Published in final edited form as:
J Phon. 2009 July ; 37(3): 297–320. doi:10.1016/j.wocn.2009.03.007.

Contextual Effects on the Perception of Duration

John Kingston[a], **Shigeto Kawahara**[b], **Della Chambless**[c], **Daniel Mash**[a], and **Eve Brenner-Alsop**[a]

[a] University of Massachusetts, Amherst

[b] Rutgers University, the State University of New Jersey

[c] Duke University

Abstract

In the experiments reported here, listeners categorized and discriminated speech and non-speech analogue stimuli in which the durations of a vowel and a following consonant or their analogues were varied orthogonally. The listeners' native languages differed in how these durations covary in speakers' productions of such sequences. Because auditorist and autonomous models of speech perception hypothesize that the auditory qualities evoked by both kinds of stimuli determine their initial perceptual evaluation, they both predict that listeners from all the languages will respond similarly to non-speech analogues as they do to speech in both tasks. Because neither direct realist nor interactive models hypothesize such a processing stage, they predict instead that in the way in which vowel and consonant duration covary in the listeners' native languages will determine how they categorize and discriminate the speech stimuli, and that all listeners will categorize and discriminate the non-speech differently from the speech stimuli. Listeners' categorization of the speech stimuli did differ as a function of how these durations covary in their native languages, but all listeners discriminated the speech stimuli in the same way, and they all categorized and discriminated the non-speech stimuli in the same way, too. These similarities could arise from listeners adding the durations of the vowel and consonant intervals (or their analogues) in these tasks with these stimuli; they do so when linguistic experience does not influence them to perceive these durations otherwise. These results support an autonomous rather than interactive model in which listeners either add or apply their linguistic experience at a post-perceptual stage of processing. They do not however support an auditorist over a direct realist model because they provide no evidence that the signal's acoustic properties are transformed during the hypothesized prior perceptual stage.

Keywords

perception; context effects; duration; short-long contrasts; speech versus non-speech; cross-linguistic comparisons; autonomy versus interaction; auditorism versus direct realism

Corresponding author: John Kingston, Linguistics Department, University of Massachusetts, 150 Hicks Way, 226 South College, Amherst, MA 01003-9274, jkingston@linguist.umass.edu, Phone: 1-413-545-6833, Fax 1-413-545-2792.

1 Introduction[1]

The auditorist and autonomous theory of speech perception advocated here distinguishes three events and associated *objects* during a trial in a typical speech perception experiment in which the listener identifies the sound presented (Kingston 2005). The listener first hears the *acoustic properties* of the stimulus, then perceives the *auditory qualities* corresponding to those acoustic properties, and finally gives a *categorical response* to those percepts. Direct realism does not distinguish perceiving from hearing because perception cannot transform the signal's acoustic properties if they are to inform the listener reliably about the articulatory gestures that produced them (Fowler 1986, 1990, 1991, 1992). An interactive model of speech perception, in which linguistic knowledge feeds back immediately[2] onto the perceptual evaluation of the signal's acoustic properties, does not distinguish categorizing from perceiving because this linguistic knowledge does not merely alter the listener's bias to give a particular response, but also the percept on which that response is based (McClelland & Elman 1986; Elman & McClelland 1988; McClelland, Mirman, & Holt 2006).

The effects of the acoustics of the target sound's neighbors, its *context*, on its categorization have been an important source of evidence in the debates between proponents of auditorism versus direct realism and of autonomy versus interaction. For duration, studies differ in what they reveal about the nature of these effects.

Kluender, Diehl & Wright (1988) found that English listeners judged silent intervals simulating stop closures as "long" or "voiceless" more often after a short than a long vowel, as well as after after a short square-wave non-speech vowel analogue. The responses to the speech stimuli could reflect those listeners' experience of the inverse covariation in duration between a preceding vowel and consonants that contrast for [voice] in English (Lisker 1957, 1986; Parker, Diehl, & Kluender 1986), but those to non-speech analogues cannot. Kluender, et al. proposed that a *contrastive* auditory transformation of the signal caused the silent interval to sound longer after both a short vowel and a short vowel analogue.[3] This parsimonious account of listeners' responses to the two kinds of stimuli has been challenged by results of two subsequent studies.

First, Fowler (1992) found that listeners responded "long" more often when the preceding square-wave vowel analogue was long, rather than short. Fowler obtained the same bias using speech stimuli, except when listeners made a voicing rather than a length judgment about the consonant. Because the contrastive effect depended on the stimuli being speech rather than non-speech and on the listeners judging voicing not length, Fowler argued that the effect of the vowel's duration of the judgment of the consonant's voicing did not arise from auditory durational contrast but instead from listeners' perceivng that the duration of the vowel gesture covaries inversely with the following consonant gesture in sequences in which the consonant contrasts for voicing. Listeners would not perceive vowel and consonant gestures in the non-speech stimuli, and their past experience with listening to speech would not create any expectation that the durations of the vowel and consonant would covary inversely when they make length rather than voicing judgments.

---

[1]The results reported in this paper were first presented in a poster at the Providence meeting of the Acoustical Society of America (Kawahara 2006a). The research reported here was supported by NIH grant R01-DC006241 to the first author, which is gratefully acknowledged.

[2]We are indebted to Holger Mitterer for pointing out the need to draw a distinction between immediate feedback from sources of linguistic knowledge to the initial perceptual evaluation of the incoming signal and delayed feedback to later categorization processes. A model may permit delayed feedback and still be autonomous in the sense intended here (see Norris, McQueen, & Culter, 2003; McQueen, Cutler, & Norris, 2006; McQueen, Norris, & Cutler, 2006).

[3]Because the literature reviewed here only considers an auditory transformation that causes the percept of the consonant's duration to *contrast* with the vowel's duration, for the present we do not consider one that would instead cause the perceived consonant duration to *assimilate* to the vowel's duration. We take up that alternative later.

Second, van Dommelen (1999) reported that Norwegian listeners also treated an interval as long more often next to a long neighboring interval in non-speech analogues, but as short more often in corresponding speech stimuli. Norwegian distinguishes short and long consonants, and the durations of preceding vowels covary inversely. Because vowels' durations differ more than the consonants' (Fintoft 1961; also §2 for details), van Dommelen varied the vowel's duration incrementally rather than the silent interval's, which was either short or long. Listeners categorized the speech stimuli as words with a short versus long consonant rather than directly judging the vowel's or consonant's length. They responded with the word having a long vowel more often when the following silent interval was short rather than long. However, when the stimuli were instead non-speech analogues, listeners judged the vowel analogue as "long" more often when the following silent interval was also long, much as Fowler's listeners had done.

In summary, one study, Kluender, et al. (1988), showed that listeners judge an interval as "long" more often next to a short neighboring interval in categorizing non-speech analogues as well as speech stimuli, and they did so when making length as well as voicing judgments. However, the other two, Fowler (1992) and van Dommelen (1999), only obtained such contrastive judgments when listeners categorized speech stimuli in terms of a distinction in their native language in which the two intervals' durations covary inversely,[4] and otherwise found that listeners categorized the target interval as "long" more often next to a long interval. Although some listeners responded differently in all these studies, the weight of the evidence apparently supports direct realism over auditorism and interaction over autonomy. This evidence does not support separating a linguistically naïve perceptual stage of processing during which the signal's acoustic properties are transformed auditorily from a later post-perceptual stage when they apply their linguistic knowledge.

One telling difficulty remains: why should listeners in Fowler's (1992) and van Dommelen's (1999) studies have judged the target interval to be long more often next to an adjacent long interval in the non-speech analogues (both studies) or when making a length judgment in response to the speech stimuli (Fowler 1992)? Neither direct realism nor interaction predicts that the durations of these two intervals should interact perceptually in any way in these circumstances. Fowler does not address this question, but, following a suggestion from Diehl, van Dommelen proposes that his listeners may instead have added the two intervals' durations together in the non-speech analogues. Fowler's listeners may have done so, too, in the circumstances where they responded "long" more often next to a long context.

We also measured the categorization of non-speech analogues as well as speech stimuli to determine whether they are both subject to contrast or addition. Our experiments examined the influence of linguistic experience on duration judgments by comparing categorization responses from listeners who speak languages that differ in how the durations of successive vowels and consonants covary, Japanese, Norwegian, Italian, and English. In separate experiments, listeners discriminated pairs of stimuli in which the durations of successive vowels and consonants (or their non-speech analogues) covaried directly or inversely, i.e. short-short versus long-long, or short-long versus long-short. Addition would exaggerate the difference between the directly covarying stimuli and make them more discriminable than the inversely covarying ones, while contrast would have the opposite effect.

Japanese, Norwegian, Italian, and English all use the duration of a consonantal constriction to convey a phonological contrast, and in all four, the duration of the preceding vowel covaries with the consonant's duration. Japanese, Norwegian, and Italian use constriction duration to

[4]Other studies present results that could be interpreted as evidence for durational contrast, notably Diehl & Walsh (1989, cf. Miller & Liberman 1979), but this evidence, too, remains controversial (Fowler 1990, 1991; Diehl, Walsh, & Kluender 1991). There is no space to discuss these data and their interpretation here (but see Brenner-Alsop, 2006).

convey a contrast between short and long consonants,[5] while in English, constriction duration is instead a correlate of the voicing contrast.[6] The languages also differ in the direction in which the preceding vowel's duration covaries with the consonant's; see the next section.

2 Crosslinguistic differences in the covariation of vowel and consonant duration

In Table 1, we present ratios obtained from measurements reported in the literature of consonant and preceding vowel durations in Japanese, Norwegian, Italian, and English. Ratios make cross-linguistic comparisons possible.

2.1 Japanese

In Japanese, a vowel (first column) is only about two-thirds as long before a short than a long consonant (Fukui 1978, Han 1994, Kawahara 2005, 2006b). Long consonants are more than twice as long as short ones (second column), and the consonant-vowel duration ratios are considerably larger when the consonant is long than when it's short (fourth versus third columns). Of the four languages, only Japanese contrasts vowels for quantity as well as consonants, although a contrastively long vowel cannot precede a contrastively long consonant inside a single morpheme (Kubozono 1999). Previous studies of how the preceding vowel's duration affects the perception of the short:long contrast disagree. On the one hand, Watanabe & Hirato (1985) and Hirata (1990) report that a shorter preceding vowel biased listeners toward "long" responses in two studies— an effect contrary to expectations grounded in the production data. However, only two listeners (apparently the authors) participated in Watanabe & Hirato's study, and the effects of preceding vowel duration were very small in Hirata's data. On the other hand, Arakawa & Kawagoe (1998) and (Ofuka, et al. 2005) found that a longer preceding vowel induces more "long" responses — an effect in accord with the production data.

2.2 Norwegian

Norwegian also contrasts short with long consonants (Fintoft 1961, Kristofferson 2000). The relative differences in constriction duration (second column) are, however, small compared to those in the preceding vowel (first column). The consonant:vowel duration ratios for short and long consonants (third and fourth columns) also overlap more than they did in Japanese. These facts indicate once again that the contrast is conveyed more by differences in the vowel's than the consonant's duration. As discussed above, van Dommelen (1999) shows that a shorter constriction duration causes Norwegian listeners to judge the preceding vowel as "long" in speech stimuli.

2.3 Italian

Vowels are consistently longer before short than long consonants in Italian, but, unlike Norwegian, the consonant duration differences are large and the vowel duration differences comparatively small (second versus first columns). As a result, the consonant:vowel duration ratios for long versus short consonants do not overlap at all (third and fourth columns). Esposito & Di Benedetto (1999) showed that the boundary between the short versus long consonant categories shifted from 165.8 ms after a vowel lasting 116 ms to 182.7 ms after one lasting 176 ms, a shift of 16.9 ms for a 60 ms change in preceding vowel duration.

[5]Short consonants are also referred to as "singletons" and long consonants as "geminates" in descriptions of these three languages.
[6]Constriction durations also differ between obstruents contrasting for voicing in Japanese (Kawahara 2005, 2006b), Norwegian (Behne & Moxness 1995), and Italian (Esposito & Di Benedetto 1999), in the same direction as they do in English.

2.4 English

Unlike the other languages, English does not contrast consonants for length. Instead, constriction duration differs between English obstruents that contrast for voicing after vowels, and the preceding vowel's duration covaries inversely (second and first columns; see also Lisker 1957, 1986; Chen 1970; Port & Dalby 1982; Kluender, et al. 1988). The consonant:duration ratios for voiceless (long) versus voiced (short) consonants also do not overlap (fourth versus third columns). Listeners categorize a stop more often as "voiceless" or "long" when the preceding vowel is short (Raphael 1981; Kluender, et al. 1988, cf. Fowler, 1992).

2.5 Summary

In Norwegian, Italian, and English, preceding vowel duration covaries inversely with consonant duration, while in Japanese, it instead covaries directly. In Japanese, Norwegian, and Italian, the consonant durations reflect a contrast for length, while in English they instead reflect the voicing contrast.[7] Japanese listeners respond "long" more often when the preceding vowel is longer, while English and Italian listeners do so when the preceding vowel is shorter, and Norwegian listeners categorize the preceding vowel as "long" more often when the following consonant is short.

2.6 Predictions of competing theories

Before presenting the experiments, we review the different predictions of the competing theories.

First, all theories predict that listeners' categorization of the speech stimuli should depend on the direction in which preceding vowel duration covaries with consonant duration, because the speech stimuli are necessarily categorized at a stage when linguistic knowledge about the direction of covariation can bias the listener's response. The theories' predictions only differ for discrimination of the speech stimuli and categorization and discrimination of the non-speech analogues. If the autonomous model is correct and a linguistically naïve, perceptual stage of processing precedes linguistically informed, post-perceptual categorization, then listeners can discriminate the speech stimuli using either the auditory qualities perceived during that stage or the categories the stimuli are assigned to during the later post-perceptual stage. Because the interactive model does not distinguish between an initial linguistically naïve, perceptual stage from a later informed, post-perceptual stage, listeners could only discriminate the speech stimuli using the categories to which they have been assigned. In this respect, the interactive model resembles classical categorical perception (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967), while the autonomous model resembles the dual-process model (Fujisaki & Kawashima, 1969, 1970; see also Pisoni, 1973, 1975).

Second, if the auditorist model is correct, then the acoustic properties of the signal can be transformed during the perceptual processing stage; for example, a short preceding vowel could contrastively lengthen the perceived duration of the silent interval. Direct realism requires that the silent interval's duration instead be perceived veridically. This requirement does not mean that the vowel's duration cannot alter the percept of the consonant's duration, but that it can only do so to the extent that the gestures which produce these two intervals are perceived to covary in a particular direction.

These two comparisons show that autonomy and auditorism are more complex theories than interaction and direct realism, because autonomy posits two stages in processing while

[7]The independent inverse covariation between vowel and consonant duration induced by the voicing contrast in Japanese, Norwegian, and Italian should play no role in our experiments because the intervals used to simulate stop closures were entirely silent.

interaction posits only one, and auditorism posits auditory transformations of the signal's acoustic properties that direct realism prohibits.

Third, because autonomy but not interaction includes a processing stage that is sensitive to the identical acoustic manipulations in the non-speech analogues as the speech stimuli but that is indifferent to the stimuli's nature, it predicts both kinds of stimuli could be categorized in the same way. Interaction does not predict similar categorization because only categorization of the speech stimuli would be influenced by linguistic experience. Because the acoustics are manipulated identically in the two kinds of stimuli and they could therefore undergo the same auditory transformation, auditorism also predicts that the non-speech analogues would be categorized like the speech stimuli. Direct realism does not because only the speech stimuli can be perceived as the products of covarying articulatory gestures.

Fourth, both autonomy and auditorism predict that listeners with different linguistic experience will discriminate the speech stimuli similarly, and they will discriminate the non-speech analogues in the same way as they do the speech stimuli, because linguistic experience can only shift the location of the criterion for dividing the consonant interval in "short" versus "long" categories but cannot also alter the continuous values of the auditory qualities put out by perceptual mechanisms. Interaction and direct realism instead predict that discrimination of the speech stimuli will depend on how vowel and consonant duration covary in the listeners' native language, and that discrimination of the non-speech analogues will not resemble discrimination of the speech stimuli, because listeners would rely on their linguistic experience of covarying vowel and consonant gestures in discriminating the speech stimuli, and that experience would not apply to discriminating the non-speech stimuli.

In short, both auditorism and direct realism predict that listeners from different linguistic backgrounds will categorize the speech stimuli differently, as do autonomy and interaction, but only auditorism and autonomy predict that they all will discriminate them similarly. Auditorism and autonomy also differ from direct realism and interaction in predicting that the non-speech stimuli will be categorized and discriminated in the same way as the speech stimuli.

In the next four sections, we present methods, results, and discussion first for the speech and non-speech categorization experiments, and then for the corresponding discrimination experiments.

3 Categorization: Method

3.1 Stimuli

3.1.1 Speech—All speech stimuli were made from tokens of the Japanese minimal pair *hato* 'dove' versus *hat:o* 'hat'. Both words were recorded in the frame sentence *korewa* __ *desu* 'this is X' by a native speaker of Tokyo Japanese (the second author). The token of the stretch *hatto desu* with the shortest preceding vowel served as the base for all continua. This token's vowel (henceforth V1) is closest in duration to a vowel that would precede a short [t]. Its duration, 63 ms, is roughly midway between the short duration of about 40 ms observed in Japanese before short consonants and the long duration of about 80 ms observed before long consonants. From this base, we extracted four intervals: [h], [a], [od], which included the preceding [t]-burst, and [esu̥], where the [u̥] is devoiced as is usual in Japanese following an utterance-final voiceless obstruent.[8]

[8]The onset of voicing marked the end of the [h] and the beginning of the [a]; the offset of voicing marked the end of the [a] and the beginning of the [t]; the stop burst marked the end of the stop closure; and the onset of voicing following the [d]'s stop burst marked the boundary between the [od] and [esu̥] intervals.

The durations of the preceding [h] and the following [od] were held constant at 78 ms and 124 ms, respectively. To create long and short preceding vowels, we lengthened the [a] interval *in-situ* by a factor of 1.4 and shortened it by a factor of 0.7, using the PSOLA function in Praat (Boersma & Weenink 2005),[9] producing a long V1 of 89 ms, and a short V1 of 49 ms. The three [a] intervals were then extracted, and their edges were ramped up and down with a raised 5 ms cosine window to ensure the interval began and ended with zero intensity. They were then inserted after the initial [h].

The members of a 7-step continuum of silent intervals ranging from 60 to 150 ms in 15 ms increments were inserted between each of the three [a] intervals and the following [od] interval, to simulate a continuum from a short to long stop, following short, medium, and long vowels. To equalize the total durations of all the stimuli and thereby prevent listeners from responding to differences in total stimulus duration, the duration of the [esu̥] interval was then adjusted in each stimulus via PSOLA.[10] The difference between the shortest and longest combination of [a] and silent interval was 130 ms (49 ms V1 + 60 ms silence versus 89 + 150 ms), and the shortest and longest [esu̥] intervals lasted 420 and 550 ms, respectively (94 ms [e] + 326 ms [su̥] versus 123 + 427 ms). Table 2 lists the durations of the constituent intervals that formed each stimulus. After the intervals were concatenated, all stimuli were scaled to 0.92 of the maximum intensity.

Figures 1a,b present spectrograms of speech stimuli with a short (Figure 1a) and long silent interval (Figure 1b).

3.1.2 Non-Speech—The non-speech analogues were constructed using filtered square waves for the consonant intervals and anharmonic complexes of sine waves for the vowel intervals. The square wave's f0 was 100 Hz, and it consisted of the first 50 odd harmonics of this frequency. The ratios of their amplitudes to f0's intensity were 1/*harmonic number* (i.e. 1/1, 1/3, 1/5, etc.). The anharmonic complexes were composed of 50 sine waves ranging in frequency from 100–16000 Hz and separated by equal natural log intervals (0.101503). Their amplitude ratios were 1/(2**component number*^{2} + 1) (i.e. 1/3, 1/9, 1/19, etc.). The analogues of the consonant intervals ([h], [d] and [s]) were attenuated relative to the vowel intervals (see below for values). Both kinds of intervals resembled the corresponding speech intervals in their overall spectral characteristics: energy was concentrated at higher frequencies in the fricative analogues but it was broadly distributed from low to high frequencies in the vowel analogues; the [t] analogues were intervals of silence; and the [d] interval had periodic energy only at very low frequencies (100 Hz). The characteristics of each portion of the non-speech stimuli are summarized in Table 3.

The vowel analogues were all windowed by an on-and-off raised cosine ramp lasting 5 ms. The durations of all intervals were identical to those listed in Table 2. Figures 1c,d show spectrograms of the two non-speech analogues whose durations match those of the speech stimuli above them in Figures 1a,b. None of the listeners reported that the non-speech analogues sounded like speech.

[9]P(itch) S(ynchronous) O(ver)L(ap and) A(dd) is a method for changing the duration of an interval while preserving its spectral properties. It first divides the signal into separate but overlapping short windows lasting from 2–4 periods centered around each period. The sequence of windows is then modified by either repeating or leaving out windows, which modifies the duration of the signal, and finally the remaining windows are recombined by overlapping and adding.

[10]Neither Kluender, et al. (1988), Fowler (1992), nor van Dommelen (1999) adjusted the other intervals in their stimuli to prevent their listeners from responding in terms of total stimulus duration.

3.2 Procedures

3.2.1 Location and equipment—The English participants were run in a sound-attenuated chamber in the Phonetics Laboratory at the University of Massachusetts, Amherst. Each participant sat at a desktop PC input-output terminal consisting of a monitor, response box, and headphones connected to a computer outside the chamber. All stimuli were output at 16 kHz from the PCs and presented binaurally to listeners through sound-isolating Sennheiser HD 280 64 Ohm headphones, at a comfortable listening volume. Cedrus SuperLab Pro software (version 2.0.4) was used to present all sound stimuli and visual cues, and to log responses. Participants responded using Cedrus RB-834 response boxes.

The Norwegian and Japanese participants were run in quiet but not sound-attenuated rooms. They listened to the stimuli over Beyerdynamic DT 250 80 Ohm headphones at comfortable listening levels, and responded using Cedrus RB-610 button boxes. Some of the Italian participants were run in the same facility and with the same equipment as the English participants, and the others were run off-campus in quiet rooms. Those run on campus used the same headphones and button boxes as the English participants, while those run elsewhere used the same headphones and button boxes as the Norwegian and Japanese participants. For participants run off-campus, a laptop PC was used to present stimuli and collect responses. Other details match those given for English above.

3.2.2 Trial and experiment structure—On each trial, a single stimulus was presented followed by the appearance of two color-coded visual response prompts on the screen. Their colors and positions matched those on the button boxes that participants used to respond. For the English participants. the response prompts were “Short” and “Long”, for Norwegian and Italian participants they were “t” and “tt”, and for Japanese participants, they were the *katakana* orthographic representations of the words “hato” and “hatto”. The response prompts remained onscreen until the participant responded or for 1500 ms, whichever came first. After the participant responded or the 1500 ms response period had elapsed, an additional 750 ms inter-trial interval (ITI) elapsed before the next stimulus was presented.

All listeners started the experiment with a training block consisting of 24 trials. These trials alternated between the short and long consonant closure endpoints in each of the three vowel contexts. After the 24 alternating trials, listeners worked through 6 more short training blocks, in which each endpoint was presented once in the three vowel contexts, in random order, for an additional 36 trials. Both sets of training trials differed from the ensuing test trials in that listeners received feedback in the form of the correct answer after they responded or after the 1500 ms response-period elapsed. The feedback was displayed on the screen for 500 ms. It was followed by the same 750-ms ITI, and then the presentation of the next stimulus. Training familiarized participants with the experimental procedures and the tempo of stimulus presentation and response collection, as taught them what counted as exemplary tokens of the short and long categories. Participants’ responses to the training stimuli were not included in the subsequent analyses.

Following training, listeners proceeded through 10 test blocks, in which they categorized stimuli drawn from the entire consonant duration continuum and with all three vowel durations. Every stimulus was presented three times per block, in random order, for a total of 30 presentations altogether.

After completing the training blocks and after each test block, participants saw a message on the screen inviting them to take a short break, and to press a button when they were ready to continue on with the next block. Following completion of the sixth test block — approximately halfway through the experiment — listeners took a longer break lasting roughly 5 minutes. The entire experimental session lasted under an hour, including both written and verbal

instructions, testing time, breaks, debriefing, and time for filling out consent forms and receipts for compensation.

3.2.3 Instructions—Before starting the experiment, listeners from languages other than Japanese were told they would hear a variety of syllable sequences and would have to decide whether the second consonant sounded like a short or long "t" (or the equivalent). The participants were also told the details of the trial structure during the training and test phases. Japanese listeners were instructed to choose between the two words, *hato* and *hatto*, but also went through the training blocks to familiarize themselves with the procedure and stimuli.

All the participants were instructed to respond as quickly as possible and were told to rest their forefingers or thumbs on the two response buttons, so they could respond without moving their arms or hands. This instruction, together with the relatively short time provided to respond (1500 ms) were intended to get listeners to respond with their first impression of the stimuli. This impression was more likely to be influenced by the acoustic details of the stimuli than one based on reflection.

3.2.4 Conditions—Half the participants from each language used the left button for "short" responses and the right button for "long" responses, and vice versa for the remaining participants. The results of statistical tests on this variable are listed in Appendix 1.

3.2.5 Non-Speech—The same procedure was used for categorizing the non-speech analogues as the speech stimuli, except all listeners used the arbitrary labels "A" and "B" for the left and right buttons, and learned these two categories during training and they were told they would be hearing a variety of sound "sequences", rather than syllables, which they were to categorize as "A" or "B". Half the listeners were taught during training that the short endpoint corresponded to the left button (the "A" response) while the other half were taught that this endpoint corresponded to the right button (the "B" response). For the sake of comparison with the speech results, we will refer to the number of "long" responses in discussing the results obtained in the non-speech as well as the speech experiments.

3.3 Participants

3.3.1 General characteristics—No listener participated in more than one experiment. No listeners reported any speech or hearing disorder, except for one English listener in the speech experiment, who despite a mild high-frequency hearing loss, responded more accurately in training than any other listener. Listeners were either paid for their participation or granted course credit.

3.3.2 Japanese—20 Japanese listeners each participated in the speech and non-speech categorization experiments. They were recruited from International Christian University, the Tokyo University of Agriculture and Technology, Chuo University, and other, non-academic venues. Although the listeners spoke different Japanese dialects, all contrast short and long consonants and use the same orthographic convention to represent them, so this dialectal diversity should have no consequences for our experiments.

3.3.3 Norwegian—10 Norwegian listeners each took part in the speech and non-speech experiments. They were recruited from the University of Tromsø community.[11] The listeners spoke different dialects, but short consonants contrast with long ones in all dialects, and the same orthographic convention is used to represent the contrast. Some participants also spoke Saami, a Finno-Ugric language.

[11]We are very grateful to Curt Rice and Carina Reinholtsen for their hospitality and assistance in conducting the experiments in Tromsø.

3.3.4 Italian—Ten participants in the speech experiment and five participants in the non-speech experiment were adult native Italian speakers recruited from the towns of Amherst and Northampton. The remaining six for the non-speech experiment were students from the Università degli Studi di Milano-Bicocca or the Scuola Normale Superiore di Pisa in Italy.[12] Data from two additional listeners in the non-speech experiment were omitted from the analysis because they apparently reversed the response buttons during the test phase.

3.3.5 English—All participants were adult native English speakers recruited from the University of Massachusetts at Amherst community. They spoke a variety of American dialects. 17 listeners were run in the speech condition, and 16 in the non-speech condition.

4 Categorization results

4.1 Statistics

Two repeated measures ANOVAs were run, one for the responses to speech stimuli, the other for the responses to the non-speech stimuli. The dependent variable was the total proportion of "long" responses across the continuum of closure durations, V1 duration was a within-subjects independent variable (short versus medium versus long), and language (Japanese versus Norwegian versus Italian versus English) and button (left "long" versus right "long") were between-subjects independent variables (the effect of button and its interactions with the other variables were negligible and are reported in Appendix 1). After we report the results of this cross-linguistic analysis, we report the results of paired-sample *t*-tests run within each language to compare the "long" response proportions between short versus medium and medium versus long V1 durations. The alpha value was reduced to 0.025 to correct for multiple tests. Results from the speech experiments are reported first, then those from the non-speech experiments.

4.2 Speech

4.2.1 Cross-linguistic analysis—Figure 2 displays the mean percentage of "long" responses for each step along the consonant duration continuum and each preceding vowel duration separately for each language. The Japanese and English listeners responded "long" more often when the preceding vowel was longer (Figures 2a,d), while the Norwegian and Italian listeners responded "long" more often when it was shorter (Figures 2b,c). The total proportion of "long" responses across the continuum and the three vowel durations differed significantly between the languages ($F(3,49) = 21.607$, $p < .001$) and all pairwise comparisons between languages were also significantly different except those between Japanese and Italian and between Norwegian and English (Japanese versus Norwegian: $t(28) = 3.346$, $p = .01$; Japanese versus English: $t(35) = 7$, $p < .001$; Norwegian versus Italian: $t(18) = 3.633$, $p = .004$; Italian versus English: $t(25) = 6.519$, $p < .001$; Japanese versus Italian: $t(28) = 0.846$, $p > .10$; Norwegian versus English: $t(25) = 2.482$, $p = .092$ – all tests are Bonferroni-corrected for multiple comparisons). V1 duration was also a significant main effect ($F(2,98) = 4.716$, $p = .011$), but of far more interest, it interacted significantly with language ($F(6,98) = 20.677$, $p < .001$). The interaction was significant because the effect of vowel duration on the frequency of "long" judgments for Japanese and English listeners was opposite that for Norwegian and Italian listeners.

4.2.2 V1 duration effects in each language—Japanese listeners (Figure 2a) responded "long" significantly more often after long V1 (total proportion of "long" responses across the

[12]We are very grateful to Maria Teresa Guasti in Milano and Pier Marco Bertinetto in Pisa and their assistants, Flavia Adani in Milano and Maddalena Agonigi, Chiara Bertini, and Irene Ricci in Pisa, for letting us use their facilities and for recruiting subjects for our experiments.

continuum: 0.689 ± 0.025 95% confidence interval) than medium V1 (0.617 ± 0.024, $t(19) = 8.427$, $p < .001$) and after medium V1 than short V1 (0.550 ± 0.029, $t(19) = 7.713$, $p < .001$). Norwegian listeners (Figure 2b) responded "long" significantly more often after the medium V1 (0.599 ± 0.088) than the long V1 (0.356, ± 0.116, $t(9) = 2.733$, $p = .023$), but there was only a non-significant trend toward more "long" responses after the short V1 (0.642 ± 0.118) than the medium V1 ($t(9) = 1.582$, $p > .10$). Italian listeners (Figure 2c) also responded "long" significantly more often after medium V1 (0.697 ± 0.043) than long V1 (0.549 ± 0.047, $t(9) = 7.143$, $p < .001$), but they actually responded "long" slightly less often after short V1 (0.677 ± 0.037) than medium V1 ($t(9) = -1.362$, $p > .10$). Finally, English listeners (Figure 2d) responded "long" more often after long V1 (0.559 ± 0.055) than medium V1 (0.451 ± 0.46, $t(16) = 4.532$, $p < .001$) and after medium V1 than short V1 (0.387 ± 0.037, $t(16) = 4.775$, $p < .001$).

4.2.3 Summary—Japanese and English listeners responded "long" to the speech stimuli significantly more often when V1 was longer, but Norwegian and Italian listeners did so when V1 was shorter. These results agree entirely with previous studies of the perception of the short:long contrast by Italian and Norwegian listeners (Esposito & Di Benedetto 1999, van Dommelen 1999), and with some but not all of the previous studies of Japanese (Arakawa & Kawagoe 1998, Ofuka et al 2005; cf. Watanabe & Hirato 1985, Hirata 1990). This difference can be readily attributed to differences in linguistic experience: the vowel preceding a long consonant in Japanese is longer than that preceding a short one, while the opposite is true in Norwegian and Italian. Differences in linguistic experience are also probably responsible for Norwegian listeners being more sensitive to vowel duration differences but less sensitive to consonant duration differences than the Italian listeners (recall Table 1). Linguistic experience cannot explain the effect of V1's duration on the English listeners' responses. Quite the contrary, the durations of these two intervals covarying *inversely* in these listeners' experience, in sequences where the consonants are obstruents contrasting for voicing. Our English listeners' bias toward more "long" responses after longer vowels is also similar to that obtained by Fowler (1992) from English listeners who categorized speech stimuli as containing a "short" or "long" consonant (cf. Kluender, et al. 1988).

4.3 Non-speech

4.3.1 Cross-linguistic analysis—Figure 3 shows that listeners from all four languages responded "long" more often as the V1 analogue's duration increased. In the analysis of listeners' responses to the non-speech stimuli, language was a significant main effect ($F(3,49) = 3.104$, $p = .035$), but none of the pairwise comparisons between languages reached significance (Japanese versus Norwegian: $t(28) = 1.962$, $p > .10$; Japanese versus Italian: $t(28) = 2.577$, $p = .069$; Japanese versus English: $t(34) = 2.261$, $p > .10$; Norwegian versus Italian: $t(18) = 0.533$, $p > .10$; Norwegian versus English: $t(24) = 0.037$, $p >$,10; Italian versus English: $t(24) = 0.556$, $p > .10$). V1 analogue duration was significant as a main effect ($F(2,98) = 65.639$, $p < .001$) but for these stimuli did not interact significantly with language ($F(6,98) = 1.364$, $p > .10$).

4.3.2 V1 analogue duration effects in each language—Japanese listeners (Figure 3a) responded "long" significantly more often after the long V1 analogue (0.655 ± 0.028) than the medium V1 analogue (0.557 ± 0.024, $t(19) = 14.734$, $p < .001$) and after the medium than the short V1 analogue (0.473 ± 0.022, $t(19) = 7.110$, $p < .001$). Norwegian listeners (Figure 3b) responded "long" significantly more often after the long V1 analogue (0.600 ± 0.053) than the medium V1 analogue (0.492 ± 0.079, $t(9) = 3.843$, $p = .004$) and after the medium than the short V1 analogue (0.440 ± 0.094, $t(9) = 3.042$, $p = .014$). Italian listeners (Figure 3c) responded "long" significantly more often after the long V1 analogue (0.581 ± 0.075) than the medium V1 analogue (0.473 ± 0.032, $t(10) = 2.982$, $p < .014$), but the proportion of "long" responses

after the medium V1 analogue did not differ significantly from that after the short V1 analogue (0.431 ± 0.058, $t(8) = 1.674$, $p > .10$). English listeners likewise responded "long" significantly more often after the long V1 analogue (0.561 ± 0.048) than the medium V1 analogue (0.506 ± 0.029, $t(15) = 2.671$, $p = .017$) and almost significantly more often after the medium than short V1 analogue (0.462 ± 0.041, $t(15) = 2.432$, $p < .028$).

4.3.3 Summary—Unlike their responses to the speech stimuli, listeners from all four languages responded "long" more often to the non-speech stimuli when the preceding V1 analogue was longer. These non-speech results resemble those obtained by Fowler (1992) from English listeners when categorizing non-speech analogues.

4.4 Comparing responses to speech versus non-speech

Responses to the non-speech stimuli did not differ as a function of the listener's linguistic experience, while those to the speech stimuli did, although only the Norwegian and Italian listeners' responses to the speech stimuli can be confidently attributed to their linguistic experience. Some mechanism other than applying their linguistic experience determined the responses of the Norwegian and Italian listeners' to the non-speech stimuli. The Japanese listeners' responses to both kinds of stimuli are ambiguous: they could reflect these listeners' linguistic experience of the direct covariation between consonant and preceding vowel duration in Japanese or the same mechanism as determined the responses of the Norwegian and Italian listeners' responses to the non-speech stimuli. Only that other mechanism could have determined the English listeners' responses to the speech stimuli, because they have no linguistic experience that would produce the observed responses.

What then could that other mechanism be? One possibility is that the perceived duration of the consonant interval *assimilated* to the duration of the preceding vowel, another is that listeners *added* the duration of the vowel to that of the consonant in judging how long the consonant was. Assimilation is a possible auditory transformation of an acoustic property that takes place during the perceptual stage prior to categorization. Addition must instead be a post-perceptual process because it operates on the continuous values put out by the perceptual stage. As a post-perceptual process, addition would apply at the same time as rather than prior to the application of linguistic knowledge, which can pre-empt it.

4.5 Evaluating the predictions of the competing theories

How do these results contribute to resolving the debates between auditorism versus direct realism or autonomy versus interaction? As noted above in §2.6, both sides of both debates predict that listeners who differ in their experience of the covariation of consonant and preceding vowel duration would, as observed, categorize the speech stimuli differently, because that experience can shift can shift the listener's criterion for categorizing the consonant as "short" versus "long" in different directions. Auditorism and autonomy but not direct realism and interaction also predict that the duration of the preceding vowel analogue would bias listeners' categorization of the silent interval in the non-speech stimuli in the same way as it did in the speech stimuli. This second prediction is disconfirmed by the Norwegian or Italian listeners' responses, which differed between the two kinds of stimuli, but not by the English and perhaps the Japanese listeners' responses, which did not differ. Auditorism and autonomy both predict that the responses to the non-speech stimuli will resemble those to speech stimuli because they both include a stage in processing during which the signal's acoustic properties may be transformed into auditory qualities, which can then be basis for categorizing the stimuli. If the signal is actually transformed during such a stage, that transformation is assimilative, as all listeners responded "long" more often to the non-speech stimuli after a longer preceding vowel analogue. That transformation must, nonetheless, be reversible when listeners categorize speech stimuli, if, like Norwegian and Italian listeners, they have experienced the consonant's

duration covarying inversely with the preceding vowel's. The possibility of reversal strikes us as implausible and theoretically extravagant and encourages us to consider the alternative, in which listeners add the preceding vowel's duration to the consonant's unless addition is pre-empted by the application of contrary linguistic knowledge. The listeners' categorization of these stimuli thus does not confirm the positive prediction of either auditorism or autonomy that the signal's acoustic properties are transformed during an initial, linguistically naïve stage in processing. Because neither direct realism nor interaction predicts that any such transformations occur, these results are apparently more compatible with these theories.

We now turn to the test of the competing predictions of these theories of speech perception provided by the discrimination experiments.

5 Discrimination: Method

5.1 Introduction

In one type of discrimination trial, the durations of the vowel and consonant intervals (or their analogues) covaried directly (vowel and consonant both short versus both long), while in the other type, these durations covaried inversely (short vowel-long consonant versus long vowel-short consonant). In both kinds of stimuli, the duration of the [esu̥] interval was again adjusted to compensate for the differences in combined vowel plus consonant durations. If the consonant's perceived duration contrasts with the preceding vowel's duration, then the inversely covarying stimuli should be more discriminable, because preceding short and long vowels would exaggerate the perceived difference between the following long and short consonants, respectively (Figure 4a).[13] However, if the consonant's perceived duration assimilates to the preceding vowel's duration, then the perceived difference between short and long consonant intervals is instead exaggerated when the preceding vowels are themselves also short and long, respectively, as in a directly covarying pair (Figure 4b). Addition would have the same effect on discriminability as assimilation, as the combined durations of the vowel and consonant are more different in the directly covarying pair (Figure 4b) than the inversely covarying one (Figure 4a).

Running the discrimination task with both the speech stimuli and their non-speech analogues tells us whether differential linguistic experience of the covariation between these two intervals influences listeners' speech responses in the same way as it did in the categorization task. Specifically, do Norwegian and Italian listeners discriminate inversely covarying speech stimuli better than directly covarying ones, while Japanese and English listeners discriminate directly covarying stimuli better, and do listeners from all languages discriminate directly covarying non-speech stimuli better?

5.2 Stimuli

The Japanese listeners' responses in the speech categorization task with the medium duration preceding vowel were used to select three closure durations for the discrimination task: a short (S) closure of 75 ms, which was categorized as "long" on less than 25% of trials with the medium V1, a long (L) closure of 105 ms, which was categorized as "long" on more than 75% of trials with the medium V1, and a medium (M) closure of 90 ms midway between these values. The three vowel durations used in the categorization experiments, 49 (S), 63 (M), and 89 (L) ms, were combined with all three closure durations to produce SS, MM, LL, SM, SL, ML, MS, LS, and LM stimuli, which were then presented in directly or inversely covarying pairs for discrimination (Table 4).

[13]If contrast or assimilation is a product of an auditory transformation, either distortion will occur regardless of the durations of the two intervals, and regardless of whether the durations of the two intervals are positively or negatively correlated.

5.3 Procedures

5.3.1 Locations and equipment—The discrimination data were collected in the same locations, using the same equipment and software as the categorization data.

5.3.2 Trial and experiment structure—Two stimuli were presented on each trial, separated by a 500 ms inter-stimulus interval (ISI). When the second stimulus finished playing, two visual cues – corresponding in orientation and color to two buttons on a response box – appeared onscreen to prompt listeners to respond. The visual prompts were the words “Same” and “Different” for English and Norwegian listeners, and the equivalent words or phrases in Japanese or Italian for listeners who speak these languages. The visual prompts remained on the screen until the listener responded or 1500 ms had elapsed, whichever happened first. After the listeners responded or the response period elapsed, the visual prompts disappeared and feedback in the form of the correct answer — “same” or “different” or their language-specific equivalents — appeared on the center of the screen for 750 ms. Feedback was given during training to help the listeners learn how to do the task, and was continued during testing to ensure that they continued to respond as accurately as possible. Finally, a 750 ms ITI followed the feedback.

Listeners started the experiment with two training blocks, each consisting of three randomly ordered presentations of every stimulus pair in which the vowel and consonant were short or long. A total of 48 training trials was presented: 2 different trials, one for inversely covarying short-long versus long-short and the other for directly covarying short-short versus long-long, plus the 2 corresponding same trials, multiplied by 2 orders by 3 repetitions by 2 blocks. Performance in the training blocks was not included in the analysis of the results.

After finishing training, listeners proceeded through 12 test blocks, which each consisted of two presentations of each different pair listed in Table 4 and an equal number of the corresponding same trials, for a total of 24 responses to each different and same pair. The entire set of trials within a given test block was randomized within each test block. All other details — total experimental duration, breaks, etc. — matched those in the categorization experiments described above.

5.3.3 Instructions—Before starting the experiment, participants were told they would hear a variety of pairs of sound sequences, and would have to decide whether the two members of that pair were exactly the same or different. They were told all the other details of the trial structure. Participants were instructed to respond as quickly as possible, so that their responses were based on their first impressions rather than subsequent reflection. One group of participants pressed the left button for “same” responses, the other the right button (statistical tests of the effects of button are in Appendix 1).

5.3.4 Procedure for non-speech stimuli—The non-speech procedure matched the procedure for speech in every detail, except for the nature of the stimuli presented, and the description of the stimuli during briefing as “sound sequences” rather than syllable sequences. Separate groups of listeners discriminated the non-speech stimuli.

5.3.5 Participants—20 Japanese speakers were recruited from the University of Massachusetts, Amherst community; 20 Norwegian participants from the University of Tromsø community; and 20 Italian participants in Milano and Pisa, with 10 listeners in each group discriminating the speech stimuli and the other 10 the non-speech stimuli. 18 and 15 English speakers were recruited for the speech condition and non-speech conditions, respectively, from the University of Massachusetts, Amherst community.

5.4 Scoring the responses

The discrimination measure, *d′*, was calculated using the differencing rule, because our experiment used a roving same-different task (Macmillan & Creelman 2005). (Corresponding proportions correct are reported in Appendix 2.)

5.5 Statistical analysis

Separate repeated-measures ANOVAs were run on the *d′* values calculated from listeners' responses to the speech versus non-speech stimuli. In these cross-linguistic analyses, covariation (direct versus inverse) and V1 duration difference (short versus medium, medium versus long, and short versus long) were within-subjects variables, and language and the button used for the "same" response (left versus right) were between-subjects variable. Subsequently, *d′* values obtained in response to inversely and directly covarying stimulus pairs were compared for short versus medium, medium versus long, and short versus long V1 duration differences within each language using paired-sample *t*-tests, with the alpha value set at 0.0167 to correct for repeated tests.

6 Discrimination: Results

6.1 Speech

6.1.1 Cross-linguistic analysis—Figure 5 displays listeners' success at discriminating the speech stimuli separately for each language. For all but one comparison, in all four languages, the directly covarying stimulus pair was more discriminable than the corresponding inversely covarying pair – the exception is the short-medium V1 pair for Italian listeners (Figure 5c). In the analysis of these data, the main effects of covariation, V1 duration difference, and language were all significant ($F(1,40) = 28.899$, $p < .001$; $F(2,80) = 123.404$, $p < .001$; $F(3,40) = 8.353 < .001$, respectively). Directly covarying pairs were more discriminable than inversely covarying ones ($d' = 2.640 \pm 0.308$ versus 2.085 ± 0.269), and *d′* values were significantly smaller for short versus medium V1 differences (1.381 ± 0.234) than medium versus long V1 differences (2.330 ± 0.342, $t(47) = 6.455$, $p < .001$), for medium versus long than short versus long V1 differences (3.376 ± 0.334, $t(47) = 10.896$, $p < .001$), and for short versus medium than short versus long V1 differences ($t(47) = 15$, $p < .001$). Performance by listeners from the four languages broke down into two groups: Japanese (3.030 ± 0.649) and Norwegian listeners (2.942 ± 0.519) discriminated the stimuli better than Italian (1.834 ± 0.566) or English listeners (1.644 ± 0.389). Japanese and Norwegian listeners did not differ significantly from one another ($t(18) = .214$, $p > .10$), nor did English and Italian listeners ($t(26) = .559$, $p > .10$), nor do, but both Japanese and Norwegian differed significantly from both Italian and English (Japanese versus Italian: $t(18) = 2.805$, $p = .046$; Japanese versus English: $t(26) = 3.706$, Norwegian versus Italian: $t(18) = 2.913$, $p = .035$ $p = .004$; Norwegian versus English: $t(26) = 4.044$, $p = .001$ – all tests are Bonferroni-corrected for multiple comparisons.)

More importantly, covariation interacted significantly with language ($F(3,40) = 3.520$, $p = .023$). Figure 5 shows that this interaction was significant in large part because Italian listeners did not discriminate the directly covarying pairs significantly better than inversely covarying ones, unlike listeners from the other three languages. V1 duration difference also interacted significantly with language ($F(6,80) = 7.153$, $p < .001$) because listeners with different native languages differed in their sensitivity to particular V1 duration differences – Figure 5a shows that *d′* values did not differ between short versus medium and and medium versus long V1 duration differences for Japanese listeners while the other panels in the figure show that they did differ between all three V1 duration differences for Norwegian, Italian, and English listeners. Finally, covariation interacted significantly with V1 duration difference ($F(2,80) = 12.926$, $p < .001$) because the advantage for directly covarying pairs over inversely covarying

ones is much greater for the short versus medium V1 difference than the medium versus long or short versus long V1 differences.

6.1.2 Interactions of covariation and V1 duration differences for individual languages—Japanese listeners (Figure 5a) discriminated directly covarying pairs better than inversely covarying ones when V1 was short in one of the intervals: short versus medium: direct d' = 3.271 ± 0.505 versus inverse d' = 1.639 ± 0.318 ($t(9) = 5.424$, $p < .001$) and short versus long: 5.0455 ± 0.424 versus 4.084 ± 0.559 ($t(9) = 5.947$, $p < .001$), but the two kinds of pairs did not differ significantly for the medium versus long V1 duration difference: 2.824 ± 0.945 versus 2.803 ± 0.376 ($t(9) = .057$, $p > .10$). Norwegian listeners (Figure 5b) discriminated directly covarying pairs more easily than inversely covarying ones only for the short versus medium V1 duration difference: 1.907 ± 0.388 versus 0.620 ± 0.843 ($t(9) = 3.458$, $p = .007$); cf. the medium versus long V1 duration difference: 3.407 ± 0.464 versus 3.258 ± 0.600 ($t(9) = 1.024$, $p > .10$) and the short versus long V1 duration difference: 4.289 ± 0.590 versus 4.169 ± 0.615 ($t(9) = .427$, $p > .10$). None of the pairwise comparisons of directly versus inversely covarying performance were even marginally significant for Italian listeners (Figure 5c): short versus medium V1 duration difference: 1.185 ± 0.633 versus 0.890 ± 0.444, ($t(9) = 0.851$, $p > .10$); medium versus long V1 duration difference: 1.680 ± 0.581 versus 1.760 ± 0.753, ($t(9) = -.322$, $p > .10$); and short versus long V1 duration difference: 2,180 ± 0.823 versus 2.523 ± 0.895 ($t(9) = -1.003$, $p > .10$). English listeners (Figure 5d) discriminated the directly covarying stimuli more easily than the inversely covarying ones for stimuli in which one of the V1 durations was short: short versus medium V1 duration differences: 1.363 ± 0.438 versus 0.304 ± 0.401 ($t(17) = 3.775$, $p = .002$) and short versus long V1 duration differences: 2.636 ± 0.541 versus 2.231 ± 0.436 ($t(17) = 2.952$, $p = .009$), but not for the medium versus long V1 duration difference: 1.802 ± 0.585 versus 1.428 ± 0.572 ($t(17) = 1.618$, $p > .10$).

6.1.3 Summary—The directly covarying stimuli were uniformly more discriminable than the inversely covarying ones, except for Italian listeners, who discriminated both kinds of speech stimuli equally well. That neither Norwegian nor Italian listeners discriminated the inversely covarying pairs better than the directly covarying ones shows that performance in this task is unaffected by their linguistic experience. Instead, the consonant's duration either assimilated perceptually to the preceding vowel's or the duration of that vowel was added to the consonant's duration.

6.2 Non-speech

6.2.1 Cross-linguistic analysis—Figure 6 shows that the directly covarying non-speech analogues were consistently more discriminable than the inversely covarying ones (2.419 ± 0.267 versus 1.248 ± 0.225), and unsurprisingly, covariation was a significant main effect ($F(1,37) = 161.521$, $p < .001$). V1 analogue duration difference was also a significant main effect ($F(2,74) = 75.868$, $p < .001$) because listeners discriminated the short versus long V1 analogue duration difference better (2.639 ± 0.311) than either the short versus medium (1.280 ± 0.213) or medium versus long V1 analogue duration difference (1.580 ± 0.265).

On the one hand, unlike the speech stimuli, language was not a significant main effect ($F(3,37) = 1.879$, $p > .10$), but, on the other hand, like the speech stimuli, covariation interacted significantly with language ($F(3,37) = 4.241$, $p = .011$). Comparison of Figure 6a with the other three panels in the figure shows that this interaction was significant because the size of the difference in performance between directly and inversely covarying pairs was greater for Japanese than the other three languages.

Covariation also interacted significantly with the V1 analogue duration difference ($F(2,74) = 6.058$, $p = .004$) for the non-speech stimuli, because the difference in discriminability of

inversely versus directly covarying pairs was larger for medium versus long and short versus long than for short versus medium V1 analogue duration differences.

6.2.2 Interactions of covariation and V1 duration differences for individual languages—For Japanese listeners (Figure 6a), directly covarying pairs were only significantly more discriminable than inversely covarying ones when one of the V1 analogues was long: medium versus long V1 analogue duration difference: 2.8055 ± 0.513 versus 0.760 ± 0.735 ($t(9) = 5.737$, $p < .001$) and short versus long V1 analogue duration difference: 4.352 ± 0.741 versus 2.264 ± 0.343 ($t(9) = 6.634$, $p < .001$). For the short versus medium V1 analogue difference, the directly covarying advantage was only marginal: 2.117 ± 0.482 versus 1.165 ± 0.723 ($t(9) = 2.537$, $p = .032$ (marginal). For Norwegian listeners (Figure 6b), the directly covarying pairs' discriminability advantage was only marginally significant for the short versus medium V1 analogue duration difference: 1.412 ± 0.460 versus 0.728 ± 0.752 ($t(9) = 2.744$, $p = .023$) and was otherwise not significant (medium versus long V1 analogue difference: 1.817 ± 1.035 versus 0.936 ± 0.814, $t(9) = 1.958$, $p = .082$, and short versus long V1 analogue difference: 2.541 ± 1.254 versus 1.624 ± 0.943, $t(9) = 2.085$, $p = .067$). Italian listeners (Figure 6c) discriminated the directly covarying pairs significantly better than the inversely covarying ones for the medium versus long V1 duration difference: 2.413 ± 0.424 versus 0.835 ± 0.497 ($t(9) = 4.635$, $p = .001$) and for the short versus long V1 analogue difference: 3.517 ± 0.541 versus 2.287 ± 0.567 ($t(9) = 7.841$, $p < .001$) but not for the short versus medium V1 analogue difference: 1.605 ± 0.464 versus 0.755 ± 0.489 ($t(9) = 2.106$, $p = .065$). English listeners (Figure 6d) discriminated directly covarying pairs significantly more easily than inversely covarying ones forpairs in which one of the V1 analogues was long (medium versus long V1 analogue duration difference: 2.088 ± 0.312 versus 1.024 ± 0.449, $t(14) = 4.027$, $p = .001$; short versus long V1 analogue difference: 2.855 ± 0.357 versus 1.672 ± 0.327, $t(14) = 6.801$, $p < .001$, cf. short versus medium V1 analogue difference: 1.500 ± 0.203 versus 0.980 ± 0.430, $t(14) = 2.205$, $p = .045$, marginal).

6.2.3 Summary—Performance was even more uniform for the non-speech than the speech stimuli: without exception, listeners from all languages discriminated the directly covarying non-speech stimuli better than the inversely covarying ones. The differences were not significant for all pairs in all languages, but there were not even any trends in the opposite direction.

6.3 Discussion

Linguistic experience does not appear to influence any listener's discrimination of speech or non-speech stimuli. The mechanism responsible for these responses appears to be the same one that determined all listeners' categorizations of the non-speech stimuli, either perceptual assimilation or post-perceptual addition.

Auditorism and autonomy both predicted that listeners would discriminate the non-speech stimuli in the same way they did the speech stimuli, while direct realism and interaction both predicted instead that they would discriminate them differently. As listeners discriminated the directly covarying pairs better than the inversely covarying pairs for both kinds of stimuli, auditorism's and autonomy's predictions are confirmed and direct realism's and interaction's predictions are disconfirmed.

7 Summary and general discussion

7.1 Summary

Table 5 summarizes the results obtained in the categorization and discrimination experiments by whether the judgments of the duration of the silent interval varied in inverse or direct proportion to the duration of the preceding vowel or vowel analogue.

For Japanese and English listeners, the likelihood of a “long” response was directly proportional to the duration of the preceding vowel or vowel analogue. The same was true of the Norwegian and Italian listeners’ categorization of the non-speech stimuli, but in their categorization of the speech stimuli, the likelihood of a “long” response was instead inversely proportional to the preceding vowel’s duration. Listeners from the four language backgrounds did not differ in their discrimination of the non-speech stimuli: all discriminated the directly covarying stimuli better than the inversely covarying ones, as did the Japanese, Norwegian and English speakers for the speech stimuli. Only Italian listeners were no better at discriminating the directly covarying speech stimuli than the inversely covarying ones. This set of findings suggests that, except in special circumstances, the judgments of silence duration were directly proportional to vowel duration.

7.2 Evaluating the predictions of competing theories once more

The auditorist model of speech perception distinguishes perceiving from hearing, unlike direct realism, while the autonomous model distinguishes perceiving from categorizing, unlike the interactive model. What support do the results reported here provide for distinguishing perceiving from either hearing or categorizing? Both distinctions would be supported if we could show that the acoustic properties of the signal were transformed perceptually even when the listeners know that those properties covary in the opposite direction in the pronunciations of speakers of their language.

The listeners’ categorization of these stimuli revealed little evidence that the signal’s acoustic properties might be transformed into auditory qualities opposite those expected from speakers’ pronunciations. Instead, it appeared that listeners added the duration of the preceding vowel to that of the consonant in determining where a stimulus fell along the consonant duration continuum, unless contrary linguistic experience prevented them from doing so. (In the next section, we develop the argument for addition rather than an assimilative perceptual transformation.) The only evidence for such stage is that some stage in processing prior to categorization must provide the quantities, specifically, the perceived durations, which listeners add together. But those quantities need not have undergone any auditory transformation prior to being added. Otherwise, the categorization data do not confirm the positive predictions of either autonomy or auditorism and thus seem to be more compatible with the less complex alternatives, interaction and direct realism.

Listeners’ discrimination of the stimuli leads to exactly the opposite conclusion. Because listeners did not discriminate the non-speech stimuli any differently from the speech stimuli, the predictions of interaction and direct realism that discrimination of the two kinds of stimuli would necessarily differ are disconfirmed. This apparent contradiction can be resolved if discrimination performance, like the categorization of all but the speech stimuli by the Norwegian and Italian listeners, is determined by listeners adding the duration of the preceding vowel or its analogue to the duration of the silent interval. In the next section, we develop the addition account in some detail.

7.3 Addition?

Addition does not actually alter an interval's perceived duration, so it does not transform the interval's acoustic properties. Perception must nonetheless be distinct from and prior to categorization and addition must be a post-perceptual process because it operates on the continuous outputs of the perceptual stage. But how are the durations of the two intervals added?

7.3.1 Simple addition—It is easy to dismiss simple addition, in which listeners respond to the summed durations of the vowel and consonant, because it incorrectly predicts that stimuli which have the same or very similar summed durations but different vowel and consonant durations would be very hard to discriminate. For example, the summed durations of the vowel and consonant intervals in the short-medium and medium-short stimuli are 138.7 and 137.8 ms, respectively, that is, their summed durations differ by only 0.9 ms, yet listeners were able to discriminate this pair of stimuli well above chance. The differences in the summed durations in the other negatively correlated stimuli are relatively larger but still absolutely small, 10.5 ms for the short-long versus long-short pair and 11.4 ms for the medium-long versus long-medium pair, but these stimuli were also discriminated well enough above chance that listeners are unlikely to have relied on this small difference in summed vowel and consonant duration alone.

The categorization data also show that listeners could not have responded simply to the summed durations of these two intervals, because they consistently assigned stimuli whose summed vowel and consonant durations are (nearly) the same to different categories, and they appeared to do so on the basis of differences in their constituent durations. The data in Table 6 shows categorization of the speech stimuli from the Japanese and English listeners and the categorization of non-speech stimuli from Norwegian and Italian listeners. Norwegian and Italian listeners' "long" responses to the speech stimuli are not shown because they decreased rather than increased as the preceding vowel got longer.

These results show that listeners from all four languages responded "long" less often as silence duration decreases, even though the increase in preceding vowel duration made up for that decrease nearly completely. Their responses were therefore not determined by the simple sum of the vowel or vowel analogue and silence durations.

Done through here.

7.3.2 Weighted addition—Instead of the vowel and silence durations being simply added together, one more ms of vowel may contribute less or more to changing the likelihood of a "long" response as one more ms of silence. Even if addition weights vowel duration differently than consonant duration, the vowel's duration may not actually transform the percept of the silence's duration, so long as the contributions to the combined judgment of the vowel's and silence's duration remain independent of one another.

We tested the hypothesis that the contributions of vowel's and silence's contributions are weighted but independent by constructing two-level hierarchical logistic regression models separately for each language's listeners' categorization of the speech and non-speech stimuli. The dependent variable was the frequency with which listeners categorized the stimuli as "short" versus "long". At the first level in the hierarchy, the independent variables were silence and vowel duration. The interaction between these durations was added at the next level.

Because the effect of vowel duration was much smaller for the shortest and longest silence durations, responses to the endpoint stimuli were left out of the analysis. Otherwise, all the analyses would have indicated that the effect of silence duration depended on vowel duration,

even though the effects of these variables might be independent for stimuli with intermediate consonant durations. Analyzing the intermediate stimuli alone thus makes it harder to mistake linear addition for a non-linear auditory transformation of the silence's duration by the vowel's.

The model including the interaction between vowel and consonant durations would fit the data better if its log likelihood ratio were significantly smaller than that at the previous level – this difference is distributed as the X^2 statistic. This model would also fit better if the coefficient (β) that represents the interaction's contribution to predicting the observed frequency of "long" responses were significantly different from zero.

Modeling began by pooling all the data obtained from a particular group of listeners in response to a particular stimulus set (e.g. all English listeners' responses to the speech stimuli). Because one listener's responses are very likely to be correlated with the responses of others in the same group, we then constructed jackknifed models in which each listener's data was left out in turn. The resulting partial estimates of the difference between the log likelihood ratios at the two levels in the hierarchy and the β values for vowel duration, silence duration, and their interaction can then be used to calculate less biased estimates of these values than in the model which includes all the listeners' data, and equally importantly, confidence intervals for these estimates (Table 7).

The criterial X^2 value for the difference between the log likelihood ratios to reach significance is 3.842, for $\alpha = .05$ in a model with 1 degree of freedom. Moreover, the confidence intervals of the jackknife estimates of the β value corresponding to the interaction should not include 0.[14]

Among the models of the pooled data, the change in the log likelihood ratio exceeds the criterial X^2 value only for the Italian listeners' responses to the speech stimuli; among the jackknifed models, it does for the Norwegian listeners' responses to the non-speech stimuli as well. However, for both cases, the 95% confidence intervals of the jackknife estimates of this change include 0. Adding the vowel by silence duration interaction does not therefore reliably improve the fit over the model in which vowel and silence duration are independent. The β value for this interaction in the models of all the data is only significant for the Italian listeners' responses to the speech stimuli, but the 95% confidence interval for this β value also includes 0. Otherwise, no β value for this interaction differs significantly from 0. Together, these findings indicate that the contributions of the vowel's and silence's durations are statistically independent of one another and thus provide little reason to reject the weighted addition model in favor of one in which vowel duration transforms the percept of silence duration.[15]

The weighted addition hypothesis is not undermined by the finding that Norwegian and Italian listeners respond "long" less often as the preceding vowel gets longer in the speech stimuli. Arithmetically, this poses no problem: each additional ms of vowel could *subtract* some amount from the combined duration of the vowel and silence, even if subtraction would be an entirely different psychological operation than addition. However, we already have a more plausible alternative account of the speech categorization data from these two languages: listeners'

[14]Both vowel and silence duration are expressed in ms in the modeling, so the sizes of the corresponding β values can be compared directly. The interaction of vowel and silence duration is expressed as the product of their individual durations, so the corresponding β values are unsurprisingly several orders of magnitude smaller.

[15]These findings also rule out an explanation in which listeners rely more on vowel duration when silence duration is ambiguous. Shifting to vowel from silence duration would, like addition, make listeners more likely to respond "long" when the preceding vowel is longer. However, such a shift in which interval listeners rely on also predicts that the effect of vowel duration should increase from the endpoints toward the middle of the silence duration continuum. If it did, then we would expect to find that vowel and silence duration interacted significantly. One could argue that we have effectively prevented this outcome by excluding responses to the endpoints of the silence duration continuum from our logistic regression models. That would imply that all the stimuli between the endpoints are equally ambiguous. A look back at the categorization functions in Figures 2 and 3 shows that this is not so.

experience of the inverse covariation between vowel and consonant duration in the speech of speakers of their native languages shifts their criterion for judging a consonant as "long" toward a shorter value after a shorter vowel.[16] Because criterion shifts in the other direction, toward a longer value after a shorter vowel, would produce results that look like the product of addition, we distinguish between their predictions in the next section.

A number of other outcomes of this modeling exercise are revelatory. For the Norwegian listeners, the β value for vowel duration is twice as large as that of silence duration for their responses to the speech stimuli, as well as being negative, while it is not significant and positive for their responses to the non-speech stimuli. This difference is yet further evidence that these listeners use a quite different mechanism to categorize the speech than the non-speech stimuli. In the models of the pooled Italian listeners' responses to the speech stimuli, the β value for silence duration is at least three times larger than the β values obtained for this variable in any other model. Moreover, the β value for the vowel by silence interaction is significant, while that for vowel duration is not. Silence duration is the primary correlate of the contrast between long and short consonants in this language, although its contribution is moderated as the preceding vowel gets longer. The difference between these β values and those obtained in modeling the Norwegian listeners' responses to the speech stimuli reinforces the conclusion that linguistic experience of speakers' pronunciations profoundly shapes how the two groups of listeners categorize such stimuli.

In this light, it is even more striking that the β values for these variables in the models of the English listeners' responses to the non-speech stimuli are so similar to those to speech stimuli. Unlike the Norwegian and Italian listeners, these listeners apparently relied on the same, non-linguistic mechanism in categorizing both kinds of stimuli.

In summary, listeners' categorization of these stimuli can largely be explained by the assumption that they added the durations of the vowel or vowel analogue and silence, although in unequal proportions. The only indication that the vowel's duration might have distorted the percept of the silence's duration was obtained from the Italian listeners' responses to the speech stimuli, but the jackknife estimates for the model of the Italian listeners' responses that included the interaction between vowel and silence duration were not reliably different from 0, so this indication is weak. Finally, differences in the weights assigned to the vowel's and consonant's durations by listeners from different languages further confirmed how closely listeners' responses correspond to speakers' productions.

7.4 Linguistic experience, criterion shifts, and addition

Linguistic experience influenced the categorization of the speech stimuli by Norwegian, Italian, and perhaps Japanese listeners. This influence can be implemented by criterion shifts: a shorter preceding vowel shifted the criterion for categorizing the following silent interval as "long" toward shorter silence durations for Norwegian and Italian listeners, and toward longer ones for Japanese listeners. A criterion shift in the Japanese direction could also have produced the English listeners' categorization of the speech stimuli, but they have no linguistic experience that would shift the criterion in this direction.

Even if criterion shifts reflecting linguistic experience can account for the Norwegian, and Italian listeners' categorization of the speech stimuli, they cannot account for their discrimination of these stimuli, which closely resembled the discrimination of these stimuli by Japanese and English listeners. A closer look at how the directly covarying stimuli might be

[16]As noted earlier, the Japanese listeners' categorization of the speech stimuli is actually ambiguous, as the increase in their "long" responses as the preceding vowel gets longer could be a product of addition or of linguistic experience of direct covariation of the consonant's and vowel's durations.

discriminated more easily than the inversely covarying ones shows that the discrimination results cannot in fact be determined by criterion shifts. Figure 7 shows the criterion shifts that would make a listener more likely to respond "long" after a longer vowel in the categorization task, i.e. to categorize the stimuli like a Japanese listener. Given that all listeners discriminated the directly covarying pairs better than the inversely covarying ones, this is the direction to consider in evaluating whether a criterion shift can account for the discrimination data.

The horizontal axis in the figure represents the distribution of perceptual values corresponding to a portion of the continuum of silent interval durations spanning a relatively short silence duration and a relatively long one. The two distributions represent the probability that the short (solid line) and long stimulus (dotted line) will be perceived as "short" and "long," respectively; that is, they are response likelihood distributions. The solid vertical line labeled "0" bisecting the horizontal axis is the criterion for the short:long decision when the preceding vowel is medium in duration. If the preceding vowel were instead short, then this criterion would shift rightward toward longer perceived durations, to the dashed vertical line labeled "S." This shift would have little effect on the likelihood that a listener would categorize a stimulus with the short silent interval (the solid distribution) as a short consonant because most of the distribution corresponding to the short silent interval was already to the left of the criterion. A shift of the criterion toward the short endpoint after a long vowel (to the dashed line labeled "L") would have an equally small effect on the likelihood that a stimulus with a long silent interval (the dotted distribution) is categorized as a long consonant because most of the distribution corresponding to that stimulus is already to the right of the criterion. These increases, of a little more than 6%, equal the areas under the solid distribution between the medium (0) and short (S) criteria and under the dotted distribution between the medium and long (L) criteria.

For the inversely covarying stimuli, the shift of the criterion from 0 to L after a long preceding vowel now substantially decreases the likelihood that the short silent interval would be categorized as "short," and that from 0 to S after a short preceding vowel likewise substantially decreases the likelihood that the long silent interval would be categorized as "long." These decreases equal the area under the dotted curve between 0 and S and that under solid curve between 0 and L. These biases, of a little more than 24%, are nearly four times as large as those produced when the vowel and silence durations covary directly.[17] These criterion shifts change the perceived category of the silence interval in the inversely covarying stimuli in a substantial number of cases rather than mildly exaggerating the strength with which it is perceived as the intended category, as they do in the directly covarying stimuli.

Despite these differences in the size and direction of the change in the probability of assigning the two silent intervals to "short" versus "long" categories, these criterion shifts still do not predict that the directly covarying stimuli will be more discriminable than the inversely covarying ones. In the directly covarying stimuli, "short" and "long" responses increase to over 99% for short-short and long-long stimuli, respectively, while in the inversely covarying stimuli, these responses decrease to just over 69%. But in the roughly 31% of trials where the short-long and long-short stimuli are misidentified as short-short and long-long, respectively, they are still identified with different categories, so they should be no less discriminable than the directly covarying stimuli. That is, changing the likelihood that the listener will assign a stimulus with a particular silence duration to a one category or the other, neither helps nor hinders discriminability.

Let us now reconsider the alternative, where the vowel and silent durations are added together. Even if the two durations contribute unequally to the sum, as indicated by the logistic regression

[17]The sizes of both changes in bias depend, of course, the separation between the response distributions and the size of the criterion shifts, but for any separation and criterion shift, the change in bias will be greater for the inversely than the directly covarying stimuli.

models, that sum will always differ more between the directly than the inversely covarying stimuli, and they should therefore always be easier to discriminate.

Besides correctly predicting the difference in discriminability between directly and inversely covarying stimuli, addition is more compatible than criterion shifts with our finding that discrimination of both the speech and non-speech analogues was uninfluenced by linguistic experience and the categories to which that experience would assign the stimuli in these experiments. The imperviousness of speech as well as non-speech discrimination performance to linguistic experience indicates that both kinds of stimuli are discriminated in terms of differences in the percepts of the continuous outputs of the auditory system rather than in terms of differences in categories. If listeners add the values of these percepts (unequally), then they should, as observed, discriminate the directly covarying stimuli better than the inversely covarying ones.

Because the quantities being added are the outputs of a prior processing stage, which we have labeled “perceptual,” addition is itself post-perceptual. If a post-perceptual stage can be distinguished from a perceptual stage on the basis of this evidence, then it is not only possible but indeed necessary to postpone the application of linguistic knowledge until the later post-perceptual stage, because if it applied earlier during the perceptual stage, it would pre-empt addition. It may appear that this is just what happened in the Norwegian and Italian listeners’ categorization of speech, but that is a product of categorization, and categorization cannot precede addition. Finally, if the basis for discrimination is the percepts of the continuous outputs of the auditory system, as those results suggest, then perceiving can be distinguished from categorizing, contrary to the central claim of the interactive theory.

In summary, the dissociation between speech categorization and discrimination, seen most clearly in the Norwegian and Italian listeners’ responses, separates perceiving from categorizing, while the absence of conclusive evidence that the auditory system transforms the signal’s acoustics fails to separate perceiving from hearing. The continuous pre-categorical output of the auditory system permits listeners to add the durations of successive intervals in the signal together when linguistic experience and the task do not pre-empt adding their values.

7.5 Limitations of the current results

The results reported here concern the perception of the duration of silent intervals, as a function of the duration of neighboring non-silent intervals. Although the two kinds of intervals differ acoustically from one another, the results indicate that listeners can add their durations together both in categorizing and discriminating the stimuli. Nonetheless, these results may not generalize to judgments of the durations of other adjacent intervals, such as the duration of formant transitions before vowels of different durations, where quite different interactions might be observed (see footnote 4 above).

It is also possible that listeners might behave differently in judging the duration of the vowel rather than the silent interval. Fowler (1992) reports that listeners were much more sensitive to differences in vowel than silence duration in stimuli like ours. This difference likely reflects a generally greater sensitivity to duration differences between filled than empty intervals: Abel (1972a,b) reports difference limens for sounds to be 4, 15, and 60 ms for baseline durations of 10, 100, and 1000 ms (or 40, 15, and 6 percent), respectively, but for silent intervals, they are 6–19 ms for a baseline duration of 10 ms (60 to 190 percent) and 61–96 ms for a baseline duration of 320 ms (17–27 percent).

7.6 Epilogue: Could the two intervals contrast after all?

We had expected to find evidence in these studies that the perceived duration of the consonant contrasted with that of the preceding vowel. The only result that looked like such evidence, Norwegian and Italian listeners' categorization of the speech stimuli, turned out instead to be probably evidence of a criterion shift determined by these listeners' experience of how these two intervals covary in native speakers' pronunciation of Norwegian and Italian.

Is there any reason still to think that a contrastive transformation might have altered listeners' duration percepts? Perhaps so. Fraisse (1963) reviews evidence which shows that in judging the durations of adjacent sounds, listeners minimize small differences between them while exaggerating large differences.[18] Fraisse attributes both tendencies to a "law of economy [in which] we spontaneously expect a stimulus to be around an average value and we tend to minimize small differences – law of assimilation – or to overestimate them if they are fairly large – law of contrast" (p. 120; see also p. 77).

Could our listeners have responded "long" more often after longer vowels and discriminated directly covarying stimuli better because they minimized perceived duration differences between adjacent intervals as a result of their being too similar? In the stimuli used in the categorization experiments, the duration of the preceding vowel (or its analogue) ranged from 49–89 ms and the duration of the silent interval ranged from 60–150 ms. The largest silence:vowel duration ratio is thus just over 3:1 (150:49) and the smallest is just under 2:3 (60:89). More important perhaps, only 6 of the 21 ratios exceed 2, so in the majority of stimuli, the durations of the two intervals may well have been similar enough that listeners did minimize the differences between them.

Results reported by Nakajima, ten Hoopen, Hilkuysen, & Sasaki (1992) encourage us to expect that we might obtain durational contrast if we used shorter vowels and longer silent intervals. On each trial in their experiment, listeners heard a target silent interval following a standard silent interval and adjusted the target's duration until it matched the standard's (each silent interval was bounded by brief tone complexes lasting 10 ms). The standard's duration ranged from 50–280 ms. In the test condition, the standard was preceded by another silent interval that always lasted 50 ms. Listeners adjusted the target's duration to a value less than the standard's for standard durations ranging from 40–140 ms and to a value greater than the standard for those lasting 200–280 ms – the target's duration was accurately matched to the standard's when it lasted 160 ms. That is, listeners perceived the standard to be shorter than it was for ratios between its duration and the preceding interval's up to nearly 3:1, and then perceived the standard as longer than it was for ratios of 4:1 and greater. This suggests that we should also look at a larger vowel:silence duration ratios.

An important open question is whether we can attribute any evidence of durational contrast that we might find with larger ratios to an auditory transformation. Moreover, the individual differences between listeners in our experiments as well as Kluender, et al.'s (1988) and Fowler's (1992) suggest the critical ratio for obtaining contrast effects could differ between listeners.

It's also possible that addition masks contrast. Suppose the listener is presented with two stimuli in which the silent interval between two vowels lasts 105 ms. In one, the preceding vowel lasts 45 ms, and in the other, it lasts 90 ms. Suppose also that durational contrast lengthened and shortened the silent interval's perceived duration by 15 ms after the short and long vowels to 90 and 120 ms, respectively. If listeners went on to add these perceived durations to the durations of the vowels, the sum would still be greater after the long vowel, 90 + 90 = 180 ms,

[18]We are deeply indebted to Kathryn Pruitt for bringing the studies discussed here to our attention.

than after the short vowel, 45 + 120 = 165 ms, even though durational contrast had operated. In fact, in this hypothetical case, durational contrast would have to lengthen and shorten perceived durations by more than the equivalent of 25 ms for the stimulus with the short vowel duration to sound longer than that with the long vowel after addition. That is, unless durational contrast could produce such large changes in perceived duration, its occurrence would be virtually undetectable after addition had applied, even if, as hypothesized, it is an obligatory perceptual transformation of the signal.

References

Abel SM. Duration discrimination of noise and tone bursts. Journal of the Acoustical Society of America 1972a;51:1219–1223. [PubMed: 5032936]

Abel SM. Discrimination of temporal gaps. Journal of the Acoustical Society of America 1972b;52:519–524.

Arakawa M, Kawagoe I. Eigo no onsetsugata to sokuon chikaku—nansensugo niyoru chikaku testuto no hookoku [English syllable structures and perception of geminacy—a report on perceptual tests using nonce words]. Journal of the Phonetic Society of Japan 1998;2:87–92.

Argiolas F, Federico M, Di Benedetto MG. Acoustic analysis of Italian [r] and [l]. Journal of the Acoustical Society of America 1995;97:3418.

Behne DM, Moxness B. Syllable- and rhyme-internal timing: postvocalic voicing and distinctive word length in Norwegian. PHONUM 1995;3:65–72.

Boersma P, Weenink D. Praat: Doing phonetics by computer Version 4.3.18. 2005

Brenner-Alsop E. Parsing time and rate normalization versus durational contrast. Journal of the Acoustical Society of America 2006;119:3241. (Abstract).

Diehl RL, Walsh MA. An auditory basis for the stimulus-length effect in the perception of stops and glides. Journal of the Acoustical Society of America 1989;85:2154–2164. [PubMed: 2732389]

Diehl RL, Walsh MA, Kluender KR. On the interpretability of speech/nonspeech comparisons: A reply to Fowler. Journal of the Acoustical Society of America 1991;89:2905–2909. [PubMed: 1918630]

van Dommelen W. Auditory accounts of temporal factors in the perception of Norwegian disyllables and speech analogs. Journal of Phonetics 1999;27:107–123.

Elman JL, McClelland JL. Cognitive penetration of the mechanisms of perception: Compensation for coarticulation of lexically restored phonemes. Journal of Memory and Language 1988;27:143–165.

Esposito A, Di Benedetto MG. Acoustical and perceptual study of gemination in Italian stops. Journal of the Acoustical Society of America 1999;106.4:2051–2062. [PubMed: 10530028]

Faluschi, S.; Di Benedetto, MG. Acoustic analysis of singleton and geminate affricates in Italian. The European Journal of Language and Speech, Feb. 2001. 2001. available at http://www.essex.ac.uk/web-sls/

Fintoft K. The duration of some Norwegian speech sounds. Phonetica 1961;7:19–39.

Fowler CA. An event approach to the study of speech perception from a direct realist perspective. Journal of Phonetics 1986;14:3–28.

Fowler CA. Sound-producing sources as the objects of perception: Rate normalization and nonspeech perception. Journal of the Acoustical Society of America 1990;88:1236–1249. [PubMed: 2229661]

Fowler CA. Auditory perception is not special: We see the world, we feel the world, we hear the world. Journal of the Acoustical Society of America 1991;89:2910–2915. [PubMed: 1918631]

Fowler CA. Vowel duration and closure duration in voiced and unvoiced stops: there are no contrast effects here. Journal of Phonetics 1992;20:143–165.

Fujisaki, H.; Kawashima, T. Some experiments on speech perception and a model for the perceptual mechanism; Annual report of the Engineering Research Institute, 29, Faculty of Engineering; Tokyo: University of Tokyo; 1970.

Fukui S. Perception for the Japanese stop consonants with reduced and extended durations. Onsei Gakkai Kaihou 1978;59:9–12.

Giovanardi, M.; Di Benedetto, MG. Acoustic analysis of singleton and geminate fricatives in Italian; WEB-SLS The European Journal of Language and Speech. 1998. p. 1-13.available at http://wrangler.essex.ac.uk/web-sls

Han M. Acoustic manifestations of mora timing in Japanese. Journal of the Acoustical Society of America 1994;96:73–82.

Hirata Y. Tango/bun reberu ni okeru sokuon no kikitori [Perception of geminate consonants at word and sentence level. Onsei Gakkai Kaihou 1990;194:23–28.

Kawahara S. Voicing and geminacy in Japanese: An acoustic and perceptual study. University of Massachusetts Occasional Papers in Linguistics 2005;31:87–120.

Kawahara S. Contextual effects on the perception of duration. Journal of Acoustical Society of America 2006a;119:3243.

Kawahara S. A faithfulness ranking projected from a perceptibility scale. Language 2006b;82:536–574.

Kingston, J. In: Frota, S.; Vigario, M.; Freitas, MJ., editors. From ears to categories: New arguments for autonomy; Proceedings of the first conference on phonetics and phonology in Iberia; Berlin: Mouton de Gruyter; 2005.

Klatt DH. Interaction between two factors that influence vowel duration. Journal of the Acoustical Society of America 1973;54:1102–1104. [PubMed: 4757455]

Kluender K, Randy D, Wright B. Vowel-length differences before voiced and voiceless consonants: An auditory explanation. Journal of Phonetics 1988;16:153–169.

Kristoffersen, G. The phonology of Norwegian. Oxford: Oxford University Press; 2000.

Kubozono, H. Mora and syllable. In: Tsujimura, N., editor. The handbook of Japanese linguistics. Oxford: Blackwell; 1999. p. 31-61.

Kusumoto K, Moreton E. Native language determines parsing of nonlinguistic rhythmic stimuli. Journal of Acoustical Society of America 1997;102:3204. (Abstract.).

Liberman AM, Cooper FS, Shankweiler DP, Studdert-Kennedy M. Perception of the speech code. Psychological Review 1967;74:431–461. [PubMed: 4170865]

Lisker L. Closure duration and the intervocalic voiced-voiceless distinction in English. Language 1957;33:42–49.

Lisker L. “Voicing” in English: A catalog of acoustic features signaling /b/ versus /p/ in trochees. Language and Speech 1986;29:3–11. [PubMed: 3657346]

Macmillan, N.; Creelman, D. Detection Theory: A User’s Guide. 2. Mahwah: Lawrence Erlbaum Associates Publishers; 2005.

Mattei, M.; Di Benedetto, MG. Acoustic analysis of singleton and geminate nasals in Italian. The European Journal of Language and Speech. 2000. http://www.essex.ac.uk/web-sls/

McClelland JL, Elman JL. The TRACE model of speech perception. Cognitive Psychology 1986;18:1–86. [PubMed: 3753912]

McClelland JL, Mirman D, Holt LL. Are there interactive processes in speech perception? TRENDS in Cognitive Sciences 2006;10:363–369. [PubMed: 16843037]

McQueen JM, Cutler A, Norris D. Phonological abstraction in the mental lexicon. Cognitive Science 2006;30:1113–1126.

McQueen JM, Norris D, Cutler A. The dynamic nature of speech perception. Language & Speech 2006;49:101–112. [PubMed: 16922064]

Norris D, McQueen JM, Cutler A. Merging information in speech recognition: Feedback is never necessary. Behavioral and Brain Sciences 2000;23:299–370. [PubMed: 11301575]

Norris D, McQueen JM, Cutler A. Perceptual learning in speech. Cognitive Psychology 2003;47:204–238. [PubMed: 12948518]

Ofuka E, Mori Y, Kiritani S. Sokuon no chikaku ni tausuru senkoo-kouzoku boin cho no eikyoo [The effects of the duration of preceding and following vowels on the perception of geminates]. Journal of the Phonetic Society of Japan 2005;9:59–65.

Parker E, Diehl R, Kluender K. Trading relations in speech and non-speech. Perception and Psychophysics 1986;39:129–142. [PubMed: 3725537]

Patel, AD.; Iverson, JR.; Ohgushi, K. Native language influences the perception of non-linguistic rhythm. In: Slifka, J.; Manuel, S.; Matthies, M., editors. From Sound to Sense (Abstract). 2004.

Pickett JM, Decker LR. Time factors in the perception of a double consonant. Language and Speech 1960;3:11–17.

Port R, Dalby J. Consonant/vowel ratio as a cue for voicing in English. Perception and Psychophysics 1982;32:141–152. [PubMed: 7145584]

Turk A, Shattuck-Hufnagel S. Word-boundary-related durational patterns in English. Journal of Phonetics 2000;28:397–440.

Watanabe S, Hirato N. The relation between the perceptual boundary of voiceless plosives and their moraic counterparts and the duration of the preceding vowels. Onsei Gengo 1985;1:1–8.

White LS. English speech timing: a domain and locus approach. University of Edinburgh PhD dissertation. 2002

Appendix 1

The tables below list the results of statistical tests of the effects of the between-subjects variable, button, and its interactions with within-subjects variables in the four experiments reported here.

Table A.1.1

Categorization. Speech stimuli.

Language	Button	Button by V1 duration
Japanese	$F(1, 18)= 1.329$, $p > .10$	$F(2, 36)=1.326$, $p > .10$
Norwegian	$F(1, 8)= 6.062$, $p = .039$*	$F < 1$
Italian	$F < 1$	$F < 1$
English	$F < 1$	$F < 1$

* Norwegian listeners who used the left button to give “long” responses gave this response more often than those who used the right button.

Table A.1.2

Categorization. Non-speech stimuli.

Language	Button	Button by V1 duration
Japanese	$F < 1$	$F < 1$
Norwegian	$F < 1$	$F < 1$
Italian	$F < 1$	$F < 1$
English	$F < 1$	$F < 1$

Table A.1.3

Discrimination. Speech stimuli.

Language	Button	Button by Covariation	Button by Difference	Button by Covariation by Difference
Japanese	$F(1,8) =3.158$, $p > .10$	$F(1,8) = 2.302$, $p > .10$	$F(2,16) = 1.03$, $p > .10$	$F < 1$
Norwegian		$F(1,8) = 1.119$, $p > .10$	$F < 1$	$F(2,16) = 1.348$, $p > .10$
Italian	$F(1,8) = 1.038$, $p > .10$	$F < 1$	$F(2,16) = 1.888$, $p > .10$	$F(2,16) = 1.128$, $p > .10$

Language	Button	Button by Covariation	Button by Difference	Button by Covariation by Difference
English	$F < 1$	$F < 1$	$F < 1$	$F(2,32) = 2.622$, $p = .088$

Table A.1.4

Discrimination. Non-speech stimuli.

Language	Button	Button by Covariation	Button by Difference	Button by Covariation by Difference
Japanese	$F < 1$	$F(1, 8) = 26.968$, $p = .001$*	$F < 1$	$F < 1$
Norwegian	$F < 1$	$F(1,8) = 1.357$, $p > .10$	$F < 1$	$F < 1$
Italian	$F(1,8) = 2.7$, $p > .10$	$F < 1$	$F(1,8) = 1.999$, $p > .10$	$F < 1$
English	$F(1,13) = 1.254$, $p > .10$	$F(1,13) = 5.963$, $p = .03$*	$F(2, 26) = 1.97$, $p > .10$	$F < 1$

* The advantage of directly over inversely covarying stimuli was larger for the Japanese and English listeners who responded "same" using the right button.

Appendix 2

The tables below list proportions correct (95% confidence intervals) for each of the discrimination experiments.

Table A.2.1

Speech.

Language	Short:Medium		Medium:Long		Short:Long	
	Inverse	Direct	Inverse	Direct	Inverse	Direct
Japanese	.59 (.03)	.80 (.05)	.73 (.09)	.75 (.05)	.88 (.05)	.94 (.02)
Norwegian	.53 (.03)	.59 (.03)	.80 (.06)	.82 (.05)	.88 (.04)	.90 (.04)
Italian	.54 (.02)	.59 (.06)	.64 (.07)	.64 (.06)	.74 (.08)	.70 (.08)
English	.51 (.02)	.57 (.03)	.61 (.05)	.64 (.06)	.68 (.05)	.73 (.06)

Table A.2.2

Non-speech

Language	Short:Medium		Medium:Long		Short:Long	
	Inverse	Direct	Inverse	Direct	Inverse	Direct
Japanese	.57 (.03)	.68 (.06)	.54 (.03)	.75 (.06)	.68 (.04)	.88 (.06)

Language	Short:Medium		Medium:Long		Short:Long	
	Inverse	Direct	Inverse	Direct	Inverse	Direct
Norwegian	.54 (.04)	.59 (.04)	.56 (.04)	.65 (.10)	.63 (.09)	.72 (.12)
Italian	.55 (.03)	.65 (.05)	.55 (.04)	.72 (.05)	.71 (.04)	.84 (.06)
English	.55 (.03)	.59 (.02)	.56 (.03)	.67 (.04)	.63 (.03)	.77 (.04)

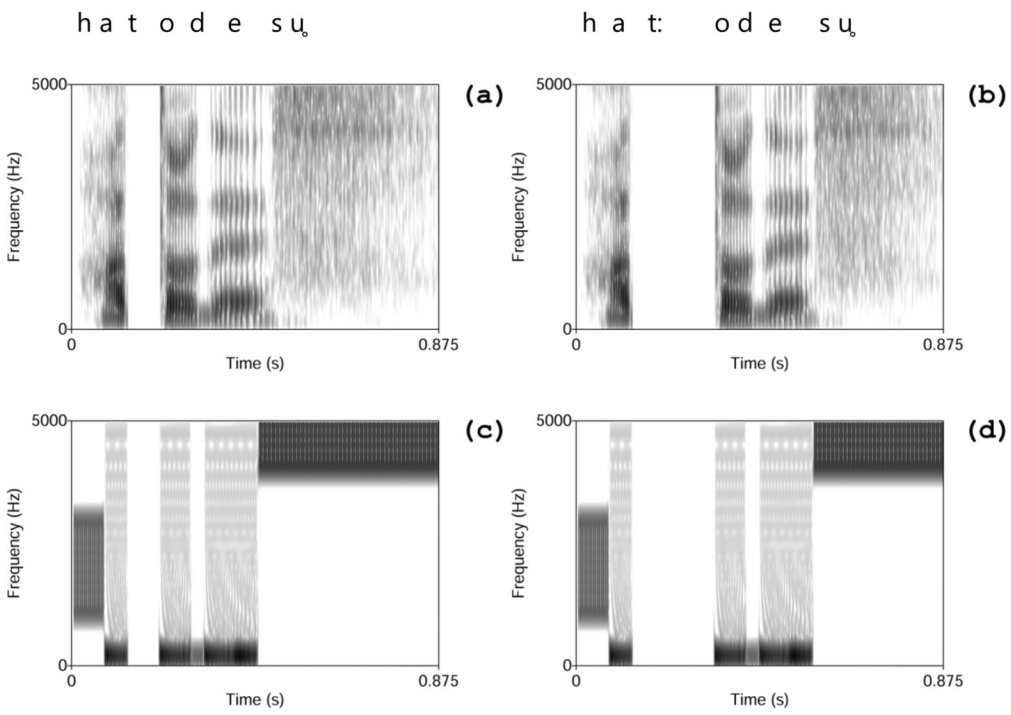

Figure 1.

Spectrograms of (a) a speech stimulus, [hatodesʊ̥], with a short closure, (b) a long closure, and (c, d) corresponding non-speech analogues. The [esʊ̥] intervals are shorter in the (b, d) than (a, c) to make up for the increase in the duration of the silent interval.

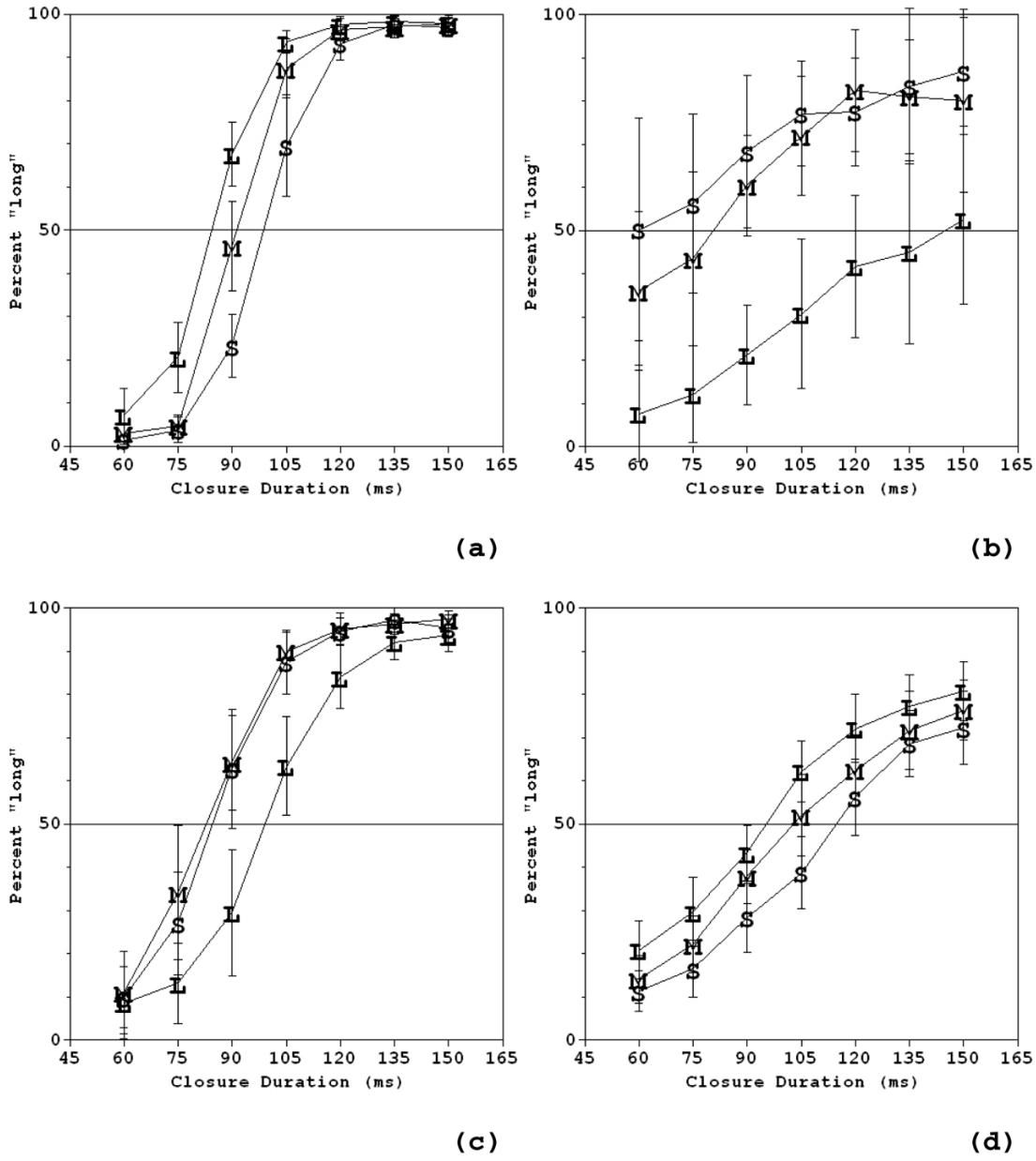

Figure 2.

Percentages of “long” consonant responses following short (S), medium (M), and long (L) vowels out of 30 trials/stimulus/listener for speech stimuli obtained from (a) 20 Japanese participants), (b) 10 Norwegian participants; (c) 10 Italian participants, and (d) 17 English participants. Error bars represent 95% confidence intervals.

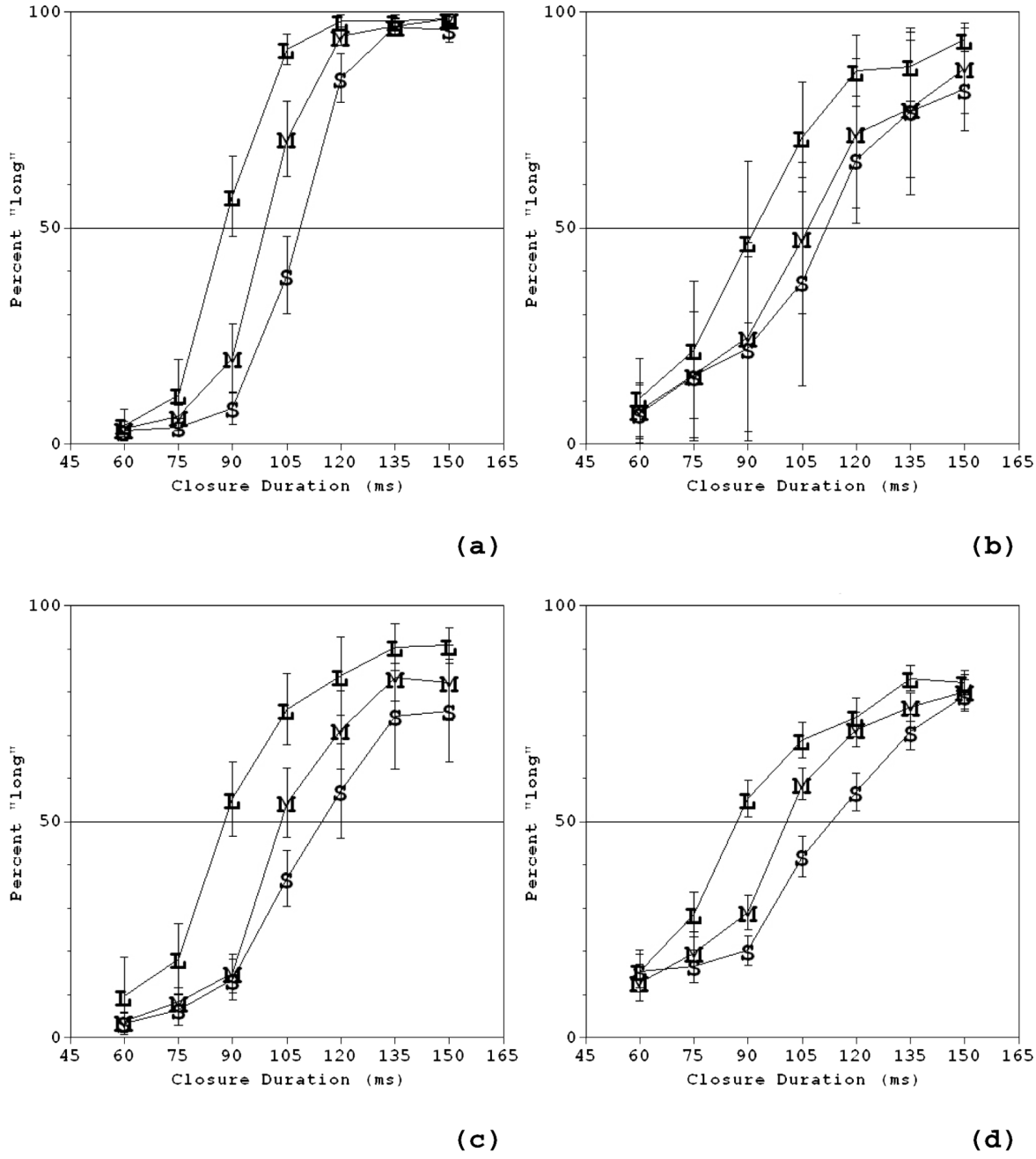

Figure 3.

Percentages of "long" responses to the non-speech stimuli following short (S), medium (M), and long (L) vowel analogues obtained from (a) 20 Japanese participants, (b) 10 Norwegian participants, (c) 11 Italian participants, and (d) 16 English participants.

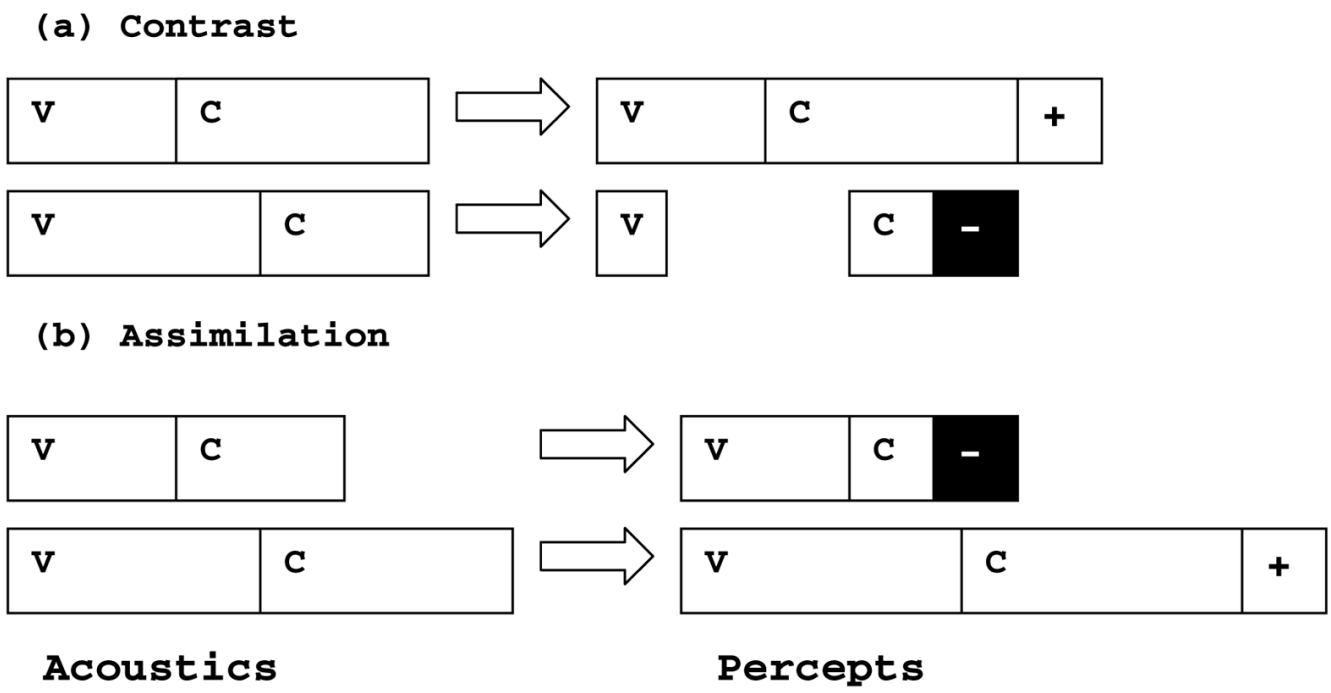

Figure 4.
Lengthening (white bar +) and shortening (black bar −) of perceived consonant duration as a function of perceptual (a) contrast with or (b) assimilation to the duration of preceding vowel.

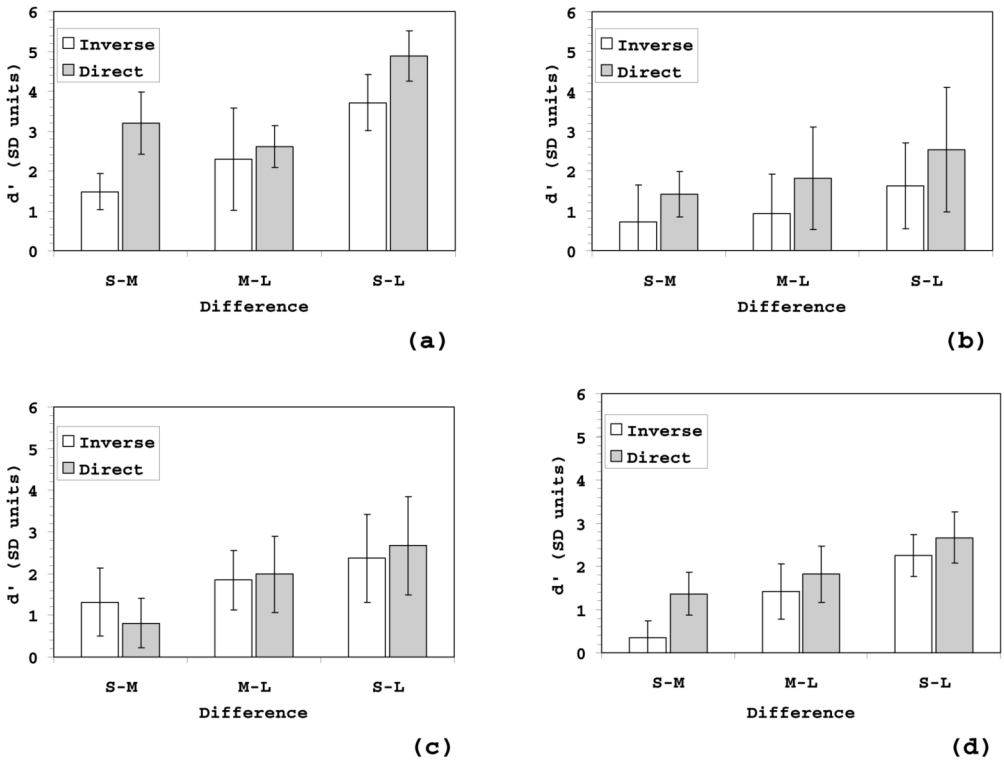

Figure 5.

Mean *d′* values with 95% confidence intervals in response to stimuli with short versus medium (S-M), medium versus long (M-L), and short versus long (S-L) vowel durations, where the consonant's duration varied inversely (white) or directly with the vowel's. (a) 10 Japanese participants, (b) 10 Norwegian participants, (c) 10 Italian participants, and (d) 18 English participants.

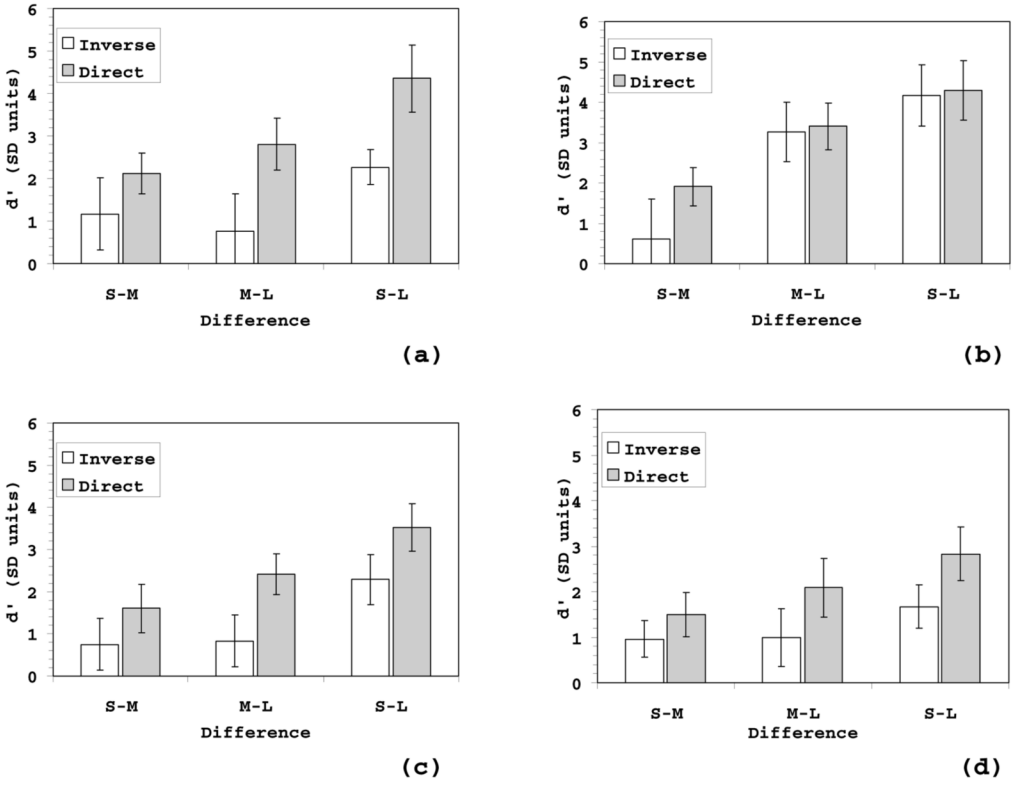

Figure 6.

Mean *d′* with 95% confidence intervals for discrimination of non-speech analogues by (a) 10 Japanese participants, (b) 10 Norwegian participants, (c) 10 Italian participants, and (d) 15 English participants.

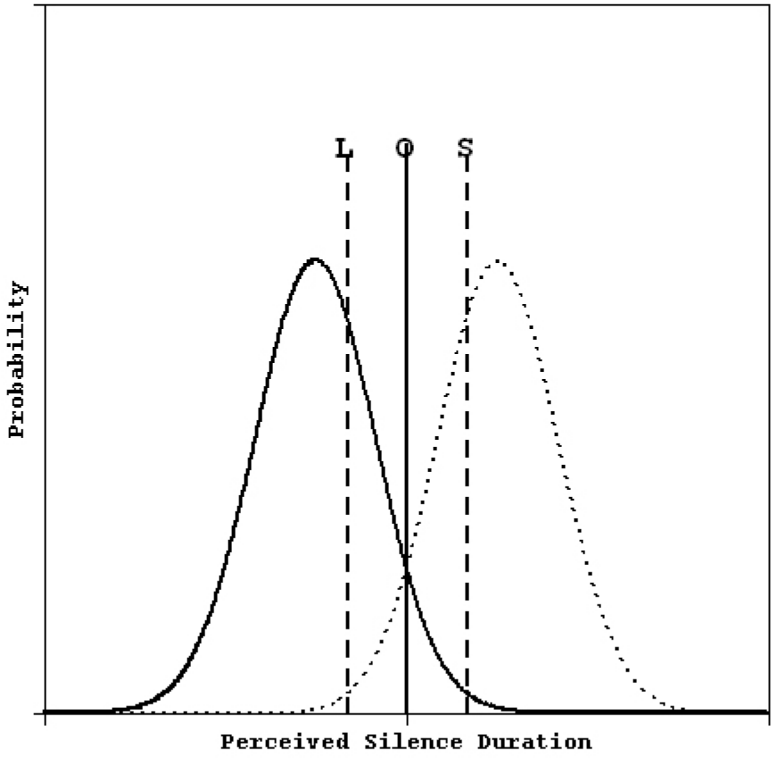

Figure 7.
Response distributions corresponding to a shorter and a longer closure duration (solid and dashed lines, respectively) and criteria (decision boundaries) corresponding to a medium (0), short (S), and long (L) duration of the preceding vowel.

Table 1

Duration ratios for vowels preceding short versus long consonants, for long versus short consonants, and for consonants to vowels when the consonant is short or long.

Language	Vowel Short:Long	Consonant Long:Short	Consonant:Vowel	
			Short	Long
Japanese	0.66–0.69	2.15–2.68	0.93–1.62	1.61–2.41
Norwegian	1.70–2.11	1.10–1.38	0.42–1.28	1.16–2.79
Italian	1.33–1.47	1.65–2.35	0.50–0.77	1.19–1.87
English*	1.32–136	1.21–1.31	0.46–0.66	0.83–1.04

* For English, the values are for voiced and voiceless stops, rather than short and long consonants, respectively. Japanese: Kawahara (2005, 2006b); Norwegian: Fintoft (1961); Italian: Argiolas, Macri, & Di Benedetto (1995), Giovanardi & Di Benedetto (1998), Esposito & Di Benedetto (1999), Mattei & Di Benedetto (2000), Faluschi & Di Benedetto (2001); and English: Raphael (1981).

Table 2

Durations of the constituent intervals in the stimuli.

	h	V1(=a)	Silence	od	esu	total
short V1	78	49	60	124	550	861
			75		535	
			90		520	
			105		505	
			120		490	
			135		475	
			150		460	
neut V1		63	60		536	
			75		521	
			90		506	
			105		491	
			120		476	
			135		461	
			150		446	
long V1		89	60		510	
			75		495	
			90		480	
			105		466	
			120		450	
			135		435	
			150		420	

Table 3

Characteristics of each interval in the non-speech stimuli.

[h]: Square wave band-pass filtered between 1000Hz and 3000Hz, that is, roughly the range of F2 and F3 in a vowel. Peak intensity at 0.1 of maximum.

[a]: Anharmonic complex (component frequencies separated by 0.101503 natural log units).

[t]: Silence.

[o]: Same as [a].

[d]: Square wave band-pass filtered between 50 Hz and 150 Hz. Peak intensity at 0.1 of maximum.

[e]: Same as [a].

[s]: Square wave band-pass filtered between 4000 Hz and 5000 Hz. Peak intensity at 0.1 of maximum. Last 50 ms ramped down to 0 intensity with a raised cosine window.

Table 4

Inversely and directly covarying stimulus pairs used in the discrimination tasks.

	Vowel-Consonant Covariation	
Difference	**Inverse**	**Direct**
Short-Medium	SM versus MS	SS versus MM
	49–90 versus 63–75	49–75 versus 63–90
Medium-Long	ML versus LM	MM versus LL
	63–105 versus 89–90	63–90 versus 89–105
Short-Long	SL versus LS	SS versus LL
	49–105 versus 89–75	49–75 versus 89–105

Table 5

Summary table of the overall results.

	Categorization		Discrimination	
	Speech	**Non-Speech**	**Speech**	**Non-Speech**
Japanese	Direct	Direct	Direct	Direct
English	Direct	Direct	Direct	Direct
Norwegian	Inverse	Direct	Direct	Direct
Italian	Inverse	Direct	Neither	Direct

Table 6

Average percentages of “long” responses to stimuli whose combined vowel and consonant durations are nearly the same, but which differ in the relative durations of each interval. Japanese and English values represent categorization of the speech stimuli, while the Norwegian and Italian values represent categorization of the non-speech analogues.

	short+long	**medium+medium**	**long+short**
language	**48.7+105=153.7**	**62.8+90=152.8**	**89.2+60=149.2**
Japanese	38.9	19.6	4.56
English	41.9	28.9	15.4
Norwegian	37.6	24.7	10.7
Italian	39.5	30.0	20.9
language	short+long	medium+medium	long+short
	48.7+120=**168.7**	62.8+105=**167.8**	89.2+75=**164.2**
Japanese	84.6	72.5	11.9
English	56.9	58.6	28.5
Norwegian	65.8	47.6	21.7
Italian	53.7	52.0	26.0
language	short+long	medium+medium	long+short
	48.7+135=**183.7**	62.8+120=**182.8**	89.2+90=**179.2**
Japanese	96.5	94.6	91.8
English	70.8	71.4	55.2
Norwegian	77.0	71.8	46.6
Italian	65.7	62.6	53.5

Table 7

J(apanese), N(orwegian), E(nglish), and I(talian); S(peech) and N(on-)S(peech). The top values in each cell were obtained from a model in which all the listeners' data are pooled, while the bottom values are obtained from jackknifing the data (see the text for explanation). Italics mark significance for the top values; significance can be inferred from the 95% confidence intervals supplied for the jackknife estimates. Values listed are the differences in log likelihood ratio (Δ-2LLR) between models in which silence and vowel duration are independent variables and those which include the interaction between them, and β values for the vowel by silence duration interaction and the independent variables representing silence duration and vowel duration.

Lg.	Stimuli	Δ-2LLR	Vowel by Silence	Silence	Vowel
J	S	0.828	−.00014	*.1321*	*.0562*
		−1.595 ± 8.222	−.00015 ± .00060	.1414 ± .0414	.0761 ± .0496
	NS	2.639	.000242	*.1075*	*.0344*
		3.0276 ± 9.2839	.000225 ± .000457	.1075 ± .0293	.0355 ± .0412
N	S	0.03	.00002	*.0174*	*−.0346*
		−0.844 ± .942	.00003 ± .00021	.0160 ± .0258	−.0343 ± .0441
	NS	2.379	.00015	*.0289*	.0063
		3.896 ± 6.707	.00017 ± .00024	.0269 ± .0334	.0043 ± .0211
E	S	1.519	−.00009	*.0375*	*−.0306*
		1.172 ± 6.127	−.00011 ± .00022	.0389 ± .0201	.0298 ± .0273
	NS	1.688	−.0001	*.0351*	*−.0243*
		1.150 ± 6.359	−.00016 ± .00020	.0307 ± .0253	.0214 ± .0269
I	S	*4.036*	*−.00031*	*.1010*	−.0053
		6.631 ± 8.355	−.00031 ± .00037	.1010 ± .0394	−.0053 ± .0291
	NS	3.111	−.00020	*.0747*	*.0460*
		3.761 ± 11.197	−.00022 ± .00045	.0751 ± .0376	.0478 ± .0338