The Relationship between Language-Specific Prosodic Timing and Intersegmental Transition Speed

—Japanese and English Vowel Sequences—

Martin Gore

(2002年10月15日 受理)

Abstract. Following a review of previous studies on diphthongs and coarticulation, evidence is presented showing radical differences between Japanese and English in the temporal properties of vowel sequences which are not adequately accounted for by conventional views of vowel-to-vowel coarticulation. It is hypothesized that these differences are related to language-specific differences in timing which may be partly prosodic in nature. Why and how such differences come about are issues which are discussed, but firm conclusions must await more exhaustive empirical findings and more detailed studies of the respective vocalic and prosodic systems.

Two studies are reported. These attempt to clarify the issues using (1) spectrography to measure actual Japanese and English diphthongs, and (2) speech synthesis to test perception. One interesting result was that Japanese listeners tolerated faster VV transitions than English listeners. There are implications for our understanding of coarticulation and prosody, especially concerning mora and non-mora timing, and also for the teaching of English in Japan. Proposals are made for more thorough investigations, including articulographic studies, which are essential to validate the present results.

1. Introduction

The relationship between language-specific prosodic timing and inter-segmental transition speed is complicated and controversial. The following sections attempt to illuminate the possible relationship between prosodic timing and vowel transition speed by focussing on the behaviour of vowels in Japanese and English. Spectrography and experiments involving the perception of computer-synthesized vowels show not only that Japanese VV transitions are faster than English, but that Japanese listeners actually tolerate faster transitions between adjacent vowels. This has various implications for future research and teaching. First, however, let us look at previous research into diphthongs and coarticulation.

2. Background

2.1. Diphthong studies

Lehiste and Peterson (1961) and Holbrook and Fairbanks (1962) offered thorough descriptions of English diphthongs. Lehiste and Peterson defined the diphthong as "a vocalic syllable nucleus containing two target positions" and showed that (especially in /ai/, /au/ and /oi/) "the transitions between targets are longer than either target position." Holbrook and Fairbanks gave detailed measurements of General American formant frequencies and amplitudes at five sampling points along the time-course of each diphthong, showing the formants to be continually and gradually changing. Kent (1992) also refers to diphthongs as "dynamic sounds in which the articulatory shape (and hence formant pattern) slowly changes during the sound's production."

In similar Japanese VV contexts, however, the transition is often so rapid that the formants seem to move almost stepwise. Few studies have been done on this contrast, though Han (1962), Beckman (1982), Hoequist (1983) and Campbell (1992) have measured formants and timing in Japanese vowels, reaching conclusions that are not entirely consitent.

2.2. Coarticulation studies

Apart from the early descriptive work outlined above, relatively few references can be found in the literature concerning the coarticulatory effects of diphthong elements upon each other within the confines of the diphthong itself. The assumption seems to be that the characteristics of vowel-to-vowel movements are relatively invariable across languages, or at least show so little variation as to be unremarkable. This is odd when one considers the number of studies that have been made on the coarticulatory effects of vowels on vowels in VCV sequences in various languages, such as Ohman (1966), Butcher (1976), Recasens (1984), Farnetani (1985) or Magen (1997).

For example, Sven Ohman (1966), describing English VCV sequences showed that in iCa sequences, formants for /i/ are affected by the following /a/ and vice versa: in other words they coarticulate even though there is an intervening consonant. This would surely be even more true for sequences without the consonant. And if this is true, any restrictions on such coarticulation would be cause for remark. Yet little has been written in this area. One reason for this may be found in the view of "coarticulation" that a number of prominent researchers have taken. One definition holds that "coarticulatory effects cannot extend from vowel to vowel or from consonant to consonant" (Gay, 1981). This view may have its origins in an early categorization of sound patterns which defined coarticulation as "the transitions between a vowel and an adjacent consonant" (Chomsky and Halle, 1968).

Fowler (1980) says "coproduction of vowels and consonants" implies that two speech gestures are produced by different articulatory systems (and that "feature spreading," which would allow coproduction by a single articulatory system, is "implausible"). She summarizes, "What makes coproduction mechanically feasible to Ohman and Perkell is that the production of vowels and consonants involve [sic] essentially different ... muscles." Fowler quotes Perkell (1969) at length, ending, "It is probable that articulation of vowels is accomplished principally by the larger, slower extrinsic tongue musculature which controls tongue position. On the other hand, consonant articulation requires the addition of the precise, more complex, and faster function of the smaller intrinsic tongue musculature." Fowler's conclusion is that "the capacity for coproduction derives from an adaptive property of speech that the two classes of articulatory gestures, consonants and vowels, are products of different (coordinated) neuromuscular systems." While this line of argument is interesting, it tends to lead the reader to a view of coarticulation which ignores the effects of adjacent vowels upon each other, and therefore also ignores the inherent possibility of limitations to such coarticulation, either within or between languages. In effect, a whole category of possible enquiry into acoustic and articulatory phonetic phenomena has become tidied out of view.

However, if the slow transitional "portamento" which typically occurs between the two elements of English diphthongs is deemed to be an effect of decreasing/increasing production of the respective elements, then it is surely admissible, in spite of the above arguments, to view this as an instance of "coarticulation," in a slightly broader sense. It is also possible to view the faster (and subjectively inaudible) transition or "jump" typical of Japanese VV sequences as an example of "coarticulation resistance" (Bladon and Al-Bamerni, 1976). Fowler (1994), reviewing Bladon and Al-Bamerni as well as Recasens' work on Catalan consonants (a very different context to the present study), defines coarticulation resistance as "an act of self-preservation," an idea which can surely be applied, with modification, to the Japanese tendency to preserve the individuality of diphthong elements. Fowler also refers to "hyperarticulation," apparently as an exceptional mode; but it is surely possible for hyperarticulation to become the predominant mode for certain sounds in certain languages. These ideas may be relevant to, for instance, Japanese /a/ and /i/, which seem to be hyperarticulated when they are adjacent (i.e. the glide one might expect is absent).

Maddieson and Emmorey (1985) report that "cross-language differences...extend to the pattern of coarticulation between semivowels and the adjacent vowels," clearly using a rather broad definition of coarticulation. Manuel (1999) defines coarticulation even more broadly, as "patterns of coordination, between the articulatory gestures of neighboring segments which result in the vocal tract responding at any one time to commands for more than one segment," but goes on to admit that "according to this definition, it is almost impossible to produce an utterance that is more than one segment long without engaging in coarticulation." The implications of these broader definitions are that phenomena captured by the narrower definitions become special cases of more general intra- and inter-segment coordination. Magen (1997) allows a similarly broad view: "A central goal of research in speech production has been the description and prediction of coarticulatory effects, defined as the articulatory or acoustic influence of one segement or phoneme on another, and resulting in the absence of a one-to-one mapping between phonemes and their output in production."

One of the major significances of the pioneering work of Ohman (and others who followed) is that it has made clear that languages differ significantly in their distribution of such coarticulatory effects. As Manuel says, "...the patterns of overlap are affected by speakers' efforts to maintain distinctions among segments. Precisely because what counts as contrastive or distinctive varies both from language to language and for different segments within a language, we would expect to find differences in coarticulatory patterns in different languages, and within a language..." One might add that "what counts as contrastive or distinctive" does so for a wide variety of reasons, not just articulatory or perceptual, but also phonological, linguistic, or even psychological, social or historical. Thus the study of language-bound coarticulation patterns reaches potentially deep down into what-it-means-to-be-English, Japanese or whatever language it is that one is investigating.

"If speakers are to maintain contrasts, we might expect them to coarticulate segments in such a way as not to destroy the distinctive attributes of those segments" (Manuel, 1999). However, when it comes to diphthong-like vowel sequences in Japanese, Japanese speakers maintain contrasts which seem at first sight to be in excess of linguistic requirements.

3. Evidence for radical differences between Japanese and English

3.1. Subjective observations

Many Japanese VV sequences have superficially similar acoustic characteristics to English. In particular, the initial and final formant frequencies for some sequences are so similar between the languages that it is not always easy to tell whether they are Japanese or English on the basis of the formant frequencies alone. This is suggested by the preliminary results of the spectrographic analysis of Japanese and English vowels which have inspired the present study (Gore 1996, 1998). There does, on the other hand, appear to be a radical difference in the way that the vowels move between the beginning and end points, which seems to be based mainly on differences in timing, but perhaps also in some measure on stress patterns -- though timing is, of course, a major component of stress (Fry 1954).

Whereas English diphthongs tend to spend a long time in the central area of the oral cavity, corresponding Japanese sequences seem to skim over this area. Moreover, Japanese students of English generally tend to pronounce English diphthongs in the Japanese way, i.e. de-emphasizing the central area, even though they may have relatively accurate English peripheral vowel sounds. It is interesting that those students who have acquired the natural "slow-gliding" English diphthongs /oi/ and /ai/ seem to have no difficulty in pronouncing English schwa. Conversely, those who cannot pronounce schwa are unable to glide slowly on these English diphthongs. This finding can be confirmed spectrographically. There may be a correlation between these two abilities.

3.1. Spectrographic evidence

One simple way to measure and show the acoustic characteristics of vowels visually is on the vowel-chart mode of the Kay Elemetrics computerised speech laboratory (Sona-Match software). When Japanese and English vowel sounds are recorded on this chart, a stark difference can be seen in the way the central area of the oral cavity is treated in the respective languages. Put simply: English vowel sounds can be seen to glide about mostly in or near the central area of the oral cavity, whereas Japanese vowels can be seen to jump from one side to the other, avoiding the centre and seeming to have relatively little effect upon each other (Gore 1996). This difference is so radical that one suspects that any inherent difference is being exaggerated by the characteristics of the software program (Sona-Match), and indeed the distinction is not so readily apparent on sound spectrographs, but it can nevertheless be revealed on spectrographs, too, with careful analysis.

Let us consider a concrete example. For many reasons /ai/ is convenient. This sound is a common diphthong with overlapping distributions of formant values in English and Japanese. It is therefore natural that when Japanese students come to pronounce an English word such as "eye" they often unconsciously substitute the Japanese "ai." Thus detailed analysis of the small but significant difference between these sequences may throw light on several aspects of the respective sound systems, particularly on the vowel systems and Japanese (mora) timing as opposed to English (stress) timing.

Compare Figures 1 and 2. Whereas English speakers tend to "undershoot" the distance between the two diphthongal elements, drawing both elements (particularly the second) towards an indeterminate mid-point, thus creating on the F1-by-F2 graph a short continuous line, Japanese speakers regularly "overshoot" this gap and give equal prominence to both elements, resulting in two widely separated but tightly knit dot groupings with no dots in the central area.

Figure 1 is a Kay CSL recording of a 20 year old male Japanese saying "ai" ("love") four times in quick succession (total duration approximately 2.5 seconds) using Sona-Match software in the vowel-chart mode. Note that the elements are widely separated. Females tended to show even clearer separation.

Figure 2 shows the word "eye" as pronounced by a native English (Southern British) speaker four times in quick succession (total duration approximately 2.8 seconds). The visual difference between the Japanese and English articulatory patterns is extraordinary, though the beginning and ending formant values overlap.

As has been mentioned in the Introduction, the Japanese "vowel-jump" phenomenon shown in Figure 1 would imply a step-wise movement of the formants, but that would surely be a physical impossibility. What then is happening? Is the tongue flying through the central area so fast that it leaves no trace? Is it somehow skirting the central area? Is there a sudden significant drop in volume as the tongue moves through that area so that nothing at all is recorded? If so, what could cause such a reduction in volume -- a split-second change in the air velocity, or in the degree of voicing, or in both, or in some other factor? Is this perhaps an effect of anti-resonances (antiformants or zeros), which can occur between the first and second formants when the soft palate is lowered during the production of a vowel, and which can cause errors in LPC analysis (Bladon, 1979; Lieberman, Blumstein, 1988; Hayward, 2000)? A movement in the tongue root is another possibility which must be tested articulographically. But then, of course, why should

Figure 1

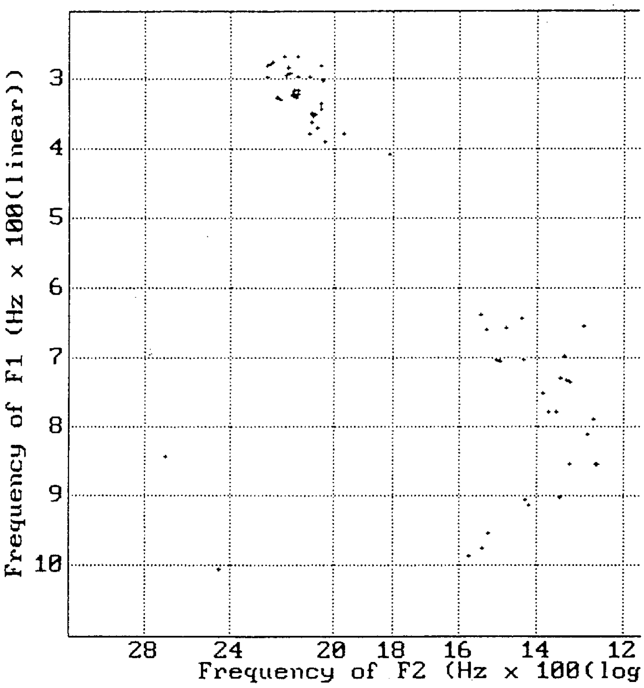


Figure 2

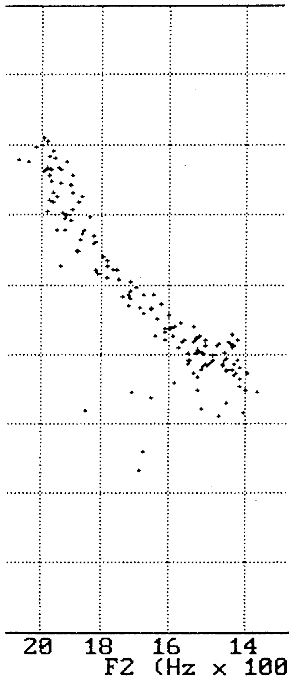
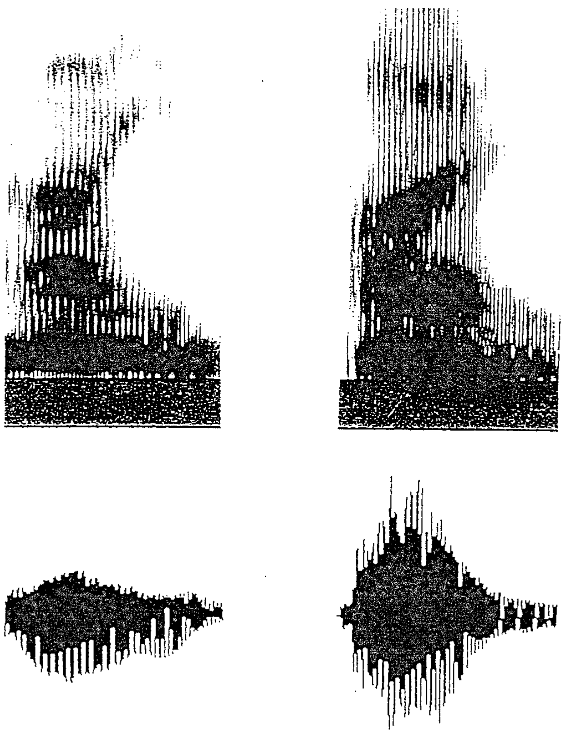


Figure 3 Kay spectrograms of /ai/, Japanese (left) and English (right)

Japanese /ai/ (left) shows sudden divergence of first and second formants, starting at the half-way point. English /ai/ (right) shows slow movement throughout the diphthong.



this "vowel-jump" need to happen? And why in Japanese? These are puzzling questions, since in many contexts no important linguistic distinction would appear to be lost by gliding through the central area, apart from the purity of the individual vowel sounds themselves. First let us look at what appears to be happening spectrographically.

Figure 3 shows Kay Real-time spectrograms (8kHz sampling) of /ai/, Japanese (left) and English (right). Japanese /ai/ (left) shows sudden divergence of first and second formants. English /ai/ (right) shows slow movement throughout the diphthong. A closer analysis of the respective spectrograms reveals that in the Japanese sequence the beginning of the glide tends to come later (about half-way through), and yet the end-point tends to come earlier than in English. No evidence can be found for a sudden reduction in volume in the fundamental frequency of either diphthong element, but a slight reduction in the sonorance or "resonance" (Fry, 1979) of formants 1 and 2 can be seen as they move rapidly from the first element to the second, followed by a recovery as F1 approaches the fundamental (F0) in the close vowel /i/. In English the transition starts earlier, the glide is slower and the volume tapers. Both intensity readings (bottom) show a single nucleus, though the Japanese contour is flatter.

4. The significance of transition time

Similar analyses of the spectrograms of various Japanese vowels show that, in Japanese, this sort of "jump" (or "fast glide") occurs most often in the long-distance transitions, /ai/, /ia/ and /oi/, /io/ (and also /oe/ and /eo/ although these are rare), whereas in the more "neighboring" transitions (/ei/, /ie/, /iu/, /ui/, /ao/, /oa/, /ae/, /ea/, /ou/, /uo/) proportionately slower gliding (like that usually found in English diphthongs) is often seen. As might be expected, /ua/ and /au/ seem to constitute an intermediate class. These observations point to a possible interesting difference in timing constraints between Japanese and English diphthongal VV sequences, which is discussed in Section 9, below.

5. Study 1 -- Spectrographic comparison of Japanese and English VV sequences.

Data on Japanese and English VV sequences was collected using a Kay Elemetrics CSL 4300B, Software Version 5.X, and a Sure SM48 microphone in a quiet, heavily carpeted and curtained room. Five Japanese males aged 18 to 24, and five Southern British males aged 24 to 60 were enrolled in the study. Spectrograms were made of each of the four VV sequences that are most similar in Japanese and English, namely /ai/, /au/, /oi/ and /ei/. Male voices were used for ease and reliability in distinguishing the formants. Each sound was repeated five times, formant values were measured on the Sona-Match vowel-chart, and averages calculated (see Formant frequencies, under Study 2, below).

6. Study 2 -- Synthesis of diphthongs

To elucidate the role of transition speed in Japanese and English VV sequences, speech synthesis software (Praat 4.0) was used to synthesize eight transitions of varying durations (with 20 millisecond increments) for each of the four diphthongal sequences common to both languages as described in Study 1, namely /ai/, /au/, /oi/ and /ei/, thus creating altogether 32 sounds from scratch. These were later tested on a panel of six volunteers who had no knowledge of the purpose or design of the experiment (three Japanese and three English speakers) singly and in random order, in a quiet, heavily carpeted room. Verbal reactions to each sound were encouraged, particularly with regard to whether the sequences sounded "English," "close to English," "Japanese" or "close to Japanese."

Intonation

A gently falling fundamental frequency was considered to be generally the most natural option (Atkinson, 1973). Accordingly, the Praat "Pitch Tier" for most of the synthesized sounds was set at 120Hz at the start (0.05 seconds), 100Hz at 0.2 seconds and 90Hz at 0.35 seconds. These values are in agreement with the fundamental frequencies listed by Holbrook and Fairbanks. "The fundamental frequency typically decreased about two and one-half tones and amplitude about 5db during the course of the diphthong" (Holbrook and Fairbanks, 1962). This fall also follows the intonation pattern of standard Japanese for actual words corresponding to the synthesized sounds as given in Kenkyusha's New Japanese-English Dictionary (1974). It is convenient that all the candidate sequences constitute complete words in both Japanese and English. (The use of units larger than words will be referred to later.) In the case of /oi/, however, the intonation needed to be flat or to have a very slight rise in Japanese in order to avoid confusion with the falling and more common Japanese word "ooi" (see Limitations, below). Such a confusion might create a problem with timing. Accordingly, /oi/ was given a flat intonation.

Formant frequencies

To create the first and second formants, the mean frequencies of the ten samples in Study 1 were used for the start and end points of each of the four sequences. (These mean frequencies are in every case within the limits of initial and final vowel values for each separate language group, and the resulting sounds were thought to be not particularly unnatural for either language as pure vowels.)

The mean frequencies (and ranges) in Herz for F1 and F2 which were used to create the first synthesized sequence to be tested, /ai/, are given below.

English (5 males) "Eye"

/a/ F1: 690 (640-710); F2: 1250 (1210-1320)

/i/ F1: 400 (350-450); F2: 2000 (1900-2120)

Japanese (5 males) "ai"

/a/ F1: 710 (650-790); F2: 1230 (1050-1380)

/i/ F1: 310 (260-360); F2: 2200 (2000-2380)

Mean frequencies of total English + Japanese (10 males)

/a/ F1: 700 F2: 1240

/i/ F1: 355 F2: 2100

The third formant was a "token" formant set at a constant 2500. This is because it was found in preliminary trials that quite wide variations in the third formant (from 2400 to 3100) had no effect at all on perception. On the other hand total absence of the third formant was immediately noticeable. This is in agreement with previous studies (Delattre et al 1952; Fant 1973; Lieberman and Blumstein 1988).

Bandwidths were the Praat default settings of 50Hz for F1, 100Hz for F2 and 150Hz for F3. Various other bandwidths including a fixed 100Hz width and widths of one-tenth of the respective formant frequency (Praat recommendation) were tried in preliminary experiments, but no perceptible difference was observed.

Total duration, and eight transition durations

The total duration of each sound was set at 300 milliseconds (from 0.05 to 0.35 seconds on the Praat graph, with 0.2 as the mid-point for convenience). Previous trials using 200ms and 400ms durations were felt to be slightly too short, especially for English, and slightly too long, especially for Japanese. These values are longer than the findings of Campbell (1992) for a Japanese professional radio announcer, but were not considered unnatural by Japanese participants in the present preliminary trials. All

American English diphthong lengths in Holbrook and Fairbanks (1962) were between 200 and 300ms (F0, F1 and F2), but both Fry (1979) and Hayward (2000) show examples of /ai/ which are approximately 400ms. A duration of 300ms was therefore felt to be a convenient compromise and not particularly unnatural for either language.

In every case, the first vowel was maintained up to the halfway point, 0.2, from where it was given a straightline glide to the second vowel which was set to begin at either 0.22, (i.e. very quickly afterwards), 0.24, 0.26, 0.28, 0.30, 0.32, 0.34 or 0.36 seconds. The decision to use eight 0.02 second increments followed much trial and error with many different scales, but is in accordance with the 20 millisecond psychoacoustic timing distinction or "difference limens" for the minimal difference that humans can perceive in time, which is thought to depend on the natural constraints of the auditory system (Hirsch 1959; Hirsch and Sherrick 1961): a difference of approximately 20 milliseconds has been shown to be necessary for auditory distinction in a variety of stimulus conditions.

The Praat "Create PitchTier" time values were set at 0 and 0.4, with the first and last 50 milliseconds devoiced using "Modify; To PointProcess; Remove points between 0 and 0.05; Remove points between 0.35 and 0.4;" etc. This has no perceptible effect on the start of the vowel sounds, but allows 0.05 seconds at the end for tapering (see Figure 4, bottom right), without which a sudden unnatural break-off may be heard.

Intensity

Praat default settings, including the recommended spectral slope of -6dB per octave, were used (Praat tutorial, Mar, 2001, Source-filter synthesis, 8 and 9), and preliminary trials produced acceptably natural sounding vowel sequences in both English and Japanese.

Synthesis and replay equipment

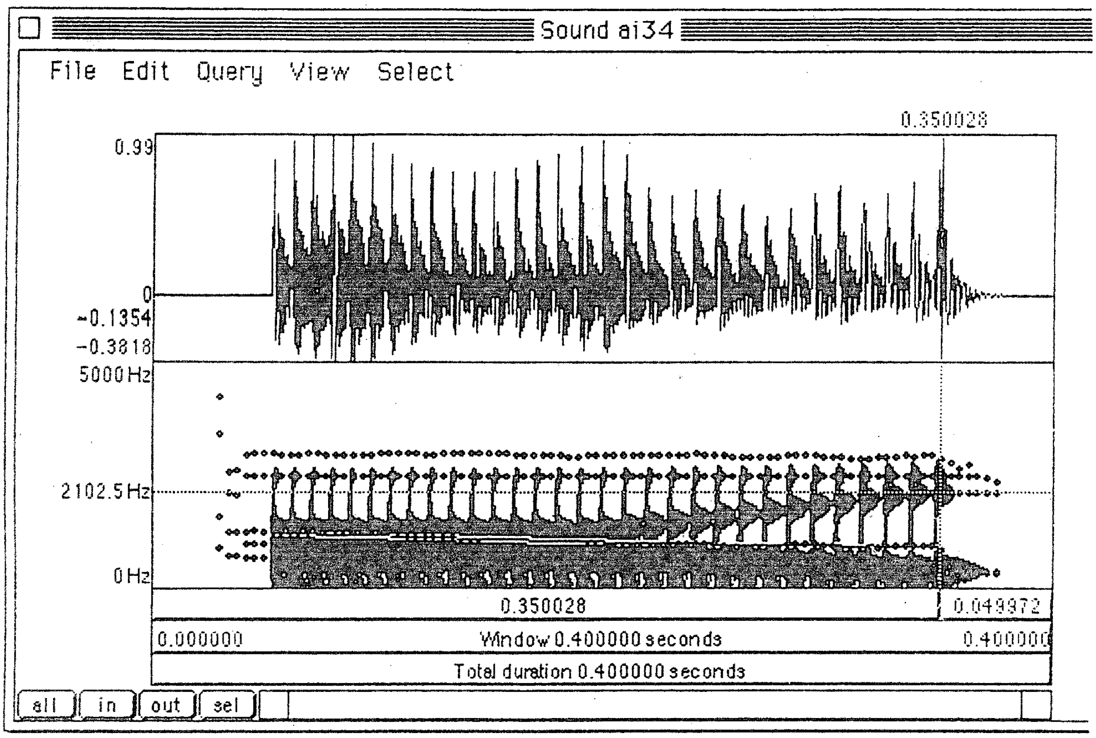
Macintosh Power PC, MacOS9.1, Praat 4.0 (www.praat.org) and a Sony SA-PC5 speaker.

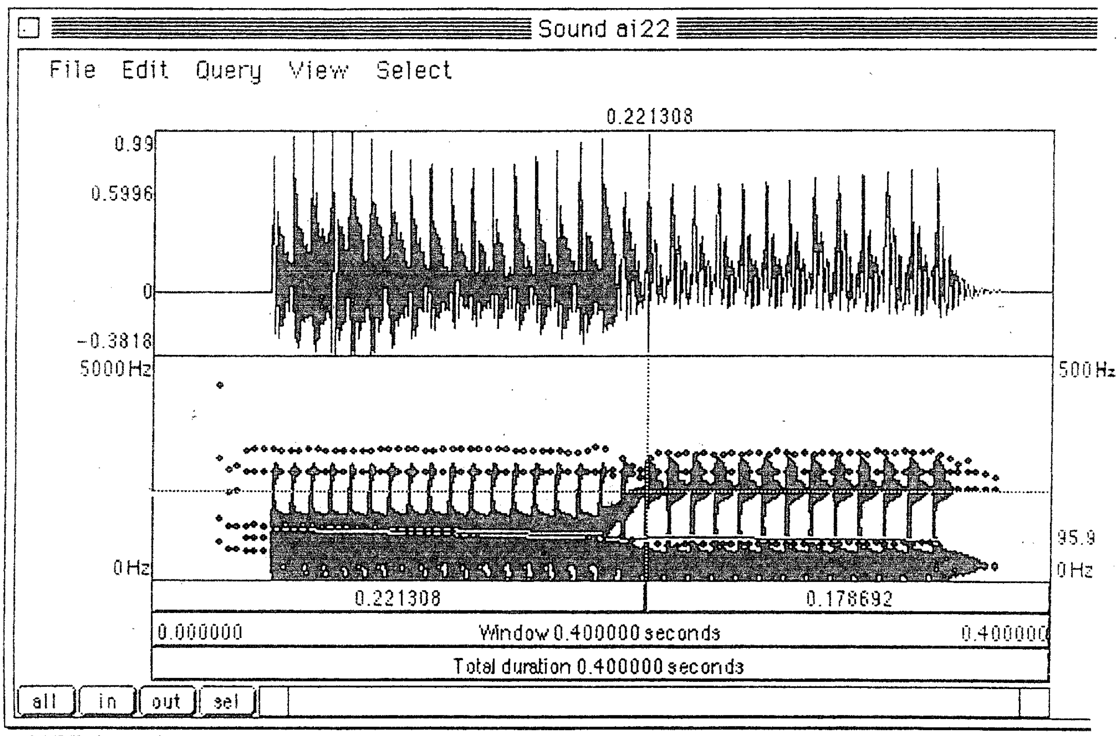
Figure 4 shows two sample Praat graphs of different transition durations for the synthesized /ai/.

Top: "Sound ai34" shows a transition duration of 0.14 seconds (0.2~0.34). The second formant reaches 2100Hz at 0.34 seconds (the cursor actually rests at 0.350028 seconds, top right).

Bottom: "Sound ai22" shows a transition duration of 0.02 seconds (0.2~0.22). The slowly descending white line shows the fundamental frequency finishing at 95.9Hz (bottom right). This right-hand scale is from 0Hz to 500Hz only, and applies to the fundamental frequency alone.

Figure 4 Two sample Praat graphs showing different transition durations for the synthesized diphthong /ai/.





7. Preliminary results

1. In Study 1, the Japanese transitions were always shorter and of more constant duration than the English transitions.

2. In Study 2, all the listeners judged the shorter synthesized transitions to be closer to Japanese, and the longer ones closer to English.

3. Two native English-speaking listeners made a distinction between 0.22, "not a natural vowel," and 0.24, "a vowel, but not English." Japanese listeners did not make this distinction, and made no comment.

4. English listeners tended to interpret the 0.22 transition as a "consonant," but were unable to identify it.

5. Both Japanese and English listeners considered 0.26 and 0.28 acceptable to both languages.

6. All listeners thought "the vowels themselves" had been changed, not the timing.

8. Discussion

As expected, shorter transitions were considered "Japanese," and longer transitions "English." Since the transition-duration increments were set in accordance with the approximate minimum perceptible time difference (0.02 seconds), it was expected that adjacent samples would not be easily distinguishable but that next-but-one samples would be readily discriminated. If, however, a clear distinction was observed between adjacent samples, that might be evidence of the sort of "category boundary" that is said to exist in other areas of sound perception such as in the voice onset time of consonants (Lisker and Abramson, 1971).

Indeed, in the present trials, a sharp distinction was reported by English speaking listeners between 0.22 ("not vowel") and 0.24 ("vowel, but not English"). Japanese listeners, however, did not perceive a distinction here. For most Japanese listeners, 0.22 (0.02 second transition) was an unremarkable diphthong sound, whereas most English listeners interpreted the abruptness of the vowel change as a consonant. That "transition tempo can distinguish stop consonant from semi-vowel from vowel of changing color," was reported by Liberman et al. (1956), but whether there is an absolute distinction between Japanese and English perception of the V/C category boundary in this borderline acoustic area is an intriguing question which must be investigated more scientifically with narrower increments and with larger numbers of listeners. In any case, it is fascinating that a minute change in the duration of the transition between the two diphthongal elements led to a perceived change in vowel formant values, and that the extent of this perceived change depended on the language background of the listener.

In Study 1, English durations tended to vary so that distant transitions were longer. In contrast, Japanese transitions appeared to have a constant duration regardless of the distance between the targets, and were always faster than English. It may be found with more exact analysis that there is a time-constraint on Japanese vowel transitions, whereby distant elements such as /ai/ require, in effect, a "jump," whereas those which are short-distance achieve the glide more slowly. This "constant-time transition constraint," if it can be verified experimentally, may throw an interesting light on mora-timing in Japanese (Beckman 1982; Hoequist 1983; Campbell 1992). In English, on the other hand, one would expect the long-distance transitions to take longer than the short-distance transitions, but this, too, requires further testing. In any case, the key to understanding the difference between English and Japanese "diphthongs" would seem, from the studies conducted so far, to be dependent primarily on timing rather than vowel quality.

More evidence is needed, but it is clear that when analysing the differences between Japanese and English vowel sequences, timing is important. Such timing is likely to be related to prosodic considerations. Some researchers have referred to the relationship between prosody and coarticulation (Hoequist 1983, Browman and Goldstein 1986, Manuel 1999), using the traditional rhythmic typology of stress/syllable/mora timing (Pike 1945). This can be summarized as follows:-

In stress-timed languages (e.g. English, German, Russian), stressed vowels are said to be shortened by following unstressed vowels and exert articulatory influence over them.

In syllable-timed languages (e.g. French, Italian, Spanish), unstressed vowels are fully articulated and the time between successive vowels tends to be fixed regardless of the number of intervening consonants.

In mora-timed languages (e.g. Estonian, Finnish, Japanese) the mora, or short syllable, (which in J may be a CV, a V, or a final nasal) is said to have a constant duration and each has approximately equal stress. Smith (1995) predicts that mora-timed languages will show "combined CV production strategies in which vowels are coordinated with preceding consonants, and not preceding vowels" (Manuel 1999). The present findings are in accord with this. Unfortunately, the boundaries to these three timing categories are notoriously difficult to define, and a detailed review of the theoretical background and its relevance to the present study will have to wait. For the moment, suffice it to say that, as Browman and Goldstein (1986) have suggested, prosody and patterns of articulatory organization seem to be closely dependent on each other.

However, what is curious about most prosody/coarticulation theorizing hitherto is that it has not focussed directly on the case of diphthongs, which show quite clear coarticulatory/prosodic contrasts between languages, particularly when one compares languages as diverse as English and Japanese. The treatment of diphthongs (along with long vowels) in traditional Japanese 5-7-5 metrical forms, such as haiku or tanka, may be illuminating in this context, because both long vowels and diphthongs have occasionally been counted as a single time unit or as two units, depending largely on the whim of the poet and the needs of meter. This may reflect an inate ambiguity of the Japanese timing system, or it may be due to the influence of Chinese poetry, or, of course, both.

Campbell (1992) reports that the Japanese "long phones" of a male radio announcer are approximately one-and-a-half times as long as the "short phones", with respective mean durations of 109ms and 66ms. His data on vowel lengths are thorough and illuminating, and relevant to the present study, but the database is of a single professional broadcaster and there is no specific comparison of short vowels with diphthongs as opposed to long vowels. The distribution of /ai/ (between verbs and nouns) is mentioned, but duration is not given, perhaps because it is considered not as a diphthong but as two short vowels. This is, of course, a legitimate view, but unfortunate from the point of view of acoustic analysis.

Though prosody and patterns of articulatory organization are clearly interdependent, theories concerning this have tended to ignore VV sequences in spite of the fact that VVs show interesting articulatory and/or prosodic contrasts when one compares languages such as English and Japanese. Four tentative accounts may be given for the difference in timing between Japanese and English diphthongal VVs: (1) Japanese requires adjacent vowel elements to have similar length; a slow transition reduces the effective length of the second element, creating an imbalance. (2) Japanese does not tolerate slow gliding; conversely English does not tolerate the very fast transitions that are tolerated by Japanese. (3) English vowels are unspecified for time (isochrony in feet only); Japanese morae are specified (isochrony in morae) so that time occupied by the glide must not encroach on the time allotted to the mora. (4) English diphthongs are more monophonemic than Japanese "diphthongs," at least as far as word-reversal, consonant insertion or "shiritori" games are concerned (see Berg 1986). However, as Figure 3 shows, Japanese diphthongs seem to have a single nucleus in terms of intensity. This controversial area is discussed in the following paragraphs.

The phonemic status of adjacent vowel morae in Japanese is a matter for debate, and whether there is a clearly defined group of "diphthongs" corresponding to the English sense of the word is a complicated issue. Phonological arguments for the existence of diphthongs in "heavy syllables" are provided by the effects on accent placement and minimum loanword size (three moras if including a diphthong, long vowel or final nasal, e.g. "daiya" not "dai") (Kubozono 2002, Haraguchi 1996). According to Kokugogaku Jiten (Dictionary of Japanese Linguistics), Japanese differentiates between vowels, diphthongs and long vowels, and the diphthongs are principally /ai/, /ei/, /oi/. One might add /au/ to this list (see below). The example given in DJL is /ei/ which can be two single vowels as in "e-iri" ("picture-included" -- a native Japanese) or a diphthong as in "ei-ri" ("profit-making" -- Chinese compound).

One interesting aspect of the so-called "diphthongs" in Japanese is that they can have pure vowel variants. The sequence /ai/ sometimes has variant forms which approach /e/ or /ee/; the sequence /ei/ nearly always becomes /ee/, (except in the first example, above); /oi/ sometimes approaches /e/ or /ee/ in some areas; and /au/ often becomes /o/ or /oo/ in old Japanese or in Kansai dialect, but this seems to be increasingly rare in standard Japanese. The persistence of such variations is contributory evidence that these sequences in these positions have a single nucleus and can be considered as single phonological units not unlike English diphthongs. With regard to the formants, we can see a sort of "default" position for reduction here, and it may be possible to say that whereas the English default is "schwa," the Japanese default is /e/, though the phonological constraints are very different. These sounds are also the pause indicators in the respective languages. Other Japanese vowel sequences do not have such variant forms, and this may be the real reason why they are not considered to be diphthongs.

When any particular /ai/ sequence in any particular word does not use the /ee/ or /e/ variant form, the two vowel elements keep their identity in excess of any overt linguistic requirements, so that there is rarely any perceptible gliding between them (at least in the speech of Japanese people educated in Japan) either through schwa or through /e/. Some gliding must occur, but this is usually too fast to be perceived as a glide, and the diminuendo conceals it. The result in these cases (the majority) is something which is often considered to be a diphthong, but which does not meet the normal English acoustic criteria for such a designation, mainly because there is no obvious gliding, at least in standard Japanese.

Given that there are these difficulties in defining diphthongs, particularly in Japanese, and that there are basic timing differences between English and Japanese, the present results can be interpreted as showing that VV (or CVV or VVC or CVVC) sequences in Japanese and English do exist which are words in themselves and which have only one nucleus and are normally thought of as "diphthongs" in both languages, where the Japanese/English timing difference has an unexpectedly strong influence on the articulatory patterns of the vowels within the sequence.

The difference between diphthongs and non-diphthong VVs in Japanese may also be defined in terms of "options." One factor that has not yet been discussed is that for "non-diphthongs" in Japanese there is an option of using a glottal stop or a slight glottal restriction, but this is not compulsory. On the other hand, with the sequences that are known as "diphthongs," a glottal stop is never an option. But still there is something causing the vowel to jump rather than glide as it would in English. Certain possibilities have already been mentioned (Section 3.1).

The present work has been done mainly on words on falling accents and has shown (Figs1, 3 and 4) a rather clear separation of the vowels in Japanese. This separation seems to be even clearer when the accent is Low-High. LHL and LHH are the two most common accentual patterns in Japanese words of three and more syllables, and having a high second syllable probably helps to maintain typically Japanese moraic distinctions, especially in VVV words such as "aoi" (blue). It may be because of this very fact that falling accents such as HLL seem to be avoided in most areas of Japan, but one must be careful not to place too much importance on falling/rising accents, because any conclusions would apply only to certain dialects. In contrast, the fast transitions which have been seen in the present investigation seem to be relatively constant throughout Japan with relatively minor variations, though this remains to be empically validated. However, the recent influence of non-Japanese accents does seem to be having an effect on transition speed. This influence is spreading slowly but steadily throughout Japan. What effect this has on the mora system and dipthong/non-diphthong VVs remains to be seen, but it is possible to predict a partial crumbling of the mora system and the growth of a diphthong/non-diphthong dichotomy more like that of English.

It would seem that few linguistic distinctions would be lost in Japanese if glides on diphthong-like VV sequences were more emphasized as in English. Those Japanese who have lived abroad (particularly in America) do seem to use more glides; but it is a distinctive characteristic of standard Japanese that perceptible gliding is consciously (or semi-consciously) avoided. However, it cannot be denied that younger Japanese are consciously trying to increase the length of the glide, but at the same time trying to keep the mora-timing. If this results in comprehension difficulties and the urge to glide is the stronger factor, there may occur a gradual erosion of the mora timing system, with more reliance being placed on stress and/or intonation instead of the traditional absolute distinction between "short" and "long."

In any case, "resistance to gliding" is largely a psychological and/or cultural phenomenon which may be in a state of flux. Many prominent people in Japanese radio and television, (younger announcers, disk-jockeys etc), probably under US and other Western influence, are using English-type glides in Japanese VV sequences. Under what conditions these glides are used must be the subject of future study.

9. Limitations of the pilot studies

All sound spectrographic investigations are limited by the equipment and software used, and by the experimental techniques. A particular problem in spectrographic analysis is the difficulty of evaluating the possible role of anti-formants which can cause errors in LPC analysis. It is therefore necessary to carry out such experiments under several different conditions, and at the same time to verify not just the acoustic properties but also the articulatory aspects, making use of articulographic equipment such as the Carstens articulograph to confirm physically the precise movements of the tongue and other relevant articulators. It is also imperative to consider the effect of speaking rate. To do this, both analysis and synthesis will need to be repeated using full sentences uttered at varying speeds, and all of the synthesized transitions will need to be embedded in full sentences in both Japanese and English.

With regard to listeners' reactions to synthesized sounds (Study 2), it is an unfortunate possibility that the relative likelihood of any particular sound in isolation being interpreted as English or Japanese may be affected by the relative frequency of that sound or word in the respective languages, or in the listener's everyday use of the language. In other words, the listener may perceive a sound as Japanese if it constitutes a word more commonly found or used in Japanese even though it sounds English. The above selection has been made to minimize that risk, but this consideration would be one reason why, for the sake of scientific objectivity, tests should also be done with whole sentences.

Full tests need to be done with native English and Japanese speakers who have extensive knowledge of both languages. However, this throws up another problem: there is a strong possibility that bi-linguals and/or advanced learners have developed their own strategies for dealing with diphthongs which are neither typically Japanese nor typically English.

Since the prime object of Study 2 was to test only one parameter, that of the speed of the transition between Vowel 1 and Vowel 2, all variations in other parameters have been excluded. Other diphthong parameters that might have been tested are formant frequencies, formant bandwidths, relative intensities and/or devoicing of the vowel(s) (Japanese /i/ and /u/ are often devoiced between voiceless consonants; devoicing may theoretically occur in other contexts, too, such as in diphthongs or triphthongs). Some reduction of resonance on the transition alone is another possibility. Though these other possibilities have not been emphasized in the present work, it will be necessary to evaluate their significance more precisely.

Extrapolating from the Kay sound spectrographic data (Study 1) and the initial reactions to the Praat synthesized sequences (Study 2), it is anticipated that the application of more detailed and rigorous experimental techniques will reveal a radical difference between the ways in which Japanese and English vowels move in diphthongs (especially in the long-distance diphthongs such as /ai/ and /oi/), and also between the ways in which Japanese and English diphthongs are perceived by the listener.

10. Conclusions

English diphthongs show a relatively slow glide, somewhat of the character of the portamento that trombone and string instrument players often use, i.e. a portamento that can to some extent be discerned and identified as an element in its own right, and indeed may characterize the whole diphthong. The portamento tends to start with the beginning of the first target vowel and continue to the end of the second, and is likely to lengthen according to the distance between the two elements. Holbrook and Fairbanks (1962) show that the overall duration of the American English diphthong does lengthen, particularly in /ai/, /au/ and /oi/, and Lehiste and Peterson (1961) report that in these three diphthongs "the transitions between targets are longer than either target position."

In similar Japanese diphthongal VV sequences (especially in long-distance transitions), the speed of the transition is often so fast as to be, subjectively, almost indiscernible. Spectrographic analysis shows it to be typically faster than in comparable English contexts, and on the Kay vowel chart (Sona-Match), groupings can be seen at each target area with few dots or none at all in the spaces between. This would seem to indicate a movement of extreme speed and/or sudden loss of intensity/voicing. At the moment, speed (transition duration) seems to be the most powerful differentiating factor. If this is so, it may be considered misleading to continue to refer to the English slow glide and the Japanese fast glide using the same ambiguous "glide" terminology. It may become expedient to class Japanese diphthong transitions under the name of "Fast Transitions," and rename the English diphthong glides as "Portamentos" along the following lines:

Fast Transitions (under 0.08 seconds), as found in Japanese diphthongs.

Portamentos (over 0.06 seconds, usually much longer), as found in English diphthongs.

If the above dichotomy of Fast Transitions versus Portamentos turns out to be an accurate prediction verifiable under various circumstances, one would expect Japanese Fast Transitions to be approximately time-constant regardless of speaking rate; on the other hand one would expect English Portamentos to vary in length in accordance with both speaking rate and total vowel duration.

This may lead to some further predictions. It would seem consistent with the evidence so far to hypothesize the existence of an "absolute time-constraint" on Japanese diphthongal transitions, whereby all Japanese diphthongs require the same "constant-time transition" between elements, so that "distant" vowels such as /ai/ require, in effect, a "jump," whereas those which are short-distance or "neighboring" achieve the glide in a more relaxed way within the time constraint and thus sound more similar to English "portamento" transitions.

As shown in Sudy 1 (Figure 1), Japanese speakers seem to give almost equal prominence to both elements of the VV sequence and to use an extremely fast transition, resulting in two widely separated but tightly knit dot groupings with no dots in the "schwa" area. The Kay CSL Sona-Match recording of a 20 year old male Japanese saying "ai" ("love") four times in quick succession showed the vowels as widely separated. Females tend to show even clearer separation. The visual difference between the Japanese and English sounds is extraordinary, and may lead to the inference that the Japanese speakers are "avoiding" the central area of the oral cavity because schwa does not exist in the Japanese vowel system.

Preliminary experiments have shown that Japanese students of English do in fact acquire a better schwa through practising long distance diphthongs /ai/ and /oi/ with a slow glide, and conversely that those who have mastered the schwa have no trouble achieving the English portamento transition (Gore 1996, 1998). This points to a correlation between schwa acquisition and slow transition timing. As Kondo (1995) argues, "production of schwa by Japanese speakers...implies the acquisition of a new coarticulatory strategy."

There is evidence from Japanese dialects that while VV never have a slow glide they may in some cases merge to form another vowel. This vowel is always another peripheral vowel, and there is never a hint of schwa. For instance /a.i/ does not become a gliding diphthong /ai/ but may tend towards /e/ or /ee/; similarly /a.u/ does not become a gliding diphthong /au/, but may become /o/ or /oo/. There are similar examples in European languages. What is interesting with the Japanese diphthongs, however, is that both forms exist side by side in many areas of Japan and yet gliding is not found to any subjectively perceptible extent. This is surely has some relation to the facts that (1) all dialects of Japanese are (predominantly) mora timed, and that (2) no dialects of Japanese allow vowel reduction to schwa. As Keating and Huffman (1984) note, "vowels that in other languages would be likely candidates for reduction...devoice, delete or shorten." However, as the present research has shown, this is unlikely to be the whole story.

References

Atkinson, J.R. (1973). Aspects of intonation in speech: Implications from an experimental study of fundamental frequency. PhD dissertation, University of Connecticut, Storrs. Quoted in Lieberman, P. & Blumstein, S.E. (1988), *Speech physiology, speech perception, and acoustic phonetics,* 200. Cambridge University Press.

Beckman, M. (1982). Segment duration and the 'mora' in Japanese. *Phonetica* 39: 113-135.

Berg, T. (1985). The Monophonematic Status of Diphthongs Revisited. *Phonetica* 42: 198-205.

Bladon, R.A.W. & Al-Bamerni, A.H. (1976). Coarticulation resistance in English /l/. *Journal of Phonetics* 4:137-150.

Browman, C.P. & Goldstein, L. (1986). *Towards an articulatory phonology.* Phonology Yearbook 3: 219-252.

Butcher, A., & Weiher, E. (1976). An electropalatographic investigation of coarticulation in VCV sequences. *Journal of Phonetics* 4, 59-74.

Campbell, N. (1992). Segmental elasticity and timing in Japanese speech. In *Speech Perception, Production and Linguistic Structure.* ATR Auditory and Visual Perception Research. Tokyo: Ohmsha.

Chomsky, N. & M. Halle. (1968). *The Sound Pattern of English.* New York: Harper and Row.

Delattre, P., Liberman, A.M., Cooper, F.S. & Gerstman, L.J. (1952). An experimental study of the acoustic determinants of vowel color: Observations on one- and two-formant vowels synthesized from spectrographic patterns. *Word* 8: 195-210.

Fant, G. (1973). Speech, sounds and features. Cambridge, MA: MIT Press

Farnetani, E. & Racasens, D. (1999). Coarticulation models in recent speech production theories.

In Hardcastle, W.J. & Hewlett, N. (eds.), *Coarticulation: theory, data and techniques.* Cambridge University Press. 31-65.

Farnetani, E., Vagges K. & Magno-Caldognetto, E. (1985). Coarticulation in Italian /VtV/ sequences: a palatographic study. *Phonetica* 42: 78-99.

Fowler, C. (1980). Coarticulation and theories of extrinsic timing. *Journal of Phonetics* 8: 113-133.

Fry, D.B. (1954). Duration and Intensity as Physical Correlates of Linguistic Stress. *Journal of the Acoustical Society of America* 27 (4). In Lehiste, I., ed. (1967), Readings in Acoustic Phonetics. Cambridge MA: MITPress.

Fry, D.B. (1979). *The physics of speech.* Cambridge: Cambridge University Press.

Gay, T. (1981). Temporal and spatial properties of articulatory movements: Evidence for minimum spreading across and maximum effects within syllable boundaries. In Myers, T., Laver, J. & Anderson, J. (eds.), *The cognitive representation of speech*, 133-141. Amsterdam: North-Holland. Quoted in Magen, H.S. (1997), The extent of vowel-to-vowel coarticulation in English. *Journal of Phonetics* 25: 187-205.

Gore, M. (1996). Formant Frequency Characteristics of English Central Vowels Acquired by Japanese University Students. *Collected Articles on the English Language*, 30. Ronsetsu Shiryo Hozonkai. Tokyo.

Gore, M. (1998). Phonetic Analysis of Japanese University Students' Pronunciation of English Vowels. *Bulletin of the Faculty of Education, Kagoshima University*, 49. 1998.

Han, M.S. (1962). *Japanese phonology: an analysis based upon sound spectrograms.* Tokyo: Kenkyusha.

Haraguchi, S.(1996). Syllable,mora and accent. In Otake, T.& Cutler, A.(eds.), Speech Research 12. *Phonological Structure and Language Processing.* Cross-Linguistic Studies. Mouton de Gruyter.

Hardcastle, W.J. & Hewlett, N., (eds.) (1999). *Coarticulation: theory, data and techniques.* Cambridge Studies in Speech Science and Communication. Cambridge: Cambridge University Press.

Hayward, K. (2000). *Experimental Phonetics.* Longman Linguistics Library.

Hirsch, I.J. (1959). Auditory perception of temporal order. *Journal of the Acoustical Society of America* 31: 759-67.

Hirsch, I.J. & Sherrick, C.E. Jr. (1961). Perceived order in different sense modalities. *Journal of Experimental Psychology* 62: 426-32.

Hoequist, C. (1983). Syllable Duration in Stress-, Syllable- and Mora-Timed Languages. *Phonetica* 40:203-237.

Holbrook, A. & Fairbanks, G. (1962). Diphthong Formants and their Movements, *Journal of Speech and Hearing Research* 5-1. In Lehiste, I. (1967), Readings in Acoustic Phonetics. Cambridge MA: MITPress.

Keating. P.A. & Huffman, M.K. (1984). Vowel Variation in Japanese. *Phonetica* 41: 191-207.

Kenkyusha's New Japanese-English Dictionary. Fourth Edition. (1974). Tokyo: Kenkyusha.

Kent, Raymond D. (1992). *The Acoustic Analysis of Speech.* San Diego: Singular Publishing Group.

Kokugogaku Jiten (Dictionary of Japanese Linguistics). (1955). Kokugogakkai. Tokyodo Shuppan.

Kondo, Y. (1995). *Production of schwa by Japanese speakers of English.* PhD dissertation, University of Edinburgh.

Kubozono, H. (2001). *Mora and Syllable.* Kobe University.

Lehiste, I. & Peterson, G.E. (1961). Transitions, Glides and Diphthongs. *Journal of the Acoustical Society of America* 33 (3): 268-277. In Lehiste, I. (1967), Readings in Acoustic Phonetics. Cambridge MA: MITPress.

Lehiste, I., (ed.) (1967). Readings in Acoustic Phonetics. Cambridge MA: MITPress.

Liberman, A.M., Delattre, P.C., Gerstman, L.J. & Cooper, F.S. (1956). *Journal of Experimental Psychology* 52 (2). In Lehiste, I. (1967), Readings in Acoustic Phonetics. Cambridge MA: MITPress.

Lieberman, P. & Blumstein, S.E. (1988). *Speech physiology, speech perception, and acoustic phonetics*. Cambridge: Cambridge University Press.

Lisker, L. & Abramson, A.S. (1971). A cross-language study of voicing in initial stops: acoustical measurements. *Word* 20: 384-421.

Maddieson, I. & Emmorey, K. (1985). Relationship between Semivowels and Vowels: Cross-Linguistic Investigations of Acoustic Difference and Coarticulation. *Phonetica* 42: 163-174.

Magen, H.S. (1984). Vowel-to-vowel coarticulation in English and Japanese. *Journal of the Acoustical Society of America* 75 (1): 541.

Magen, H.S. (1997). The extent of vowel-to-vowel coarticulation in English. *Journal of Phonetics* 25: 187-205.

Manuel, S. (1999). Cross-Language Studies: relating language-particular coarticulation patterns to other language-particular facts. In Hardcastle, W.J. & Hewlett, N. (eds.), *Coarticulation: theory, data and techniques*. Cambridge University Press. 179-199.

Ohman, S.E.G. (1966). Coarticulation in VCV utterances: spectrographic measurements. *Journal of the Acoustical Society of America* 39: 151-168.

Pike, K.L. (1945). *The Intonation of American English*. Ann Arbor, MI: University of Michigan Press.

Perkell, J. (1969). *Physiology of speech production: Results and implications of a quantitative cineradiographic study*. Cambridge, MA: MIT Press.

Recasens, D. (1984). Vowel-to-Vowel coarticulation in Catalan VCV sequences. *Journal of the Acoustical Society of America* 76: 1624-1635.

Racasens, D. (1987). An acoustic analysis of V-to-C and V-to-V coarticulatory effects in Catalan and Spanish VCV sequences. *Journal of Phonetics* 15: 299-312.

Sound spectrographic equipment used in Pilot Study 1: Kay Elemetrics CSL 4300B, Software Version 5.X, Sure SM48 microphone, Spectrogram software and Sono-Match software purchased from Kay Elemetrics Corp., 12 Maple Avenue, Pine Brook, NJ 07058 USA. **Speech synthesis equipment used in Pilot Study 2:** Macintosh Power PC, MacOS9.1, Praat 4.0 (www.praat.org) and a Sony SA-PC5 speaker.