatively, results for male versus female NSPs were (4.1%, 2.5% 2.1% versus 2.4%, 2.8%, and 2.0%), respectively. It was observed that under the experimental conditions, male NSPs were better than female NSP at 10 dB, while female NSPs were slightly better than male NSPs at 15 dB.

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

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 nonrecognition errors. The tendency of differences in recognition accuracy between males and females can be attributed to several factors including user characteristics (age, gender), the language (vocabulary size), and the channel and environment (noise), 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 accurately recognize 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 the differences in recognition accuracy among the 100 Japanese phrases and the Japanese similar pronunciation phrases was due to the difference in database size. A smaller database (of similar pronunciation phrases) has an increased chance of better recognition compared to a much larger database (of 100 Japanese phrases), in this case. In the latter, an increased number of misrecognitions and false recognitions is often recorded as a result.

5. CONCLUSION

The paper proposes running spectrum analysis (RSA) with certain passbands for noisy speech recognition. Performance of speech recognition for short Japanese phrases was compared with those obtained by running spectrum filtering (RSF). Experiments

were conducted for various passbands, and the results showed an advantage over the RSF method. Filtering was 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), can be useful in ASR. In this study, RSA on a 35 frequency component range showed better performance than RSA on a 7 frequency component range used in other related research studies. Under noisy conditions, various types of RSA showed different performance for male and female speakers. It was discovered that in the case of male speakers, system performance was influenced m ostly by the RSA type, while for that of female speakers, the performance relied mostly on the SNR. In future, we plan to evaluate our proposed method on recognition of childrens speech and develop a recognition system that can distinguish between a child's voice and that of an elderly person.

ACKNOWLEDGEMENT

The authors would like to thank Raytron, Inc, Japan, for fruitful discussions. This study was supported in part bythe 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 Motlíček, "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.