2007 International Symposium on Nonlinear Theory and its Applications
NOLTA'07, Vancouver, Canada, September 16-19, 2007

Robust Speech Recognition Based on Running Speech Spectrum on Critical Band Intensity

Nongnuch SUKTANGMAN[*], Kraisin SONGWATANA[*] and Yoshikazu MIYANAGA[†]

[*]Faculty of Engineering, King Mongkut's Institute of Technology, Ladkrabang 3-2 Chalongkrung Road, Ladkrabang, Bangkok 10520, Thailand.

[†] Graduate School of Information Science and Technology, Hokkaido University, Kita 14 Nishi 9, Kita-ku, Sapporo-shi, Hokkaido, 060-0814, Japan. Email: [†] miya@ist.hokudai.ac.jp

ABSTRACT

In this report, we introduce the new results of robust automatic speech recognition (ASR) based on features of speech spectrum on Bark scale. The robustness is improved by adding running spectrum filtering (RSF) techniques and dynamic range adjustment (DRA) to the features. Bark scale is a psychoacoustics measurement on human hearing property and speech features extraction processes consists of four steps: (1) auto-regressive model (AR model), (2) critical band intensity (CBI), (3) logarithm CBI into discrete cosines transform, DCT (log(CBI)). The detailed information feature is extracted by RSF and DRA from DCT spectra of log CBI sixteen dimension parameters vectors, sixteen dimension parameters of delta, parameter voice energy and a parameter of delta energy. In ASR, the utterance signal-to-noise ratio (SNR) for the speech signal is first extracted speech features for recognition and decoded via acoustic hidden Markov models (HMMs) trained with clean data. We explore the noise robust property of the total system and thus several noise circumstances were considered 0 dB SNR to 20 dB. The recognition rates are improved in our experiments by above 27% at 0 dB SNR, 30% at 10 dB SNR and 7% at 20 dB SNR.

Keywords: Robust Speech Recognition, Bark scale, CBI, RSF, DRA.

1. INTRODUCTION

Speech recognition systems trained in quiet environments are subjected to performance degradation in the presence of ambient acoustic noise. The degradation is mainly attributed to the mismatch between clean acoustic models and noisy speech data. Considerable efforts have been made to reduce this mismatch and improve recognition accuracy in noisy conditions [1]. Generally speaking, noise-robust algorithms are applied in the front-end feature domain or in the back-end model domain.

In the front-end feature domain, spectral subtraction [1] and [2], is a commonly used method for noise suppression where additive noise spectrum is estimated and subtracted from the noisy speech spectrum to recover the clean speech spectrum. These techniques typically work at the spectral level of the extracted feature by trying to rid of the effect of external noise on the spectrum. A relatively new technique called RSF processing [3] and [4], which has shown to be quite successful for noise robust speech recognition, tries to remove those noise components in the power spectrum whose temporal properties are quite different from that of the speech component. Band-pass filters, with bandwidths equal to the bandwidths of the temporal characteristic of the speech component, are applied to each frequency band of the spectrum, to get rid of the noise components.

Most contemporary ASR systems attempt to incorporate some of these features [5]-[9]. In some conventional speech recognition system, speech spectrum envelops are calculated from auto-regressive model (AR model) using minimum mean square estimation (MMSE) method. The log spectrum envelops are also employed as speech features. They are converted to Mel-frequency Cepstrum components and Perceptual Linear Predictive Coefficients components, [5]-[7]. As a unique spectrum distortion measure, Bark scale has been studied [5], [8], [9]. This scale is based on human physiological and psychological property. Bark scale is recognized as a suitable scale for recognizing many auditory phenomena, such as perception of loudness and timbre. These processes provide good performance in many languages. In this report, the coefficients of DCT (log(CBI)) from AR are used as feature representation of speech signals in the ASR process.

These features are applied on isolated word speech recognition experiments using HMM. This report is organized as follows: section 2 explains the feature extraction; in section 3 the robust parameters for speech feature by RSF and DRA is discussed; and in section 4 the speech recognition experiments are reported. The concluding remark is presented in Section 5.

2. FEATURES EXTRACTION

Fig.1 shown the total process proposed in this report. The speech data are first segmented into frame of 300 samples where its time length 27.21 ms with 11.025 kHz sampling rate. Each frame of speech is represented by the parameters vector of DCT (log(CBI)), and the robust speech parameters by RSF and DRA. These features are applied to HMM for training and recognition.

2.1 Auto-Regressive Model (AR model)

In Fig.1, the speech data are pre-emphasized and then Hamming window is applied. In the linear acoustics model of speech production [10] and [11], the speech signal is produced by filtering the excitation signal $u(k)$ with a time-varying linear filter (the vocal tract) $s(k)$ as shown in Fig.2. The AR model coefficients are extracted by using MMSE. For a given speech sample at time *k*, the output signal is assumed to be

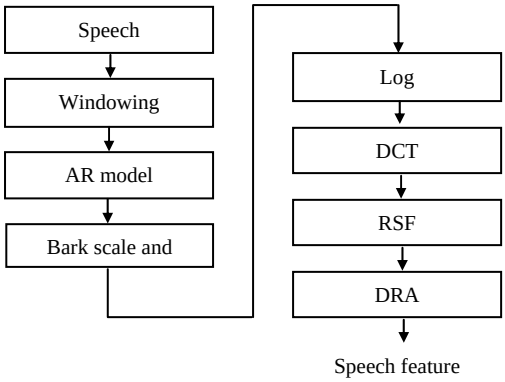

Fig.1: *Block Diagram of a Proposed Speech Recognition System.*

$$s(k) = \sum_{i=1}^{q} b_i s(k-i) \tag{1}$$

where b_i ($i = 1,2,...,q$) are the estimated coefficients of the AR model. q is the order of the Linear Predictive Coefficients. From (1), we get

$$H(z^{-1}) = \frac{S(z^{-1})}{U(z^{-1})} = \frac{1}{1 - \sum_{k-1}^{q} b_k z^{-k}} \tag{2}$$

where $H\left(z^{-1}\right)$ is a Z-transfer function of linear speech production model whose output is an observed speech. $S\left(z^{-1}\right)$ and $U\left(z^{-1}\right)$ are the output speech function and an excitation signal function in z-domain respectively.

Eq.(2) represents the spectrum envelop of MMSE when $z^{-1} = e^{-j\omega}$ and $\omega = 2\pi f$. The value f denotes frequency [Hz]. The spectrum envelop represents an approximation of a linear speech production model. For this report, several experiments are carried out with high accuracy recognition. MMSE with 15-th or more predictive order to speech spectrum envelops are mapped on to the Bark scale and extracted features to high accuracy recognition.

2.2 Bark scale and Critical Band Intensity (CBI)

The Bark scale is a psychoacoustics spectrum measure whose property corresponds to human hearing. In other words, it is based on the fact that our hearing system analyzes speech with critical bands intensity (CBI). The concept of critical band has been developed [8], [9].

Some experiments have shown that critical bands are narrower at the region of low frequencies than at the region of

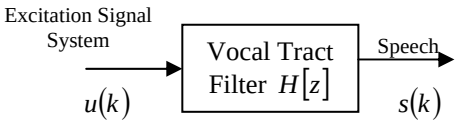

Fig.2: *Linear prediction model of speech.*

high frequencies. The critical bands are analogous to the band of a spectrum analyzer with variable center frequencies and bandwidth.

Based on the measurements by Zwicker [8], the Bark scale is approximately expressed in terms of the linear frequency by

$$\beta = 13\arctan(0.76\times10^{-3} f) + 3.5\arctan(0.13\times10^{-3} f)^2 \tag{3}$$

The range of human auditory frequency spreads from 20 to 20,000 Hz. It covers approximately 25 critical bands on Bark scale. For example, the lowest critical band is represented by β = 1 [Bark]. f_c, the center frequency of a critical band, is = 50 Hz when β = 1. The corresponding critical bandwidth Δf can be expressed by

$$\Delta f = 25 + 75\left[1 + 1.4\times10^{-6} f_c^2\right]^{0.69} \tag{4}$$

which is approximately = 100 Hz.

In this report, the underlying sampling rate is set to be 11,025 kHz with a bandwidth of 5.5 kHz. Accordingly, there are 18 critical bands as listed [8].

The intensity of voice in the critical band (α_m) can be calculated by

$$\alpha_m = \int_{f_{l,m}}^{f_{u,m}} \frac{d\alpha}{d\beta} d\beta \tag{5}$$

where $f_{l,m}$ and $f_{u,m}$ are the lower and upper band frequencies of the *m*-th critical band, respectively. The α represents the spectrum energy of MMSE spectrum envelop, $|H\left(z^{-1}\right)|$. In other words, α_m represents the integrated power of MMSE spectrum envelope in the *m*-th critical band. The samples of 18 CBIs are shown as in Fig.3.

2.3 Discrete Cosines Transform on Logarithm CBI DCT (log(CBI)) and Deltas

It is generally believed that logarithmic function is sensitive to certain types of noise and signal distortions [13]. To increase the dynamic range of CBI, logarithm of the CBI is used. Fig.3 shows an example of mapping of CBI values on to log scale. It shows that the small-valued information on the CBI is enhanced to in the log(CBI), reflecting the importance of them. Let the logarithmic form of α_m be

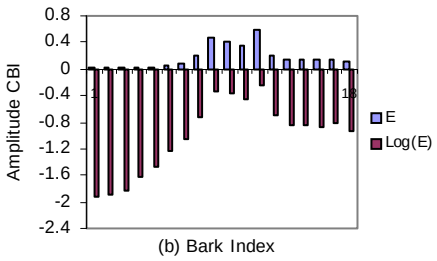

Fig.3: *Critical Band mapping between linear frequency scale and Bark scale.*

$$\delta_m = \log_{10}(\alpha_m) \tag{6}$$

where m = 1 to 18. The Discrete Cosine Transform (DCT) of δ_m is calculated by

$$\mu_m = \frac{2\gamma_m}{M}\sum_{k=0}^{M-1}\delta_k \cos\left(\frac{\pi(2k+1)m}{2M}\right),\ m = 0,1,...,M-1 \tag{7}$$

where

$$\gamma_m = \begin{cases}\sqrt{\frac{1}{2}} & \text{for } m = 0\\ 1 & \text{otherwise}\end{cases}$$

The performance of a speech recognition system can be greatly enhanced by adding time derivatives to the basic static parameters. The delta coefficients are computed using the following regression formula

$$D_m^n = \frac{\sum_{k=-\theta}^{\theta} k\mu_m^{n+k}}{2\sum_{k=-\theta}^{\theta} k^2},\quad 1 \le m \le 18 \tag{8}$$

where D_m^n is the delta of DCT(Log(CBI)) at fram n, computed in terms of the corresponding static parameters $\mu_m^{-\theta}$ to μ_m^{θ}. The value of θ = 2 for this report, 4 terms are used to compute each delta component, two before and two after.

3. Running Spectrum Filtering (RSF) [3],[4] and Dynamic Range Adjustment (DRA) [3],[4]

The method combines running spectrum filtering, (RSF) based on the different time variations of the power spectrum between speech and noise, and post processing to reduce background noise. RSF is a filter which is realized in the modulation spectrum and can effectively reduce noise when applied to the running spectrum of speech. These techniques are used for both additive and convolution noises. FIR filter has been used to do band-pass filtering. Speech components in modulation frequency domain [2]-[4] are dominant around 4 Hz and anything out of the range of 1-12 Hz can be regarded as noise. RSF applies high order FIR filter (typically 240 orders) [4] to realize sharp modulation frequency cut off. Thus RSF realizes effective feature

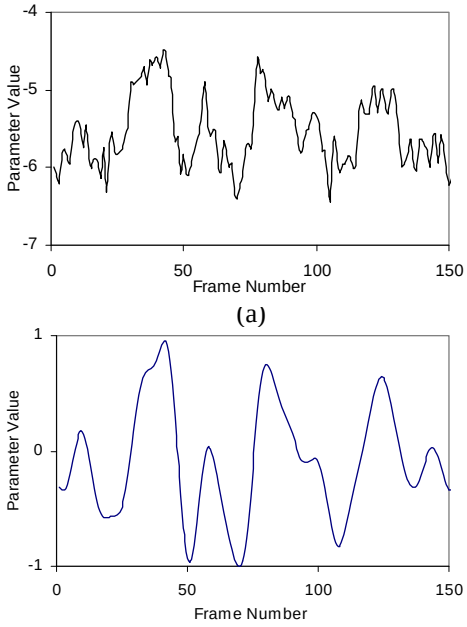

Fig.4: *Indicates speech feature of the 1th order of log CBI-DCT , (a) value of speech feature in 0 dB SNR and (b) feature after RSF and DRA.*

extraction and can be applied in practical speech recognition system. Each filter-bank magnitude component ζ_i, where i is discrete time index filter of RSF and filtered filter-bank magnitudes ζ_i are produced. DCT (log(CBI)), μ_m is performed on the ζ_i, where $1 \le m \le 18$. The result is obtained as robust speech feature.

Dynamic range adjustment (DRA) is applied to reduce the differences between amplitudes of clean speech and noisy speech. DRA adjusts the dynamic range by normalizing the amplitude of a speech feature to its maximum amplitudes as follows.

$$\overline{p}_i(t) = p_i(t) / \max_{j=1,...,m}\left|p_j(t)\right|,\quad i = 1,...,m \tag{9}$$

where $p_i(t)$ denotes an element of the feature vector, m denotes the dimension, t denotes the frame number and $\overline{p}_i(t)$ is defined as normalized parameter of feature.

An example of DCT(log(CBI)) from speech signal with SNR of 0dB is shown in Fig. 4(a). After the process of RSF and DRA, a cleaner speech features is shown in Fig. 4(b). RSF and DRA help to remove unnecessary parts (speaker characteristic and background noise) of the speech required for recognition.

4. Speech Recognition Experiment

4.1 Conditions on experiment

Experiments were carried out on the recognition of isolated Thai word. The Thai syllables used were the name of public institution in Thailand. The speech data were recorded in a quiet room, sampled at 11.025 kHz. The duration of each frame of speech was

Table 1. Comparison between recognition performances conventional robust speech

Noise Name	DCT(log(CBI))			DCT(log(CBI)) with DRA			DCT(log(CBI)) with RSF/DRA		
	0dB	**10dB**	**20dB**	**0dB**	**10dB**	**20dB**	**0dB**	**10dB**	**20dB**
White noise	1.38	21.03	92.16	2.35	21.04	95.96	18.92	81.16	96.26
Pink noise	1.75	74.60	98.95	42.04	88.64	99.09	48.54	94.76	99.73
HF channel noise	1.57	11.29	72.30	1.61	8.42	76.67	6.38	60.44	88.24
Speech babble	4.39	65.93	89.32	6.90	59.86	90.90	17.49	76.46	92.82
Factory floor noise	2.49	73.83	97.30	4..42	70.58	96.48	27.69	88.74	97.71
Jet cockpit noise	1.48	31.40	94.43	1.38	37.94	96.83	18.14	82.95	97.30
Destroyer engine room noise	1.48	19.15	81.27	7.59	24.90	85.55	23.22	83.43	92.85
F-16 cockpit noise	2.39	54.12	95.54	1.58	53.29	95.55	35.01	89.71	97.55
Military vehicle noise	37.48	92.28	95.98	19.57	92.64	97.65	76.95	96.37	98.90
Tank noise	56.63	85.88	87.35	68.47	92.66	95.02	83.87	95.24	97.05
Machine gun noise	52.55	72.76	79.69	73.52	85.73	90.36	84.39	92.17	96.02
Car interior noise	45.82	72.98	77.58	72.68	93.68	95.53	94.47	98.35	98.81
Average	**17.45**	**56.27**	**88.49**	**25.17**	**60.78**	**92.96**	**44.59**	**86.65**	**96.10**

27.21 ms (300 points) with an overlap of 9 ms (100 points) between successive frames. The number of word was limited into 72. A 20-state Hidden Markov Model (HMM) was used for speech recognition. In the stage of training, speech data from 10 male and 10 female Thai speakers were used. Each person uttered 72 Thai words twice. In the stage of recognition, we have used 5 unspecific speaker of the male and female in the experiment. We explore the noise robust property of total system and thus several circumstances of noise were considered with SNR ranging from 0 dB to 20 dB. 12 additive noises were selected from NOISEX-92. The test were done with DRA only and both RSF and DRA.

Several experiments are carried out with these conditions. Each speech feature vectors has 34-dimensional parameters consisting of 16 parameters of DCT (log(CBI)) ,16 delta of DCT(log(CBI)), a parameter of logarithm power and a parameter of delta logarithm power. Recognition results are shown in Table 1.

4.2 Recognition results

In Table 1 all values represent average accuracy rate (%) of experiment recognition. “DCT (log(CBI))” means a simple speech recognition method in which there is no noise robust algorithm. “DCT (log(CBI)) with DRA” means that only DRA is applied in speech feature recognition. “DCT (log(CBI)) with RSF/ DRA” means that both DRA after RSF are applied.

The last row on the Table 1 shows the accuracy averaged from several types of noise from 0 dB SNR to 20 dB. With no noise robust technique, the average accuracy for SNR of 0 dB, 10 dB, and 20 dB SNR are above 17%, 56% and, 88% respectively. When, DRA is applied in speech features, the average accuracy improve to 7%, 4% and 4 %, respectively. The best performance is seen when both RAF and DRA are applied. The average accuracy for the latter are above 44%, 86% and 96 %, respectively.

6. CONCLUSION

In this report, we have proposed noise robust algorithm in DCT (log(CBI)). In this new speech feature, MMSE 15th predictive order to speech spectrum envelops are mapped on to the Bark scale forming eighteenth-critical bands. Logarithm of CBIs is further transformed using DCT and delta values with respect to time are used. RSF and DRA are applied to enhance the speech feature for recognition. The comparison is provided and the effectiveness of RSF and DRA in noisy speech is shown. The recognition accuracy are improved by 27%, 30% and 7% for SNR 0 dB, 10dB and 20 dB, respectively.

REFERENCES

[1] S. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-27, no. 2, pp. 113–120, 1979.

[2] H. Hermansky and N. Morgan, “RASTA processing of speech,” IEEE Trans. Speech Audio Process., vol. 2, pp. 578–589, 1994.

[3] Naoya Wada, Noboru Hayasaka, Shingo Yoshizawa and Yoshikazu Miyanaga, “Robust Speech Recognition with Feature Extraction Using Combined Method of RSF and DRA”, ISCIT2004, pp. 1001-1004, 2004.

[4] N. Hayasaka, Y. Miyanaga and N. Wada, “Running spectrum filtering in speech recognition,” SCIS Signal Processing and communication with Soft Computing, Oct 2002

[5] L. Rabiner and B.H. Juang, “Fundamentals of Speech Recognition”, Prentice Hall, 1993.

[6] L.R Rabiner, “A Tutorial on Hidden Markov Models and Selected Application in Speech Recognition”, Proceedings of the IEEE, vol. 77, no. 2 , pp. 257-287, Feb. 1989.

[7] H. Hermansky, “Perceptual linear predictive (PLP) analysis for speech”, J. Acoust. Soc. Amer., Vol. 87, No. 4, pp. 1738 – 1752, 1990.

[8] E. Zwicker, H. Fastl, “Psychoacoustics: Facts and Models” Second Edition, Springer, pp 158-170, 1999.

[9] Smith, J. O. and Abel, J. S., “Bark and ERB Bilinear Transforms”, IEEE Trans. Speech & Audio Proc., vol. 7, no. 6, pp. 697-708, Nov. 1999.

[10] L. R. Rabiner and R. W. Schafer, “Digital Processing of speech signal”, Prentice-Hall, p443, 1978.

[11] Sadaoki Furui, “Digital Speech Processing, Synthesis and Recognition”, Second Edition, Marcel Dekker, 2001.

[12] K. Songwatana, K. Dejhan, Y.Miyanaga, K. Khanthavivone “A vowels recognition model for Laotian language using transfer function on bark scale and hidden Markov modeling”, NSIN 2005, pp. 36-39, May 2005.

[13] W. Joseph, Picone, “Signal Modeling Techniques in Speech Recognition”, IEEE, Vol. 81, NO. 9, September 1993.