Corpus-based analysis of production and perception of Japanese English in view of the entire phonemic system of English

Nobuaki Minematsu[*†], **Gakuto Kurata**[*], and **Keikichi Hirose**[**]

[*]Graduate School of Information Science and Technology, University of Tokyo
†Royal Institute of Technology (KTH, Stockholm)
[**]Graduate School of Frontier Sciences, University of Tokyo
{mine,gakuto,hirose}@gavo.t.u-tokyo.ac.jp

ABSTRACT

In this paper, a method is proposed to automatically estimate the phonemic structure embedded in the pronunciation. Using two speech databases of American English and Japanese English, their structures are extracted and compared. The extraction is done only by regarding inter-phoneme relations and ignoring absolute locations of the individual phonemes in acoustic space. Comparison of the structures, however, clearly shows well-known habits observed in Japanese English. The relations also enable another interesting analysis to investigate whether the observed phonemic structure is well matched with the structure of vocabulary of English. Experimental results of the analysis show that acoustically-defined lexical density of Japanese English is approximately double that of American English.

1 INTRODUCTION

What kind of pronunciation should be pursued in language learning ? In English education in Japan, the criterion seems to have been changed from increasing acoustic similarity of the pronunciation to the native one to gaining its intelligibility. The first criterion is a sufficient condition to the second one, which is a requisite condition to the first one. Most of the pronunciation training, however, can be viewed as articulation training of the individual phonemes and this paradigm is closely related to the first criterion. Although the articulation training is required, the authors think that a method to measure the intelligibility of the pronunciation should be devised with speech technologies.

In this paper, the intelligibility of the pronunciation is defined as the easiness of accessing a metal lexicon with given utterances. Factors influencing the mental lexical access have been discussed by many researchers[1, 2], but in this paper, only the segmental factors are focused. We propose a method to estimate the phonemic structure embedded in the pronunciation[3]. The obtained structure ignores the absolute positions of the individual phonemes in acoustic space and only represents the inter-phoneme relations. This structure can be visualized as a phonemic tree diagram and it is well matched with their properties discussed in phonetics.

The phonemic structures of American English (AE) and Japanese English (JE) are compared and acoustic space of JE is found to be reduced by confusing different phonemes. The acoustic space reduction can be regarded as the lexical density increase. Using the two phonemic structures of AE and JE, a simulation is carried out to estimate the density increase based on the cohort theory of word perception. Results show that the lexical density of JE is about double that of AE.

2 ANALYSIS OF JE PRODUCTION

2.1 Speech material of JE and AE

Male speech samples of AE and JE were used. For the AE samples, WSJ speech database was used, which contained about 26,000 sentence utterances spoken by 245 speakers. As for the JE speech samples, about 8,500 utterances spoken by 68 speakers were used. The JE database was built based upon random selection of university students to avoid biased distribution of pronunciation proficiency. Recording of the JE was done using reading sheets with phonemic/prosodic symbols and was repeated until the speaker judged that he could do the correct pronunciation[4]. The resulting database still contained a large number of errors[3].

2.2 Training acoustic models of AE and JE

From the AE and JE material, HMM-based acoustic models were trained separately. In this analysis, monophones with diagonal matrices were adopted to facilitate visualization of results of the analysis. To build the HMMs, phonemic transcriptions of the training data were required, which were automatically generated by looking up PRONLEX lexicon. 12 MFCCs, 12 ΔMFCCs and Δpower were extracted as acoustic parameters. Table. 1 shows a phoneme set used here.

2.3 Phoneme tree diagrams of AE and JE

An HMM is composed of a number of states and several transitions between two states. Distance between a state and another was calculated using Bhattacharyya distance measure, which is derived based upon information theory. Two distance matrices were made for AE and JE HMM sets. These distance matrices enabled us to draw two tree diagrams of the entire

Table 1: Phoneme set used in the analysis

b, d, g, p, t, k, jh, ch, s, sh, z, zh, f,
th, v, dh, m, n, ng, l, r, w, wh, y, hh,
iy, ih, eh, ey, ae, aa, aw, ay, ah, ao,
oy, ow, uh, uw, er, ax

ISBN 1-876346-48-5

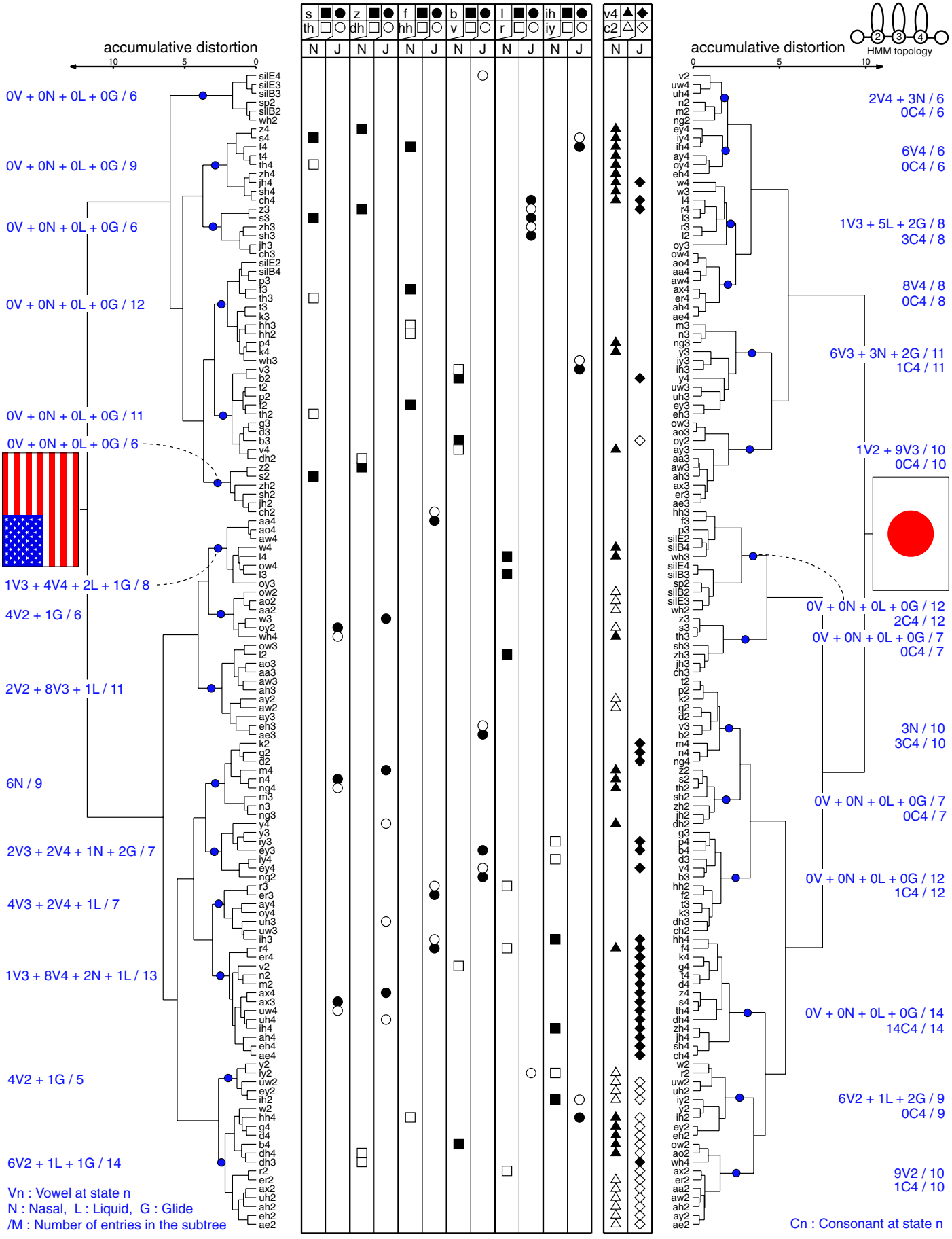

Figure 1: Two tree diagrams for the entire phonemes of American English and Japanese English

ISBN 1-876346-48-5

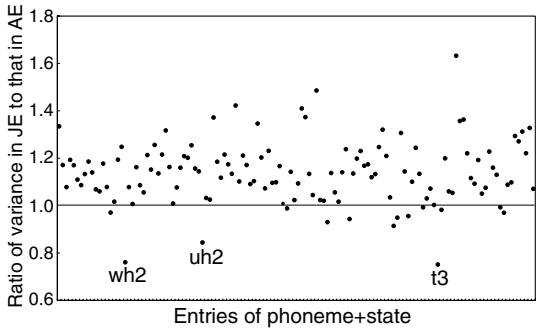

Figure 2: Ratio of variance in JE to that in AE

Table 2: Ratios of state distance in JE to that in AE

pair	s2	s3	s4	pair	s2	s3	s4
/r/&/l/	0.18	0.18	0.10	/hh/&/f/	0.31	0.27	0.58
/s/&/th/	0.09	0.03	0.10	/b/&/v/	0.94	0.79	0.40
/s/&/sh/	0.28	0.34	0.55	/ih/&/iy/	0.22	0.20	0.15
/th/&/sh/	0.23	0.32	0.49	/ih/&/y/	0.17	0.29	0.62
/z/&/zh/	0.36	0.49	0.76	/uh/&/uw/	0.26	0.23	0.30
/z/&/dh/	0.20	0.24	0.35	/ae/&/aa/	0.24	0.28	0.62
/z/&/jh/	0.21	0.37	0.62	/ae/&/ah/	0.26	0.17	0.12
/zh/&/jh/	0.31	0.32	0.56	/aa/&/ah/	0.26	0.13	0.46
/zh/&/dh/	0.27	0.31	0.38	/er/&/ah/	0.08	0.09	0.16
/dh/&/jh/	0.19	0.20	0.44	/er/&/aa/	0.17	0.12	0.26
/n/&/ng/	0.74	0.59	0.50	/er/&/ae/	0.15	0.09	0.22

phonemes of AE and JE. Ward's method, a method of hierarchical clustering, was adopted for the tree generation. Figure. 1 shows the tree diagrams of AE and JE. Leaf nodes correspond to states of the HMMs.

2.4 Comparison between the two tree diagrams
Difference in magnitude of variances: Figure. 2 shows ratios of averaged variances over MFCCs in JE to those in AE. Clearly shown, JE's variances are larger than those of AE although the JE database only contained carefully read speech. This is considered due to inter-speaker variation of pronunciation proficiency.

Phoneme pairs confusing to Japanese: State distances between phonemes confusing to Japanese are shown in Table. 2, where distances are normalized to 1.0 in the case of AE. It can be definitely said that JE's acoustic space is largely reduced and the phoneme confusions can also be seen in the tree diagrams.

Vowel insertion between consonants: State-4s of consonants and state-2s of vowels are clustered into a single subtree only in the JE tree. This is because of the well-known Japanese habit of vowel insertion.

Schwa sounds: The five nearest phonemes to schwa are obtained and shown in Table. 3. Various vowels are found in AE but most of the phonemes are mid and low vowels in JE. Considering Table. 2, these vowels are acoustically realized as a Japanese vowel of /a/.

Some other findings were also obtained by comparing the two trees but they are not described here due to limit of the space. The structural differences were also effectively used in Japanese English speech recognition. Interested readers should refer to literatures [3] and [5].

Table 3: The five nearest phonemes to schwa

state	1st	2nd	3rd	4th	5th
ax2/AE	ih2(0.68)	uh2(0.73)	d4(0.75)	ah2(0.76)	eh2(0.86)
ax3/AE	ih3(0.87)	uh3(0.88)	eh4(0.93)	ae4(0.94)	uw4(0.96)
ax4/AE	uw4(0.69)	ih4(0.72)	uh4(0.76)	ah4(0.80)	eh4(0.84)
ax2/JE	ae2(0.46)	ah2(0.51)	aa2(0.51)	ay2(0.65)	aw2(0.69)
ax3/JE	ah3(0.57)	ae3(0.61)	aa3(0.72)	aw3(0.80)	uh3(0.87)
ax4/JE	ah4(0.54)	ae4(0.61)	aa4(0.73)	aw4(0.78)	uh4(0.86)

3 ANALYSIS OF JE PERCEPTION

Reduction of the acoustic space of the JE pronunciation immediately indicates increase of the lexical density or the lexical confusedness. In this section, the confusedness is viewed as the segmental unintelligibility and, using cohort theory, one of word perception models, the increase is quantitatively estimated.

3.1 Cohort theory of word perception
Original cohort theory characterizes a human process of perceiving an isolated word as a simple left-to-right process[1]. When the initial phoneme of the input word is perceived, a set of words starting with the phoneme are activated in brain. The number of the activated words is reduced by the subsequent input of phonemes and finally reaches one, which is the end of the word perception. Cohort means a set of the activated words.

3.2 Acoustic unit used for the simulation
In this work, the intelligibility of the pronunciation is defined as the easiness of the mental lexical access and, as told above, the lexical confusedness is viewed as the segmental unintelligibility. In other words, the intelligibility is considered here as perceptual representation of goodness of the pronunciation. Cohort theory is often discussed with phonemes as its acoustic unit. To apply the theory to the simulation, however, it should be definitely more adequate to use a perceptual unit of English, which is syllables. In this case, syllabification of phoneme sequences is consequently required and it can be done with `tsylb` software[6].

3.3 Estimation of the cohort size in AE and JE
To estimate the cohort size, the entire vocabulary (lexicon) has to be defined and, in this work, PRONLEX pronunciation dictionary was used. The dictionary has about 100,000 word entries, which of course contain words whose baseforms are identical, such as walk, walked, and walking. Estimation of the cohort size is often done by using string-based distance measure and, in this case, distance between any two different phonemes is assumed to be the same. But, in the current work, the distance measure is based on acoustics, which is calculated by comparing two phoneme HMMs using Bhattacharyya distance measure.

Syllabification of all the words in the dictionary showed that there were about 10,000 different kinds of syllables located at the beginning of words. In this simulation, only the word-initial syllables were focused. For each of the different word-initial syllables, the number of

ISBN 1-876346-48-5

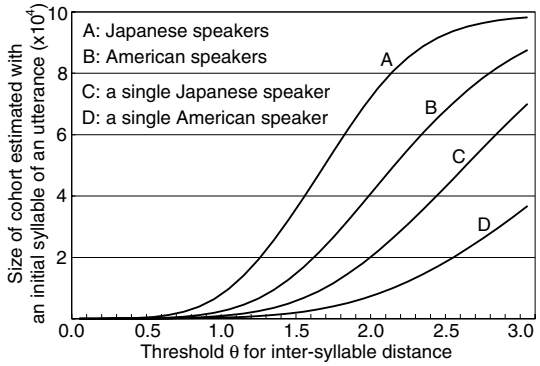

Figure 3: Estimated cohort size of AE and JE

words starting with that syllable or a syllable acoustically close to that syllable was calculated. Distance between two syllables was defined in the following manner and the acoustically close syllables were those distant only by less than threshold θ from the input syllable. A phoneme HMM is composed of 3 states and a syllable is represented as a sequence of phoneme HMMs, which means a syllable is composed of n states. Distance between two syllables was obtained by DP matching between the two state sequences, where state-to-state distance was calculated as Bhattacharyya distance. In the DP matching, local path constraint was not used because the number of phonemes in a syllable can be varied from 1 to 8. Furthermore, state transition probabilities in HMMs were also ignored. This means that temporal characteristics of phonemes were ignored. To reflect them on the DP matching, state sequences have to be converted to frame ones, which can be done with HMM-based speech synthesis techniques, if required.

3.4 Results and discussions

The simulation showed how many words are activated in brain when the initial syllable of each word of the entire vocabulary is perceived, namely, cohort size. Of course, the cohort size is calculated by a function of threshold θ. If θ is set to be smaller, the number of acoustically close syllables is decreased and the cohort size is also reduced. Smaller θ indicates more intensive listening, where a small acoustic difference is perceived by listeners but they easily get tired of listening.

Figure. 3 shows the averaged cohort size over the word entries in the PRONLEX dictionary. Four kinds of the averaged size are calculated by using A) JE speaker independent HMMs, B) AE speaker independent HMMs, C) JE speaker dependent HMMs, and D) AE speaker dependent HMMs. The first two sets of HMMs are the same as those used in building the tree diagrams. Each of the other two sets were trained with 100 sentence utterances of a single speaker. A typical Japanese student and an American teacher of English. Since the size of the dictionary is limited (about 100,000 words), the slope of the cohort size curves in cases A) and B) becomes smaller when θ is larger. Ignoring these findings caused by the experimental conditions, the cohort size of speaker independent JE models is about double that of AE models and the size of a Japanese student is also double that of an American teacher of English. If a listener tries to listen to JE while limiting the cohort size, he/she has to set θ to a smaller value, which means intensive and tough listening has to be done.

In this paper, two analyses were done with AE and JE. It should be noted that the both analyses are based upon acoustics, but not based upon acoustic matching between students and teachers. The first analysis compared two phonemic structures extracted from HMMs and the second investigated the degree of accordance between the extracted phonemic structure and the lexical structure. Further, by regarding the intelligibility as perceptual representation of goodness of the pronunciation, the cohort size (degree of the confusedness) is estimated. As told in section 1, the pronunciation training, currently done in school, is the articulation training. As some researchers pointed out[7], foreign accented pronunciation (bad articulation) dose not always decrease the intelligibility. In this meaning, the pronunciation with good (healthy) phonemic structure may be pursued in the training. At least in the two analyses, the absolute positions of the phonemes in acoustic space, in other words, information on manner and place of the articulation was completely ignored.

4 CONCLUSIONS

Two kinds of corpus-based analyses of AE and JE were carried out. By capturing only the inter-phoneme distances, the embedded phonemic structure in the pronunciation was estimated and the accordance between the phonemic structure and the lexical structure was discussed. Currently, we are doing similar analyses on 100 male and 100 female students and 20 Americans separately in the speech database, which also has pronunciation proficiency labels assigned by English teachers. We're very interested in relations between the proficiency labels and the embedded structures.

REFERENCES

[1] W. D. Marslen-Wilson *et al.*, "The temporal structure of spoken language understanding," Cognition, vol.8, pp.1–71 (1980)

[2] S. Amano, *et al.*, "Estimation of mental lexicon size with word familiarity database," Proc. ICSLP'2002, pp.2119–2122 (1998).

[3] N. Minematsu *et al.*, "Corpus-based analysis of English spoken by Japanese students in view of the Entire phonemic system of English," Proc. ICSLP'2002, pp.1213–1216 (2002).

[4] N. Minematsu *et al.*, "English speech database read by Japanese learners for CALL system development," Proc. LREC'2002, pp.896–903 (2002)

[5] N. Minematsu *et al.*, "Integration of MLLR adaptation with pronunciation proficiency adaptation for non-native speech recognition," Proc. ICSLP'2002, pp.529–532 (2002).

[6] http://www.nist.gov/speech/tools/tsylb2-11tarZ.htm

[7] J. Flege, "Factors affecting the pronunciation of a second language", Keynote of PLMA'2002 (2002).