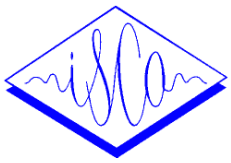

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

COMPARATIVE STUDY OF SPECTRAL REPRESENTATIONS IN MEASURING THE ENGLISH /r/-/l/ ACOUSTIC-PERCEPTUAL DISSIMILARITY

Kiyoaki Aikawa and Reiko A. Yamada

ATR Human Information Processing Research Laboratories
2-2 Hikaridai, Seikacho, Sorakugun, Kyoto 619-02 Japan

ABSTRACT

The talker dependency of the correct response rate (CRR) on English /r/-/l/ identification by Japanese listeners has already been reported. This paper shows that the talker dependency of the CRR can be explained by the acoustical dissimilarity (ADS) between an /r/ and an /l/ measured by the dynamic-cepstrum. The dynamic-cepstrum is a new spectral representation which simulates time-frequency forward masking. Nine spectral representations including weighted-cepstrum, mel-cepstrum, and delta-cepstrum were compared in terms of correlation between the CRR and the ADS. The ADS measured by the dynamic-cepstrum showed the best correlation with the CRR. The experimental results imply that Japanese listeners tend to identify /r/ or /l/ using the succeeding vowels affected by co-articulations.

1. INTRODUCTION

The correct response rate (CRR) for English /r/-/l/ identification by Japanese listeners depends on the talker of the database [1, 2]. If the difference in CRRs is caused by the acoustical properties of the stimuli, the talker dependency can be explained by measuring the acoustical difference between an /r/ and an /l/. Each member of an /r/-/l/ pair having a larger acoustical difference is assumed to be more easily identified. To measure the acoustical difference, it is necessary to use a spectral representation capable of simulating the auditory response. One such spectral representation is the dynamic-cepstrum.

The dynamic-cepstrum is a new spectral representation which incorporates time-frequency two-dimensional forward masking and provides excellent automatic speech recognition performance [3]. The time-frequency masking model was derived from experimental results on masking patterns evoked by a pure tone masker [4]. The dynamic-cepstrum can emphasize formant transitions and voice onsets like an auditory system. This feature simulating an auditory-specific response suggests the applicability of the dynamic-cepstrum to talker dependency analysis [5]. This paper compares the performance of the dynamic-cepstrum with that of conventional spectral representations in measuring the /r/-/l/ acoustical dissimilarity (ADS). The compared representations are: cepstrum, weighted-cepstrum [6], mel-cepstrum [7], delta-cepstrum [8, 9], and their combinations.

2. DYNAMIC-CEPSTRUM

The dynamic-cepstrum is derived from a time-frequency masking model. A masking pattern evoked by a pure tone shows a relatively sharp peak at the masker frequency for a short period immediately after the masker is turned off. The masking pattern attenuates and loses its sharpness as a function of the elapsed time after the turn-off time. This process can be modeled by a spectral smoothing lifter as a function of the masker-signal time interval [5]. Thus, the masking pattern evoked by a speech sound is obtained by the sum of the preceding spectra smoothed depending on the time delay between the preceding and the current times. The masked spectrum is obtained by subtracting the masking pattern from the current instantaneous spectrum.

The dynamic-cepstrum is defined as the inverse Fourier transform of the masked spectrum. Given the kth cepstrum time-sequence $c_k(i)$, the kth dynamic-cepstrum at time i is given by

$$b_k(i) = c_k(i) - m_k(i) \quad (1)$$

$$m_k(i) = \sum_{n=1}^{N} c_k(i-n) l_k(n) \quad (2)$$

where $m_k(i)$ denotes the inverse Fourier transform of the masking pattern. $l_k(n)$ denotes the lifter gain to be multiplied to the kth cepstrum coefficient at n frames before. N limits the masking duration. A Gaussian shape is employed for the lifter gain, which is given by

$$l_k(n) = \alpha \beta^{n-1} \exp\left(-\frac{k^2}{2(q_0 - \nu(n-1))^2}\right) \quad (3)$$

where the constant q_0 denotes the standard deviation of the Gaussian and ν its decreasing rate. The constant β is the masking decay rate and α is the initial decay.

3. DELTA-CEPSTRUM

The delta-cepstrum is a conventional dynamical feature parameter [8, 9]. The kth delta-cepstrum at time i is obtained by

$$\Delta c_k(i) = \frac{\sum_{n=-L/2}^{L/2} w(n) c_k(i+n) n}{\sum_{n=-L/2}^{L/2} w(n) n^2} \quad (4)$$

where $w(n)$ is a temporal window of size L. The delta-cepstrum represents only the transitional component of a running spectrum. Therefore, the delta-cepstrum distance $D_{\Delta Cep}$ is used in combination with an instantaneous spectral distance such as the cepstral distance D_{Cep} as

$$D = (1-\lambda) D_{Cep} + \lambda D_{\Delta Cep} \quad (5)$$

where λ is the combination weight.

Table 1. Number of minimal pairs in each 5 phonetic context: initial singleton position (IS), initial consonant cluster position (IC), intervocalic position (IN), final singleton position (FS), and final consonant cluster position (FC).

Category	Phonetic Context	Exp. 1	Exp. 2
IS	r/l VC	21	13
IC	C r/l V..	32	24
IN	..V r/l V..	6	5
FS	..V r/l	19	15
FC	..CV r/l C	15	11

4. BASE PARAMETER

The base parameter is defined here as the spectral parameter from which the dynamic-cepstrum and the delta-cepstrum are calculated. The cepstrum is a base parameter.

Another possible base parameter is a weighted-cepstrum, which represents a spectrum with emphasized peaks [6]. The kth weighted-cepstrum is given by

$$g_k = \begin{cases} k\,c_k & k < k_s \\ k_s\,c_k & k \geq k_s \end{cases} \qquad (6)$$

where c_k is the kth cepstrum coefficient. k_s is the quefrency threshold which makes the weight saturate to avoid over emphasis on noisy spectral components. In this paper, k_s is fixed to 5 [6].

A mel-cepstrum can be yet another base parameter. The mel-cepstrum represents a spectrum on the mel frequency scale. A mel-cepstrum can be calculated from the cepstrum by the phase function of the bi-linear transform [7].

Nine spectral parameters are derived by the combination of three base parameters and three spectral parameters: dynamic-cepstrum, delta-cepstrum, and the original base parameter. For example, a weighted dynamic-cepstrum is one of the nine parameters.

5. PERCEPTION EXPERIMENT

5.1. Speech Materials

The speech database consisted of 93 English minimal-pair words contrasting /r/ and /l/ [1]. Words in each pair contrasted /r/ and /l/ in five positions as shown in Table 1. This word set was spoken by five native talkers of American English. The speech database was the same as that used in [1, 2, 10]. Talkers 1,3 and 5 were male and talkers 2 and 4 were female.

Two kinds of /r/,/l/ identification experiments were conducted. One is the experiment without training and the other with training [2]. The subjects were required to listen to single words and to answer what word they heard by selecting one of the members of a minimal pair (e.g., "red" or "led" with stimulus /rɛd/) after listening to a stimulus through a headphone.

5.2. Experiment 1: without training

The stimulus sound was given to listeners in a random order from the 93 pairs uttered by five talkers (930 trials in total). The subjects were 11 native Japanese who had never resided in a foreign country. The average CRR was obtained for each talker. Figure 1 shows the difference of the average CRRs among the talkers. This talker dependency pattern is similar to that reported in [1]. This figure shows that Japanese listeners should be able to relatively easily identify /r/s and /l/s uttered by talker 4.

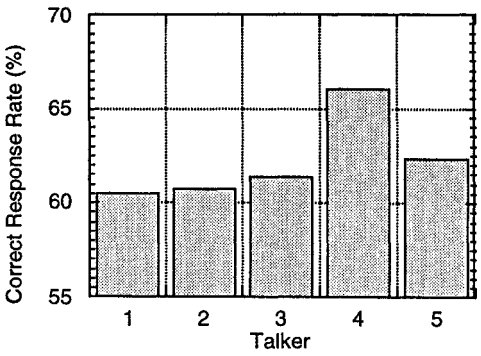

Figure 1. Talker dependency pattern of the correct response rate in English /r/-/l/ identification by Japanese listeners (Exp. 1, all phonetic contexts).

5.3. Experiment 2: with training

In order to verify the result, data was also obtained from identification experiments with training [2]. In one training session, 68 minimal pairs uttered by a single talker were presented. The same stimulus appeared twice (272 trials in total) in random order. Each subject received 45 training sessions resulting from nine series of five training sessions. The five talkers appeared in a fixed order from talker 1 to talker 5. The stimulus speech sounds were the same as those used in Exp. 1. The subjects were 13 native speakers of Japanese different from the subjects in Exp. 1.

It was found that the CRR improves through the training. In order to normalize this trend, the training effect was removed from the CRR. Let u_i denote the CRR at session i, then the logarithmic error rate,

$$\ln(1 - u_i) \qquad (7)$$

linearly decreases. Therefore, the identification error can be approximated by an exponential decay function.

Let v_i denote the log error rate after removing the training effect. Since a talker appears once every five sessions in the training session sequence, the estimated CRR before training for talker j is given by

$$x_j = 1 - \exp\left(\frac{1}{9}\sum_{k=1}^{9} v_{((k-1)\times 5+j)}\right) \qquad (8)$$

$$v_i = \ln(1 - u_i) + \kappa i \qquad (9)$$

where κ denotes the average error reduction rate per session. The κ was 0.02412 for the phonetic contexts IS, IC and IN (Table 1). The suffix k is the repetition number for each talker.

The trend-compensated CRR for this experiment is shown in Figure 2. This talker dependency pattern is mostly identical to that obtained in Exp. 1.

6. ACOUSTICAL DISSIMILARITY

The acoustical dissimilarity (ADS) of an /r/-/l/ minimal pair is measured by the average matching distance of the minimal pair using a dynamic time warping (DTW) algorithm [11]. Spectral dissimilarity is given by the Euclidean distance between corresponding frame parameters. If a minimal pair is spoken by the same talker in the same recording environment, the matching distance reflects only the difference between the /r/ and the /l/ in the minimal pair.

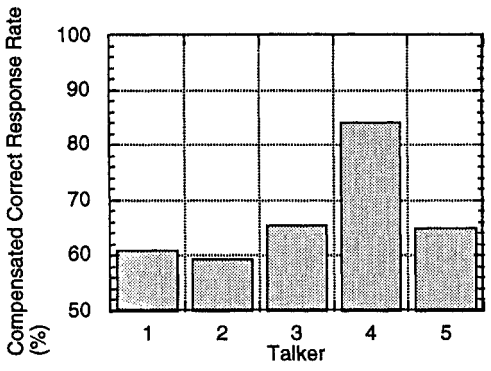

Figure 2. Estimated CRRs by compensating the training effect for phonetic contexts IS, IC and IN (Table 1) (Exp. 2).

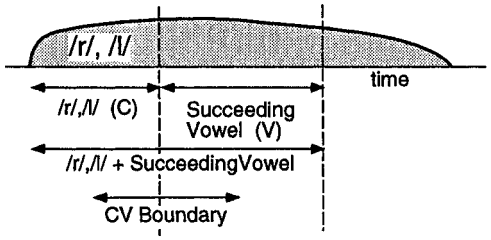

Figure 3. Portions where the acoustic dissimilarity of a minimal pair are measured.

The speech data was sampled at 10 kHz. The spectrum was estimated through 16th order linear prediction analysis every 10 ms. A spectrum was represented by a 32nd order cepstrum. All speech data was converted to three base parameters: LPC (Linear Predictive Coding) cepstrum, weighted-cepstrum, and mel-cepstrum. The dynamic-cepstrum and delta-cepstrum were calculated from each of these three base parameters. The dynamic-cepstrum time-sequences were calculated from the 32nd order cepstrum time-sequences under the conditions $N = 4$, $q_0 = 18$, $\nu = 1$, $\alpha = 0.3$, and $\beta = 0.7$ [3]. The temporal window size for the delta-cepstrum was 70 ms and its shape was isosceles triangular. The cepstrum/delta-cepstrum combination weight was $\lambda = 0.95$.

7. CRR-ADS CORRELATION

7.1. Analysis of Experiment 1

The ADS was measured for all of the /r/-/l/ minimal pairs and was averaged for each talker. The talker dependency pattern of the ADS measured using the cepstrum parameter was not similar to that of the CRR shown in Figure 1. On the other hand, the talker dependency pattern of the ADS measured using the dynamic-cepstrum showed a high correlation with that of the CRR.

The following part investigates partial dissimilarities of the minimal pairs in the phonetic contexts IS, IC and IN (Table 1); each minimal pair shows a good coincidence at the speech portions other than the /r/ and /l/ portions.

The ADS was measured for whole-word and various portions: /r/ or /l/, its succeeding vowel, /r/ or /l/ with the succeeding vowel, and the boundary region from the /r/ or /l/ to the succeeding vowel. In this paper, the boundary between /r/ or /l/ and a succeeding vowel is defined as the approximate midpoint of the transition from /r/ or /l/ to the vowel. Figure 3 illustrates these portions schematically.

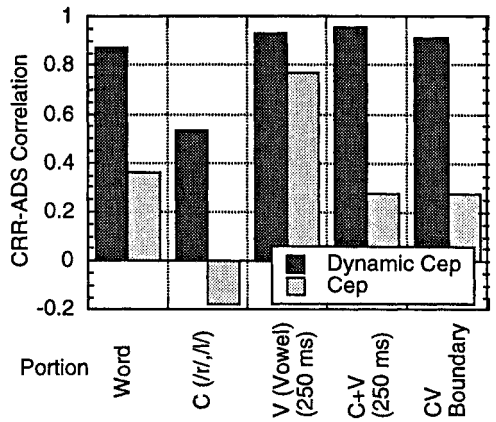

Figure 4. Correlation between partial ADS and CRR (Exp. 1).

Figure 4 compares the correlations between the CRR and the ADS for each of the portions described above. The succeeding vowel period was 250 ms. The CRR-ADS correlation is high at the portion that includes an /r/ or /l/ and the succeeding vowel. The dynamic-cepstrum shows a better correlation than the cepstrum for every portion.

The nine spectral representations were compared in terms of how well the measured ADS values matched the results of the perceptual experiments. CRR-ADS correlations were analyzed for the portion of /r/,/l/ followed by various lengths of succeeding vowel regions.

Figure 5 plots the CRR-ADS correlation at the CV portion using cepstrum as the base parameter. Figures 6 and 7 show CRR-ADS correlations when using the following base parameters: the weighted-cepstrum and mel-cepstrum, respectively.

These results demonstrate that the ADS measured using the dynamic-cepstrum shows a very good correlation with the perceptual CRR. The combination of cepstrum and delta-cepstrum (Cep+ΔCep) shows a higher CRR-ADS correlation than the cepstrum; however, the correlation is lower than the dynamic-cepstrum's. This relation is common among the three base parameters: cepstrum, weighted-cepstrum and mel-cepstrum.

The CRR-ADS correlation becomes higher as the attached succeeding vowel period lengthens, and indicates a highest value at the vowel length of 250 ms. This implies that an /r/-/l/ minimal pair that is easily identified by Japanese shows the spectral difference over a long period after /r/ or /l/.

7.2. Analysis of Experiment 2

The Experiment 2 studied the correlation between the estimated CRR in the training experiment and the ADS for the minimal pairs in the phonetic contexts IS, IC and IN (Table 1). Figure 8 shows the CRR-ADS correlation for the /r/,/l/+succeeding vowel portion using cepstrum as the base parameter. The obtained results are almost the same as those in Exp. 1, although the subjects in this perceptual experiment were different from those in Exp. 1.

8. CONCLUSIONS

The acoustical /r/-/l/ dissimilarity (ADS) measured by the dynamic-cepstrum showed a very good correlation with the correct response rate (CRR) in the identification of American English /r/ and /l/ for native speakers of Japanese. The CRR-ADS correlation using the dynamic-cepstrum was higher than those measured using other conventional spectral parameters. This result implies that the dynamic-cepstrum has the potential to explain the talker

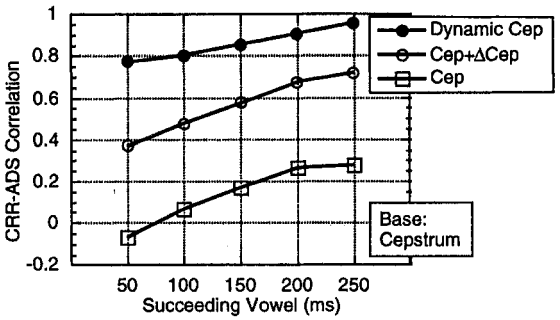

Figure 5. Correlation between CRR and ADS for /r/,/l/ + succeeding vowel using cepstrum as the base parameter (Exp. 1).

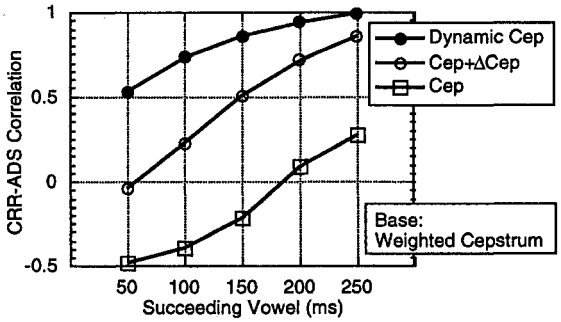

Figure 6. Correlation between CRR and ADS for /r/,/l/ + succeeding vowel using weighted-cepstrum as the base parameter (Exp. 1).

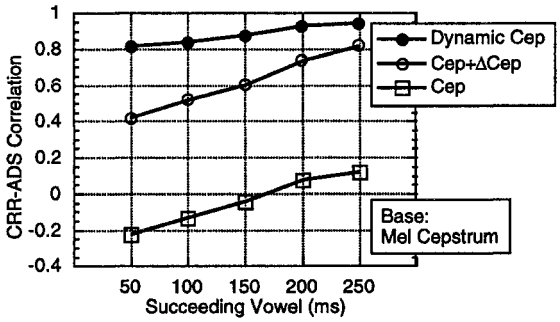

Figure 7. Correlation between CRR and ADS for /r/,/l/ + succeeding vowel using mel-cepstrum as the base parameter (Exp. 1).

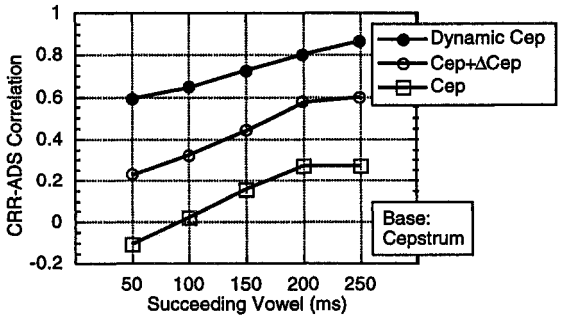

Figure 8. Correlation between compensated CRR and ADS for /r/,/l/ + succeeding vowel (Exp. 2).

dependency of the CRR obtained by a perceptual /r/,/l/ identification test. The CRR-ADS correlation was very high at the portion that included an /r/ or /l/ and a relatively long period (250 ms) after that. This suggests an interesting tendency in that Japanese listeners identify English /r/ or /l/ from a long period including the /r/ or /l/ and its succeeding vowels.

ACKNOWLEDGMENT

The authors would like to thank Prof. David. B. Pisoni of Indiana University for providing the speech database, and Dr. Yoh'ichi Tohkura for his valuable suggestions.

REFERENCES

[1] J. S. Logan, S. E. Lively and D. B. Pisoni, "Training Japanese listeners to identify English /r/ and /l/", *J. of Acoust. Soc. of Am.*, vol. 89, no. 2, pp. 874–886 (1991).

[2] R. A. Yamada, "Effect of extended training on /r/ and /l/ identification by native speakers of Japanese", *J. of Acoust. Soc. of Am.*, vol. 93, no. 4, Pt.2, p. 2391 (Apr. 1993).

[3] K. Aikawa, H. Singer, H. Kawahara and Y. Tohkura, "A dynamic cepstrum incorporating time-frequency masking and its application to continuous speech recognition", *Proc. ICASSP'93*, vol. II, pp. 668–671 (Apr. 1993).

[4] E. Miyasaka, "Spatio-temporal characteristics of masking of brief test-tone pulses by a tone-burst with abrupt switching transients", *J. Acoust. Soc. Jpn*, vol. 39, no. 9, pp. 614–623 (in Japanese) (1983).

[5] K. Aikawa, H. Kawahara and Y. Tohkura, "Dynamic cepstral parameter incorporating time-frequency masking and its application to speech recognition", *J. Acoust. Soc. Am.*, vol. 92, no. 4, Pt.2, p. 2476 (Oct. 1992).

[6] Y. Tohkura, "A weighted cepstral distance measure for speech recognition", *IEEE Trans.*, vol. ASSP-35, no. 10, pp. 1414–1422 (1987-10).

[7] A. V. Oppenheim and D. H. Johnson, "Discrete representation of signals", *Proc. of the IEEE*, vol. 60, no. 6, pp. 681–691 (1972).

[8] S. Sagayama and F. Itakura, "On individuality in a dynamic measure of speech", *Acoust. Soc. Jpn. meeting*, pp. 589–590 (1979-06).

[9] S. Furui, "Speaker-independent isolated word recognition using dynamic features of speech spectrum", *IEEE Trans.*, vol. ASSP-34, no. 1, pp. 52–59 (1986-02).

[10] K. Aikawa and R. A. Yamada, "A new masked spectrum representation applied to English /r/-/l/ dissimilarity measurement.", *J. Acoust. Soc. Am.*, vol. 94, no. 3, Pt. 2, p. 1864 (Oct. 1993).

[11] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition", *IEEE Trans.*, vol. ASSP-26, no. 1, pp. 43–49 (1978).