INTELLIGIBILITY AND ACOUSTIC CORRELATES OF JAPANESE ACCENTED ENGLISH VOWELS

Diane Kewley-Port[1], Reiko Akahane-Yamada, and Kiyoaki Aikawa[2]

ATR Human Information Processing Research Laboratories
2-2, Hikaridai, Seika-cho, Soraku-gun Kyoto 619-02, Japan
e-mail: kewley@indiana.edu, yamada@hip.atr.co.jp, aik@nttspch.hil.ntt.jp

ABSTRACT

To produce near-native American English (AE) vowels, Japanese speakers must extend their five vowel system with at least six new vowels. Three experiments have been conducted to acquire both perceptual and acoustic measures about Japanese accented English (JE)vowels. Six non-back vowels of AE in several phonetic environments were recorded from four Japanese male talkers with moderate English skills. Intelligibility was assessed by a panel of six Americans as the percent of JE vowels identified as intended. The first experiment was an open-set identification task. Two vowels, /i/ and /e/, were fully intelligible (>98%) while others ranked from 81% (/ɛ/) to 23% (/ʌ/) intelligible. The second experiment used minimal-pair responses to assess intelligibility in terms of three acoustic properties of vowels, the spectral target, dynamic formants and duration. The results indicated that the spectral property was not communicated effectively, dynamics were partially effective, and duration was used effectively. In the third experiment a stepwise multiple linear regression analysis determined how these acoustic properties contributed to the intelligibility. For the two vowels analyzed, /ɪ/ and /æ/, the direction and extent of the formant movement of JE vowels in relation to the AE targets was the vowel property that contributed the most to intelligibility.

1. INTRODUCTION

To produce near-native American English (AE) vowels, Japanese speakers must acquire at least six new vowels to extend their five vowel system. This report describes new techniques to obtain specific knowledge about deficits in the production of AE vowels by Japanese talkers and the results from a limited set of vowels on the contribution of three acoustic properties to vowel intelligibility. The long-range goal is expand the data set and to use the knowledge gained to suggest training methods for Japanese to improve the intelligibility of their English.

Presumably AE vowels produced by Japanese talkers who have low intelligibility have some spectro-temporal properties that differ from native AE vowels. Acoustic properties of vowels are typically derived from spectral measures of F1 and F2. Recently Bradlow [1] described differences among the vowel productions of two languages with 5-vowel systems, Greek and Spanish, and the equivalent English vowels in an F1 X F2 plane. These *spectral* differences measured at a single time interval are but one way of characterizing the spectral-temporal properties of English vowels. *Dynamic* differences in formant movement inherent to vowel production also contribute to vowel identity ([5], [2]). In addition it is well-known that English vowels are differentiated by *durational* differences, particularly for short-long vowel pairs (e.g. /i-ɪ/). In order for Japanese to produce near-native AE vowels, the acquisition of at least the appropriate spectral, dynamic and durational properties of vowels is required. Three studies of Japanese accented vowels were conducted that include perceptual assessment of intelligibility, acoustic measurements and a description of the relation between the two.

2. SPEECH DATABASE

The first step in the proposed research was to develop a suitable database of Japanese accented English (JE) utterances. The data set was designed to facilitate comparisons with existing vowel data by Hirahara and Kato [3], Bradlow [1] and Hillenbrand [2]. Six non-back vowels of AE, /i ɪ e ɛ æ ʌ/, were selected such that the set contained three pairs that contrasted in their durational properties. Two phonetic contexts were selected, isolated vowel or vowel embedded in the syllable, /bVt/. There were two sentence contexts, none (citation form), the frame, 'I say ___ on the tape', yielding four possible utterance types and a total of 48 utterances.

The purpose of the experiment was to examine intelligibility of JE vowels. Therefore talkers were needed with some variability in English pronunciation. Following an informal screening test for spoken English, four male talkers were selected who appeared to have moderate English skills. All had from six to eight years of English studies, read English in their technical jobs, but spoke English less than one hour/day. Two repetitions of the 48 utterances were randomized for the talkers. In addition to the utterances, simple rhyming words were written next to the utterances to remind the talkers of the appropriate pronunciation of the English vowel in order to avoid reading errors. Recordings were made in an anechoic chamber with a condenser microphone (SONY ECM77). Recordings were directly digitized at a sampling rate of 22.05 kHz with 16-bit resolution through DAT-LINK+ and stored onto the hard disk of a Sparc Station. An American and a Japanese listener monitored the recordings and had the talkers repeat utterances (less than 5%) in which reading or other errors might have occurred.

---

[1] Currently at Indiana University.
[2] Currently at NTT.

2. EXPERIMENT 1: OPEN-SET IDENTIFICATION

2.1 Procedure

The purpose of experiment 1 was to assess the intelligibility of the JE vowels when utterances differed only in vowel quality. A single panel of six Americans who spoke English as their native language were listeners in all three experiments. These phonetically-naive listeners were identified from a small number available within a reasonable distance of the laboratory. The duration of their stay in Japan ranged from 2 months to 19 years. Given the possibility that they listened to JE in special way due to their experience, instructions stressed that vowels should always be judged relative to the pronunciation of AE broadcasters. While potentially the judgements of the listeners were biased, debriefing indicated that since the utterances had no meaningful context, listeners felt comfortable judging vowel pronunciation as instructed.

The stimuli for experiment 1 were the 48 utterances from each talker, plus an additional sentence type. The 'target extracted' type was just the vowel or /bVt/ syllable edited from the sentence frame. The 72 utterances from each of four JE talkers were presented three times each. NeXT workstations were used to control the experiment. Utterances were presented over STAX SR Lambda signature headphones. The responses were displayed as buttons on the screen, using keywords, such as 'ate' and 'out', for all 15 AE vowels including diphthongs and /ɚ/. Listeners were given three brief familiarization tasks with native AE vowels to learn the keyword responses. Subsequently the JE utterances were presented to the listeners blocked by talker.

Listener Responses

	i	I	e	ɛ	æ	ʌ	ɑ	ɚ
i	97	1	2					
I	64	28	6	1				
e			100					
ɛ	2		0	81	1			1
æ			15	13	49	16	20	2
ʌ			1		36	23	39	1

Table 1. Confusion matrix for vowels intended by Japanese talkers (rows) and the vowel responses from the English listeners (columns) in percent.

2.2 Results

Responses from the six individual listeners were analyzed and compared to one another in several ways. This evaluation demonstrated that performance was consistent across the listeners. Thus results of this and subsequent experiments are presented as the group average of the percent intelligibility of the vowels identified as intended. Even though 15 vowel responses were available, only eight were used as shown in the confusion matrix in Table 1. Only one response outside of the set of intended vowels was frequently used, /a/. The overall percent intelligibility was 63%. A three-way anova was conducted to examine the effects of phonetic context, sentence context and vowel type on intelligibility. Vowels in /bVt/ context were significantly more intelligible (68%) than isolated vowels (58%, $F(1,5)=8.62$, $p<0.05$). There was no effect of the three sentence contexts on intelligibility. Apparently the particular sentence frame used did not interfere with the correct production of the AE vowels, while the more natural /bVt/ syllable context did facilitate production. Intelligibility of individual vowels varied significantly ($F(5,25)=28.05$, $p<0.0001$). Two vowels, /i/ and /e/, were fully intelligible (>98%) while others ranked from 81% (/ɛ/) to 23% (/ʌ/) intelligible (Table 1). Clearly the intelligibility deficits for Japanese producing AE vowels ranged from none to severe.

3. EXPERIMENT 2: MINIMAL-PAIR IDENTIFICATION

The second experiment was a speech perception task designed to determine what spectro-temporal properties account for the reduced intelligibility of the JE vowels. The first premise underlying this perception study is that the spectral, dynamic and durational properties of AE vowels can be quantitatively defined based on the recent analyses of AE vowels by Hillenbrand et al. [2] as follows: (1) Spectral similarity of vowels based on a single, steady-state time slice; (2) Dynamic formant movement from two time slices at 20% and 80% of the vowel duration; and (3) Duration of the vowel. The six vowels from this study may be then grouped into categories for each property as follows. Vowels in a F1 X F2 plane cluster /i ɪ e/ as spectrally similar 'high' vowels, and /ɛ æ ʌ/ as similar 'mid' vowels. Formant movement is measured as the Euclidian distance in the F1 X F2 space between the 20% and 80% measures. Vowels /i e æ ʌ/ have dynamic formants (movement > 150 Hz) while /ɪ/ and /ɛ/ are static. Vowels /i e æ/ have long duration (> 110 ms) in contrast to the short duration of /ɪ ɛ ʌ/.

The second premise of this experiment is that sensitive measures of intelligibility may be obtained in minimal-pair, forced choice tasks [4]. To assess the contribution of each of the three acoustic properties to intelligibility, every pair of vowels can be labeled as similar or dissimilar for each acoustic property. For example, /ɪ-e/ are spectrally similar, have similar dynamics and are dissimilar in duration. In the perception task, one vowel is presented and listeners must choose one of the minimal-pair responses. To assess the contribution of the three vowel properties to JE intelligibility, subgroups for each property of the minimal-pair responses were formed that contrast on one of the spectral, dynamic or durational properties of vowels, but are matched on the other two properties. The mean intelligibility over all vowels judged in the subgroup measures how the contrasting property contributes to intelligibility. For example, if intelligibility is high for spectrally dissimilar subgroup, but low for the spectrally similar subgroup, then the acoustic property for spectral targets is not being effectively produced.

3.1 Procedure

Only the citation form of the /bVt/ syllables from experiment 1 was used in this experiment. Fifteen minimal-pair responses were chosen from 30 possible pairs to represent all the spectral confusions that occurred in Table 1 such that each vowel occurred 5 times in the set. Test procedures were the same as for the first experiment except responses on the NeXT display consisted of one of the minimal pairs displayed as keywords on buttons. Three new familiarization tasks with AE vowels acquainted the same listener panel with the minimal-pair responses.

3.2 Results

Intelligibility was again calculated as the percent of JE vowels correctly identified as intended. Surprisingly, overall intelligibility in experiment 2 was 94%, a 30% increase over experiment 1. Apparently listeners were much better at identifying the intended JE vowel when only one utterance type (/bVt/ in citation form) was heard and the response choice was narrowed from 15 to two. The mean intelligibility for the similar and dissimilar subgroups was calculated as shown in Table 2. The minimal-pair subgroups each consisted of four pairs of vowels. For example, for the spectral property, the pairs /i-ɪ/, /æ-ʌ/, /ɛ-ʌ/ and /ɪ-e/ were spectrally similar, and contrasted with pairs /e-ɛ/, /i-ɛ/, /ɪ-æ/ and /i-æ/ that were dissimilar, but pairs across subgroups did not contrast for the other two properties. Given the high overall intelligibility, the somewhat arbitrarily selected criterial values of 93%(high) and 85% (low) were used as to interpret the results. Vowels pairs that were spectrally dissimilar were very intelligible (98%), while similar pairs had low intelligibility (85%). This suggests that Japanese talkers were not able to produce sufficient differences between AE vowels that were spectrally similar. That is, it appears difficult for the Japanese to add new vowels to their five vowel space when the AE vowels are clustered close together. For the property of dynamic (versus static) formant movement, mean intelligibility was the same (90%) and was not remarkable since it fell between the high and low criteria. That is, according to Hillenbrand et al. [2], AE pairs like /ɪ-ɛ/ and /æ-ɛ/ have contrasting dynamics, and yet the levels of intelligibility were not very high for these pairs. This result suggests that Japanese were sometimes, but not always, able to produce the appropriate dynamic contrasts.

Property	**Mean Intelligibility (%)**	
	Similar Pairs	**Dissimilar Pairs**
Spectral	84.8	98.0
Dynamic	90.4	90.5
Duration	96.7	96.4

Table 2. Analysis of the subgroups of minimal pairs for each vowel property.

Finally, for the duration property, intelligibility was equally high, 96%, whether the pairs contrasted in vowel length or not. This result suggests that Japanese talkers produced the correct short or long durations for AE vowels. Since the Japanese language has a vowel length contrast (one versus two mora), these results imply that this contrast can be easily transferred to the correct production of short-long vowels in English. In summary, the purpose of this refined intelligibility task was to determine what vowel properties contributed to the intelligibility deficits observed in experiment 1. Results indicated that the spectral property was not communicated effectively, dynamics were partially effective, and duration was used effectively according to the perceptual judgements of the AE listeners.

4. EXPERIMENT 3: ACOUSTIC CORRELATES

4.1 Procedure

The goal of the final experiment was to discover the acoustic metrics that might correlate with the reduced intelligibility observed in the perceptual tasks. Given that a range of intelligibility must be observed in order to examine these correlations, only three vowels, /ɪ æ ʌ/, with lower intelligibility were candidates for further analysis. Therefore, it was decided to augment the data set from experiment 2 with a similar minimal-pair task concentrating on the three vowels. Additional utterance types, vowel only and /bVt/ extracted from the sentence, were also included as stimuli. The minimal-pair response task and listener panel (less one subject who did not return) from experiment 2 were the same. Only six vowel pairs were used for responses.

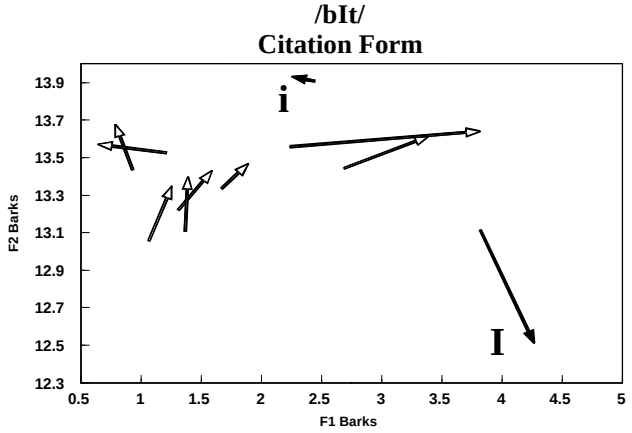
4.2 Results

Mean intelligibility was calculated for each vowel. Combined with the results of experiment 2 there were six data sets of 16 vowels each, /ɪ æ ʌ/, in the two phonetic contexts, vowel only and /bVt/. Conceptually the output of this perception task was the rank ordering the intelligibility of each utterance in a data set for use in subsequent correlations. The ordering for /ɪ/ was for the response pair /i-ɪ/ that had the most frequent confusion in Table 1. The ordering for /æ/ and /ʌ/ were both from the response pair /æ-ʌ/, vowels that were mutually confusable in Table 1.

4.3 Metrics for Acoustic Properties

In order to determine the contribution of each acoustic property to intelligibility, various acoustic metrics were examined. Analyses began with the traditional spectral target theory of vowel production as the steady-state values in a F1 X F2 plane. For the /bVt/ syllables, AE targets were calculated as averages from Bradlow's [1] measurements of /bVt/ produced in a sentence frame. For the vowel-only productions, targets were calculated from a more neutral, /hVd/, environment in citation form used by Hillenbrand et al. [2]. The Euclidian distance was calculated between each JE utterance and the corresponding AE target in the F1 X F2 plane in either Hertz or Barks. Correlations were calculated between these distance measures and mean intelligibility for the six data sets (n=16). Since a small spectral distance means that the JE vowel was close to the AE target, negative correlations were expected if spectral distance contributes to intelligibility. Two things were immediately observed. First, the correlations for /ʌ/ were positive, not negative as expected. Apparently the minimal-pair intelligibility data for /ʌ/ was flawed for the correlation analyses because the minimal-pair response for /ʌ/ was /æ-ʌ/, but Table 1 showed that /ʌ/

was most often confused with /a/, not /æ/. Therefore, /ʌ/ was dropped from further analysis. Second, the correlations were much higher for Bark than Hertz distances for /ɪ/, although nearly identical for /æ/. Given that F2 for /ɪ/ was the most affected by the Bark transform, metrics based on Barks were selected for this analysis.



The metric for dynamic formant movement was based on analyses proposed by Hillenbrand et al. [2] for a F1-F2 vector between 20% and 80% of the syllable duration. These vectors are shown on the figure for /i/ and /ɪ/ (filled arrows) for the average of the AE male vowels in Barks. The vector for /i/ is very short which was the basis for categorizing it as a static vowel. Also shown on the figure are the vectors for the eight /bIt/ syllables in citation form. It can be seen that these vectors differ both in their spectral distance from the AE vector for /ɪ/, and in the direction and extent of the dynamic formant movement. A formant dynamics metric was calculated as the scalar product between JE and AE vowel vectors after translating both vectors to the origin to remove spectral distance. Formant dynamics is a large, positive value when the direction and extent of the two vectors are similar, zero when they are perpendicular, and negative when the directions are opposite.

For duration, a simple metric of total vowel duration (ms) was the obvious choice. A comparison of the effective use of durations in the JE vowels in /bVt/ context were made with corresponding AE vowels from Bradlow [1]. Mean durations of the short-long AE pairs were statistically different (t-test for paired comparison, $p<0.05$). For the JE pairs, mean duration was significantly different for /i-ɪ/ and /e-ɛ/, but not /æ-ʌ/. Specifically the JE /æ/ vowels were produced with inappropriately short durations.

4.4 Acoustic Correlates

A stepwise multiple linear regression analysis was performed to determine how the acoustic metrics contributed to intelligibility. The three metrics of spectral distance (Barks), formant dynamics and vowel duration (ms) were input to Statistica to predict the mean intelligibility of /ɪ/ or /æ/ in the /bVt/ utterances. For /ɪ/, the regression ($R = .81$) showed that formant dynamics were only slightly more important than spectral distance, but that duration made only a small contribution towards intelligibility. For /æ/ ($R = .87$), formant dynamics were also the most important vowel property, duration contributed significantly to intelligibility and the contribution of spectral distance was not statistically significant. Thus, for this very limited data set in the context /bVt/, the direction and extent of the formant movement of Japanese accented vowels in relation to the American English targets was the most important vowel property that contributed to intelligibility. Although spectral distance by itself correlated significantly with intelligibility for both /ɪ/ and /æ/ ($r > 0.64$), it covaried sufficiently with formant dynamics such that the independent contribution of spectral distance to intelligibility was not significant. The contribution of duration to intelligibility was only significant for /æ/, where duration was sometimes inappropriate, but not /ɪ/ where duration was correctly produced. Overall, the results of this study of acoustic correlates compared to results from the perception task (experiment 2) are in partial agreement: the spectral and dynamic properties of vowels are more important to the intelligibility of JE vowel than duration. More refined comparisons await a larger body of data.

In summary, the present experiments demonstrated that vowel intelligibility may vary significantly depending on the vowels present in the native language. Results from a new intelligibility task using minimal pairs suggests that the contribution of three vowel properties to intelligibility may be assessed perceptually. Specific acoustic metrics were proposed for the vowel properties, spectral target distance, formant dynamics and duration. Results from a limited data set indicate that these properties contribute differentially to intelligibility. Altogether this investigation presents a new approach to understanding of the relation between second language intelligibility and the correlates of acoustic properties that signal phonemic distinctions.

5. REFERENCES

1. Bradlow, A.R., *Language-specific and universal aspects of vowel production and perception: A cross-linguistic study of vowel inventories.* Cornell University (unpublished manuscript), New York, 1993.

2. Hillenbrand, J., Getty, L.A., Clark, M.J., and Wheeler, K., "Acoustic characteristics of American English vowels", *J. Acoust. Soc. Am.*, 97: 3099-3111, 1995.

3. Hirahara, T., and Kato, H., "The effect of *F0* on vowel identification", In *Speech Perception, production and linguistic structure.* IOS Press, Tokyo, 1992.

4. Kent, R.D., Weismer, G., Kent, J.F., & Rosenbeck, J.C., "Toward phonetic intelligibility testing in dysarthria", *J. Speech Hear. Disorders*, 54: 482-499, 1989.

5. Nearey, T.M. "Static, dynamic, and relational properties in vowel perception", *J. Acoust. Soc. Am.*, 85: 2088-2113, 1989.