

CITY UNIVERSITY OF HONG KONG  
香港城市大学

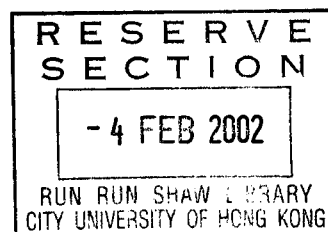
Reliable Bands Guided Similarity Measure  
for Noise-Robust Speech Recognition  
抗噪声语音识别中的一种基于  
可靠频段的相似性度量

Submitted to  
Department of Electronic Engineering  
电子工程学系  
in Partial Fulfillment of the Requirements  
for the Degree of Doctor of Philosophy  
哲学博士学位

by

ZHANG Bo  
张波

February 2001  
二零零一年二月



# Abstract

Under noisy conditions, due to the redundancy of the speech signal, there are some spectral bands (Reliable Bands) whose local SNR's are high enough to be used effectively by a recognizer. A novel, phonetically motivated Reliable Bands Guided similarity measure (RBG measure) is proposed in this study. It has the following features. Firstly, for reference spectrum, frequency bands which have larger absolute energy or sharper spectral peaks are marked as reliable bands. They are to be given more weight than the other bands in the definition of the RBG measure. Secondly, within each reliable band, similarity between formant positions and formant shapes of test spectrum and reference spectrum is explicitly modeled. Thirdly, to accommodate the phoneme restoration phenomenon, the similarity scores for some Mandarin consonants are calculated separately, by Induction Models. Lastly, the measure can automatically emphasize spectral bands whose amplitudes change abruptly, which normally contain more reliable dynamic features of the speech signal. Both the RBG measure and the Parallel Model Combination (PMC) method are tested on a speaker-independent, continuous Mandarin digit string recognition task, under 15 noisy conditions. Noises are drawn from the NOISEX-92 database. The RBG measure shows an average 4.22% word accuracy score below the PMC method above 0 dB. However, it outperforms the PMC method by 8.82% at 0 dB. More importantly,

the RBG measure does not rely on accurate background noise modeling, which is a difficult task in itself.

# Contents

<b>Abstract</b>	<b>i</b>
<b>Acknowledgement</b>	<b>iii</b>
<b>Contents</b>	<b>v</b>
<b>List of Figures</b>	<b>viii</b>
<b>List of Tables</b>	<b>xiii</b>
<b>Notation</b>	<b>xv</b>
<b>1 Introduction</b>	<b>1</b>
1.1 General Situation of Speech Recognition . . . . .	1
1.2 Speech Recognition Technologies . . . . .	3
1.3 Scope of this Study . . . . .	8
<b>2 Computational Auditory Scene Analysis</b>	<b>10</b>
2.1 Overview of Bregman's Work . . . . .	10
2.2 Data-Driven Systems . . . . .	14
2.3 Prediction-Driven Approach . . . . .	18

2.4	Speech Recognition based on Auditory Scene Analysis . . . . .	26
<b>3</b>	<b>Noise-Robust Techniques</b>	<b>28</b>
3.1	Robust Speech Representations . . . . .	28
3.1.1	RASTA-PLP . . . . .	29
3.1.2	Cepstral Mean Normalization . . . . .	32
3.1.3	Segmental Cepstral Mean Normalization . . . . .	32
3.2	Recognition Based on Partial Information . . . . .	33
3.2.1	Noise Masking . . . . .	34
3.2.2	Missing Data Techniques . . . . .	35
3.3	Robust Distance Measures . . . . .	40
3.3.1	Cepstral Projection Measure . . . . .	40
3.3.2	Factored Spectrum Similarity Measure . . . . .	41
3.4	Noise Compensation Methods . . . . .	44
3.4.1	Spectral Subtraction . . . . .	44
3.4.2	Wiener Filtering . . . . .	45
3.4.3	Joint Cepstral Compensation . . . . .	49
3.4.4	Stochastic Matching . . . . .	54
3.4.5	Model Composition and Decomposition . . . . .	56
3.5	Discussion . . . . .	60
<b>4</b>	<b>The Reliable Bands Guided Similarity Measure</b>	<b>63</b>
4.1	Generation of Template Set . . . . .	63
4.1.1	Monte Carlo Generation . . . . .	66
4.1.2	Generation of Template Set . . . . .	71
4.2	The RBG Similarity Measure . . . . .	78
4.2.1	Spectral Dynamics . . . . .	80
4.2.2	Calculation of the Min-Vector . . . . .	82

4.2.3	Reliable Bands of the Input Signal . . . . .	83
4.2.4	Definition of the Reliable Bands Guided Similarity Measure . . . . .	86
4.3	Other Techniques for Noise-Robustness . . . . .	96
4.3.1	Induction Modeling . . . . .	97
4.3.2	Duration Modeling . . . . .	99
4.3.3	Voiced/Unvoiced Feature . . . . .	105
<b>5</b>	<b>Experimental Results</b>	<b>110</b>
5.1	Database Preparation . . . . .	110
5.2	Baseline System . . . . .	117
5.3	Implementation of the Parallel Model Combination Method . . . . .	122
5.4	Experimental Setup of the RBG Similarity Measure Method . . . . .	128
5.5	Experimental Results . . . . .	130
<b>6</b>	<b>Conclusions and Future Work</b>	<b>135</b>
6.1	Conclusions . . . . .	135
6.2	Future Work . . . . .	136
	<b>Bibliography</b>	<b>139</b>

# List of Figures

1.1	General training and recognition procedure (based on Sankar and Lee 1996).	7
2.1	Demonstration of the old-plus-new rule in speech perception (reproduced from Bregman and Ahad, 1995, demonstration 37). Intensity of each harmonic is represented by the darkness of the harmonic.	13
2.2	Competition of the sequential grouping rule and the onset-offset synchrony rule (reproduced from Bregman and Ahad, 1995, demonstration 27).	14
2.3	Block diagram of the data-driven segregation system by Brown and Cooke (1994).	15
2.4	Blackboard-based architecture of the prediction-driven approach (after Ellis, 1996).	20
2.5	Correlogram slice sampled at a particular time $t$ of a voiced segment in male speech.	21
2.6	Signal model of the noise-cloud element.	22
2.7	Signal model of the click element.	23

2.8	Sound events detected by the prediction-driven approach (after Ellis, 1996). . . . .	25
3.1	Diagram of the PLP analysis (the upper part) and the RASTA-PLP analysis (the lower part). . . . .	29
3.2	LPC spectrograms (the upper ones) and J-RASTA-PLP spectrograms (the lower ones) of a Mandarin utterance “3 9 1y 2 2 7” [san tɕiou iou əɾ əɾ tɕi] under a clean condition (the left ones) and a noisy condition (the right ones). The digit “1” may be pronounced as either [i] or [iou]. In the latter case, we represent it as “1y” as shown here. The noise is the “hfchannel” noise which is recorded from an HF radio channel (to be described in section 5.1). $J = 10^{-6}$ is used for generating the J-RASTA-PLP spectrograms. . . . .	31
3.3	The interpolating function $f(\text{SNR}_i)$ with $\alpha = 1.0$ and $\beta = 3.0$ . . . . .	52
3.4	General training and testing procedure (based on Sankar and Lee 1996, reproduced from figure 1.1 for convenient viewing). . . . .	55
3.5	The concept of three dimensional Viterbi decoding (after Varga and Moore 1990). . . . .	57
3.6	The general principle of the Parallel Model Combination method (after Gales 1995). . . . .	58
3.7	Non-iterative Parallel Model Combination (after Gales 1995). . . . .	59
3.8	Data-driven Parallel Model Combination method (after Gales 1995). . . . .	61
3.9	Wideband spectrogram of noises recorded in a bar-room. . . . .	62
4.1	The overall procedure to generate template set. . . . .	65
4.2	20,000 samples generated by the Monte Carlo method for the two-dimensional random vector $Z$ , with each sample being represented by a dark pixel in the figure. . . . .	71

- 4.3 A template generated from the second state of the triphone HMM  $s-an+\text{t}\text{c}$ . The HMM has three emitting states and is trained from male speech. Both the reference log-spectrum  $S(\omega)$  and the frequency weighting vector  $F(\omega)$  have 26 elements, with each element corresponding to one mel-frequency channel. . . . . 79
- 4.4 The min-vector of a female utterance “3 9 1y 2 2 7” [san  $\text{t}\text{c}$ iou iou  $\text{a}\text{r}$   $\text{a}\text{r}$   $\text{t}\text{c}$ ' i] recorded under clean condition. . . . . 84
- 4.5 The min-vector (solid line) of a female utterance “3 9 1y 2 2 7” [san  $\text{t}\text{c}$ iou iou  $\text{a}\text{r}$   $\text{a}\text{r}$   $\text{t}\text{c}$ 'i], corrupted by the “hfchannel” noise at SNR = 6 dB. The average log-spectrum of the noise is shown as the dashed line. . . 84
- 4.6 Determination of the reliable bands of the input signal. **(a)** Log-spectrogram of a female utterance “3 9 1y 2 2 7” [san  $\text{t}\text{c}$ iou iou  $\text{a}\text{r}$   $\text{a}\text{r}$   $\text{t}\text{c}$ 'i] recorded under clean condition. **(b)** Log-spectrogram of the same utterance corrupted by the HF radio channel noise at SNR = 6 dB. Each dashed vertical line labels the onset of the word which is attached to the line. The frequency axis is in the mel-scale. Poles which are selected into the vector  $P_t(k)$  are marked by the diamonds. 87
- 4.7 Illustration of the calculation of the center frequency of the nearest formant to  $\omega$  in the reference log-spectrum  $\hat{S}(\omega)$  and the input log-spectrum  $O(\omega)$ . . . . . 89
- 4.8 The function  $f_P(x)$  with  $\omega_b = 0.3$ ,  $\theta = 10.0$  (the solid line) or  $\theta = 40$  (the dashed line). . . . . 90
- 4.9 The function  $f_E(x)$  with  $\alpha = 0.1$ ,  $\beta = 0.8$  (the solid line) or  $\beta = 2.5$  (the dashed line). . . . . 92
- 4.10 The function  $f_E(x)$  with  $\beta = 0.8$ ,  $\alpha = 1.0$  (the solid line) or  $\alpha = 2.0$  (the dashed line). . . . . 92

- 4.11 The RBG similarity measure between an input log-spectrum  $O(\omega)$  of a noisy [tʃ] sound, and a template generated from a state of [tʃ]’s triphone model. The similarity function  $\rho(\omega)$  is reduced by a factor of 20 for clearer view. Formant positions of  $O(\omega)$  and  $\hat{S}(\omega)$  are labeled by  $P_l(k)$  and  $P_r(k)$  respectively. . . . . 94
- 4.12 The RBG similarity measure between the input log-spectrum  $O(\omega)$  of the noisy [tʃ] sound, and another template generated from sound [a], without making use of the min-vector (i.e., setting  $O_{min}(\omega_k)$  to zero in Eq. (4.26)). Again, the similarity function  $\rho(\omega)$  is reduced by a factor of 20 for clearer view. . . . . 95
- 4.13 The optimal RBG similarity scores calculated for all frames in a female utterance “3 9 1y 2 2 7” [san tʃiou iɔu əɾ əɾ tʃi]. The utterance is corrupted by the HF radio channel noise at SNR = 6 dB. . . . . 96
- 4.14 Duration distributions of the second states of the base phones [p] (the solid line) and [əɾ] (the dashed line), with their “expectation durations” marked by the vertical dashed lines. . . . . 103
- 4.15 The utterance “7” [tʃi] (transcribed by the upper label) is mis-recognized as “7 4” [tʃi sɪ] (transcribed by the lower labels). The longer dashed lines label the onsets of the word “7” and “4”, while the shorter dashed lines label the onsets of 7’s phoneme [i] and 4’s phoneme [ɪ]. The utterance is corrupted by the white noise at SNR = 6 dB. . . . . 106
- 4.16 The narrow-band spectrogram and the pitch contour of a female utterance “3 9 1y 2 2 7” [san tʃiou iɔu əɾ əɾ tʃi], which is corrupted by the babble noise at SNR = 6 dB. Each vertical line labels the onset of the word which is attached to the line. . . . . 108

4.17	The narrow-band spectrogram and the pitch contour of a female utterance “2 1y 1y 9” [ər iəu iəu tɛiəu], which is corrupted by the HF radio channel noise at SNR = 0 dB. . . . .	109
5.1	Wideband spectrograms of the noises. For each noise source, only the segment in [1 second ~ 3 second] is shown. . . . .	113
5.2	Wideband spectrograms of the noises (continued). For each noise source, only the segment in [1 second ~ 3 second] is shown. . . . .	114
5.3	LPC spectrogram of a female utterance “9 5 0 9 0 3 8” [tɛiəu u liŋtɛiəu liŋsan pa]. Each dashed vertical line labels the onset of the word which is attached to the line. . . . .	118
5.4	LPC-spectrogram of the digits “1” (the left one) and “4” (the right one) of a male speech, uttered in isolation. The second formant of 4’s phone [ɹ] is much lower than that of 1’s phone [i]. . . . .	118
5.5	Performance comparison between the baseline system, the DPMC method and the RBG measure method. The average word accuracy rates in the figure are taken from table 5.2, table 5.4 and table 5.5. . . . .	134

# List of Tables

1.1	Scope of this study. In the table, <i>perplexity</i> is defined as the average number of possible words following a particular word for a recognition task. . . . .	9
3.1	Word error rates of the Factored Spectrum Similarity Measure method (taken from Kopec and Bush 1989). The “baseline” system is a recognizer in which the similarity measure between an input LPC-spectrum and a reference LPC-spectrum is simply defined as the Euclidean distance between the two cepstra. . . . .	43
3.2	Word accuracy rates (in percentage) of the cepstral compensation algorithms (taken from Acero 1993). . . . .	54
5.1	Recognition results obtained when the mixture number is increased.	121
5.2	Word accuracy rates (%) of the baseline system tested on the noisy speech database. . . . .	123
5.3	Phoneme recognition accuracy rates (in percentage) for different values of $N$ . . . . .	129

- 5.4 Word accuracy rates (%) of the DPMC method when tested under the clean condition and the noisy conditions. Under each of the three SNRs, the noises are sorted in descending order by the word accuracy rates associated with them, with the average word accuracy rate shown in the last row. . . . . 133
- 5.5 Word accuracy rates of the RBG similarity measure method when tested under the clean condition and the noisy conditions. . . . . 134