

Title: Contextual Effects on the Perception of Duration

Authors: John Kingston, Shigeto Kawahara, Della Chambless, Daniel Mash, Eve

Brenner-Alsop

Affiliation: University of Massachusetts, Amherst

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

Abstract: In this paper, we report the results of experiments designed to test the competing predictions of direct realist vs auditorist models and of autonomous vs interactive models of speech perception. They do so by comparing Japanese, Norwegian, Italian, and English listeners' identification and discrimination of speech and non-speech analogue stimuli in which the durations of a vowel and a following consonant or their analogues were varied orthogonally. Vowel duration covaries directly with following consonant duration in Japanese but inversely in Norwegian, Italian, and English, so these experiments test the top-down application of linguistic experience on these tasks. The non-speech analogues are used to test the hypothesis that the acoustic properties of the signal are transformed during their passage through the auditory system. The identification of the speech stimuli by the Japanese, Norwegian, and Italian listeners but not the English listeners reflected their linguistic experience. Listeners' identification of the non-speech stimuli and their discrimination of both the speech and non-speech stimuli did not differ as a function of their linguistic experience, but instead appear to reflect their having added the two intervals' durations in responding to these stimuli and tasks. These findings delineate the limited scope of the top-down application of linguistic knowledge and thus support an autonomous model of speech perception over an interactive one. Because addition is post-perceptual rather than auditory, these results do not support an auditorist model over a direct realist one.

Key-words: perception, context effects, duration, singleton-geminate contrasts, speech vs non-speech, cross-linguistic comparisons

1 Introduction¹

In the auditorist theory of speech perception, three mental events occur during a trial in a speech perception experiment (Kingston 2005). The listener first hears the acoustic properties of the stimulus, then perceives the auditory qualities corresponding to those acoustic properties, and finally responds to those percepts. An object is associated with each event: acoustic properties with hearing, auditory qualities with perceiving, and a category with responding. There are two challenges to using this model to understand how listeners behave in such an experiment, one practical and the other theoretical. The practical challenge is that the experimenter has access to the first and last objects: the acoustic properties of the stimulus and the response the listener gives, but not to the second: the perceived auditory qualities. The perceived auditory qualities must instead be inferred from the relationship between the response and the acoustic properties, that is, the psychoacoustics, as well as from the psychophysics of the listener's task. The theoretical challenge is that theories of speech perception disagree profoundly about whether this model's three stages can really be distinguished, as well as whether speech perception is as strictly bottom-up and feed-forward as in the model we have just proposed. As most of this paper is a discussion of evidence that may be used to address the practical challenge of making sensible inferences about the perceived auditory qualities, we turn briefly here to a sketch of the theoretical challenge.

On the one hand, direct realism (Fowler 1986, 1990, 1991, 1992) denies that perceiving can be distinguished from hearing because sense experience is information about events in the world. If the sensory apparatus were to transform sense experience, the information in that experience would be distorted and rendered partly or completely unusable as a means of determining what event produced it. In other words, direct realism asserts that the auditory system is transparent to the acoustic properties of the stimulus. On the other hand, models of speech sound recognition such as TRACE (McClelland & Elman 1986; Elman & McClelland 1988; McClelland, Mirman, & Holt 2006), in which linguistic knowledge feeds back onto the perceptual evaluation of the signal's acoustic properties, deny that perceiving can be distinguished from responding because this linguistic knowledge does not merely alter the listener's bias to give a particular response, but also alters the perception on which that response is based.

The effects that a sound's acoustics have on the identification of its neighbor have been an important source of evidence in arguments developed by proponents of all these theories. In the case of duration, the acoustic property studied here, a listener's likelihood of judging a silent interval to be "long" can vary inversely or directly with the duration of a preceding vocalic (or vowel-analogue) context, but there is no resolution in the literature as to which behavior is more general. Previous studies have also left unanswered the question of whether a sound actually alters the perceived duration of a neighboring sound, and if so,

whether a sound's duration causes a neighboring sound's duration to be perceived as different from its own, i.e. to *contrast*, or similar to its own, i.e. to *assimilate*. Either process would be a perceptual distortion of the sounds' acoustic durations. The results presented in the literature could be accounted for equally well by criterion shifts induced by particular experimental demands or by linguistic experience. Our experiments are the first attempt to distinguish the empirical predictions of these two alternatives, and we conclude that listeners default to *adding* the durations of adjacent intervals when they judge how long they last.

We now review previous investigations of how a vowel's duration affects listeners' judgments of the duration of a following consonant. Kluender, Diehl & Wright (1988) found that English listeners judged silent intervals simulating stop closures as "short" or "voiced" more often after a long than a short vowel, and appealed to a perceptual-auditory mechanism to explain the effect. Kluender, et al. interpreted the length and voicing judgments as equivalent because voiced stops have shorter closure durations than voiceless ones, and English listeners use this difference to identify whether a stop closure is voiced (Lisker 1957 1986; Parker, Diehl, & Kluender 1986). Because the duration of the preceding vowel differs in the opposite direction — longer before the shorter closures of voiced stops — its duration could have influenced listeners' judgments of the silent interval's duration. However, Kluender et al. also found that listeners responded "short" more often to a given silence duration after a long square-wave non-speech analogue of the vowel. They argued that

listeners' linguistic experience cannot explain this finding because the non-speech stimuli do not sound at all like speech. They instead propose that an auditory transformation of the signal caused the silent interval to sound shorter after both a longer vowel and a longer vowel analogue. In this view, the auditory transformation distorts the silent interval's perceived duration *contrastively* with respect to the preceding vowel or vowel analogue's duration. These results and the theory behind them have been referred to as "durational contrast."

Fowler (1992) attempted to replicate Kluender et al's non-speech results, but found that listeners responded "short" more often when the preceding square-wave vowel analogue was short, rather than long. In these results, it appears that the silent interval *assimilates* to rather than contrasting with the preceding interval. Fowler obtained the same result using speech stimuli except when she asked listeners to treat the difference in silence duration as a correlate of the voicing contrast. In that case alone, she obtained an inverse relationship between the "long" (or unvoiced) response and the length of the preceding context, as Kluender et al. did with both their speech and non-speech stimuli. Otherwise, listeners' likelihood of a "long" response varied directly, regardless of whether the stimuli were speech or non-speech. Because these findings depended on the character of the stimuli, speech rather than non-speech, and the judgment, voicing rather than length, Fowler argued that the effect did not arise from an auditory transformation but instead from listeners' perception of the inverse covariation between the durations of the vowel and following consonant gestures in

sequences in which the consonant contrasts for voicing. Listeners would not perceive vowel and consonant gestures in the non-speech stimuli, and would not expect the durations of the vowel analogue and silent interval to covary inversely.

Van Dommelen (1999) reported that Norwegian listeners also treated an interval as long more often next to a long neighboring interval in non-speech stimuli, but as short more often in speech stimuli. Norwegian distinguishes short and long consonants (henceforth “singletons” and “geminate”), and the durations of preceding vowels covary inversely with the consonants’ durations. In fact, the preceding vowels differ more substantially in duration than the consonants themselves (Fintoft 1961; see §2 for details). For this reason, unlike Kluender et al. or Fowler, van Dommelen varied the preceding vowel duration incrementally rather than the silent interval’s duration, which had just two values — short and long. Listeners identified the speech stimuli as a word with a singleton (*mate* [ma:tə] “to feed”) or a geminate (*matte* [mat:ə] “mat”) rather than directly judging the length of the vowel or consonant. They judged the vowel to be “short” more often when the following silent interval was long rather than short. However, when the stimuli were instead non-speech analogues, listeners judged the vowel analogue as “short” more often when the following silent interval was short. The linguistic judgment that van Dommelen’s listeners made in response to the speech stimuli resembles the voicing judgment to speech stimuli made by one group of Fowler’s listeners in tapping into listeners’ experience of the inverse covariation between

vowel and following consonant duration in Norwegian.

Kluender, et al. found the number of “long” responses to be inversely proportional to the preceding vowel analogue’s duration in non-speech stimuli, and when listeners made a non-linguistic as well as a linguistic judgment of the speech stimuli, but Fowler and van Dommelen obtained inversely proportional judgments only when the stimuli were speech and the judgment linguistic.² The results leave open the question of whether inversely or directional proportional responses are more general, and whether either effect is attributable to an auditory distortion or linguistic experience.

Importantly, van Dommelen himself suggests a different alternative: that the vowel analogue’s duration may not have altered the evaluation of the following silent interval’s at all, but that listeners may instead have simply added the durations of two intervals together. In the conditions where target judgments vary directly with contextual duration in Fowler’s experiments, listeners may also have added the durations of the vowel’s or vowel analogue’s duration to the silent interval’s duration.

Our experiments examine 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. Like the studies just discussed, we also collected responses to non-speech analogues from other listeners. These responses can inform us as to whether the signal is transformed as it passes through the

auditory system because linguistic knowledge should exert no top-down influence on duration judgments of stimuli that do not sound at all like speech (Kingston 2005; also Norris McQueen, & Cutler 2000; cf. Kusumoto & Moreton 1997; Patel, Iverson, & Ohgushi 2004). If the auditory transformations of the perceived durations of successive signal intervals are universal, their influence should be discernable in responses not only to non-speech analogues but also to speech signals, even for listeners who speak languages that differ in how vowel and following consonant durations co-vary. That is, auditory transformations should exert a bottom-up influence on the durational percept even if the ultimate response to that percept is influenced by linguistic experience.

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 vs long-long, or short-long vs long-short. If the perceived duration of one interval assimilates perceptually to its neighbor, that will exaggerate the difference between the directly covarying stimuli and make them more discriminable than the inversely covarying ones. Contrast would have the opposite effect.

The directly and inversely covarying stimuli should, however, be equally easy to discriminate if the basis for discriminating them is the categories with which their constituent intervals are identified (Kingston 2005). At the categorization stage, the duration of a neighboring interval can only shift the criterion for deciding whether an interval is short or

long, and such shifts cannot exaggerate the difference between the directly covarying stimuli anymore than that between the inversely covarying ones.

We collected categorization and discrimination data from listeners whose native language is Japanese, English, Norwegian, or Italian. All four languages use the duration of a consonantal constriction to convey a phonological contrast, and in all four, the duration of the preceding vowel depends on the consonant's duration. Japanese, Norwegian, and Italian use constriction duration to convey a contrast between singletons and geminates, while in English, constriction duration is instead a correlate of the voicing contrast. 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. The languages also differ in the particular way in which the preceding vowel's duration covaries with the consonant's, as described in the next section.

2 Crosslinguistic differences in the covariation of vowel and consonant duration

2.1 Japanese

Table 1 shows that in Japanese, a vowel is shorter before a singleton than a geminate (Fukui 1978, Han 1994, Kawahara 2005, 2006b). Previous studies concerning the effect of

preceding vowel duration on the perception of the singleton:geminate contrast do not accord with one another. On the one hand, a shorter preceding vowel was found to bias speakers toward geminate responses (Watanabe & Hirato 1985, Hirata 1990) — an effect that runs contrary to expectations grounded in the production data just cited. How robust the effect is remains unclear as 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. On the other hand, two more recent studies have shown that a longer preceding vowel can induce more geminate responses (Arakawa & Kawagoe 1998, Ofuka, et al. 2005) — an effect that conforms to the production data.

*** Please put Table 1 here. ***

2.2 Norwegian

Norwegian also contrasts singleton and geminate consonants (Kristofferson 2000). The differences in constriction duration are, however, rather small, while those in the preceding vowel are rather large (Table 1), which suggests that the contrast is conveyed more by differences in the duration of preceding vowels, which are long before singletons and short before geminates. The consonant:vowel duration ratios for singletons and geminates also overlap, which indicates once again that it is more the differences in the vowel's rather than the consonant's duration that conveys the contrast. As discussed above, van Dommelen's

(1999) data show that a longer constriction duration causes Norwegian listeners to judge the preceding vowel as shorter. This result is expected if listeners' behavior is attuned to speakers' behavior, given that vowel and following consonant duration covary inversely in the Norwegian production data.

2.3 Italian

A series of studies by Di Benedetto and her colleagues have shown that vowels are consistently longer before singletons than geminates and that the consonant duration differences are large (Table 1). The values in the table indicate that, unlike Norwegian, the singleton:geminate contrast is conveyed more by consonant than preceding vowel duration in Italian, although it is enhanced by complementary differences in the duration of the vowel. Esposito & Di Benedetto (1999) tested the effect of preceding vowel duration on the perception of the singleton:geminate consonant contrast by Italian listeners and found that the boundary between the two 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. In other words, Italian listeners labeled more of a consonant duration continuum as "short" after a long vowel than after a short one, giving duration judgments that varied inversely with the duration of the preceding vowel.

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 (Table 1; see also Lisker 1957, 1986; Chen 1970; Raphael 1981; Port & Dalby 1982; Kluender, et al. 1988). The differences in vowel duration are as small as those observed in Italian, while the consonant duration differences are as small as those in Norwegian. The consonant:vowel duration ratios for voiced stops are as small as those observed for singletons in Norwegian and Japanese, but those for voiceless stops are not as large as those observed for geminates in any of the other three languages.

2.5 Summary

In Norwegian, Italian, and English, preceding vowel duration complements consonant duration, while in Japanese, preceding vowel duration increases with consonant duration. In Japanese, Norwegian, and Italian, the consonant durations reflect a contrast for length, while in English they instead reflect the voicing contrast.³ Finally, only Japanese contrasts vowels for quantity as well as consonants, although a contrastively long vowel cannot precede a geminate consonant inside a single morpheme (Kubozono 1999).

Because these languages differ phonologically and phonetically, listeners who speak

them may differ in how they respond to stimuli differing in consonant duration when the duration of the preceding vowel is varied orthogonally. Specifically, Japanese listeners may respond “long” or “geminate” more often when the preceding vowel is longer, while listeners from the other three languages may do so when the preceding vowel is shorter — some evidence confirming this prediction is discussed in the review above. Preceding vowel duration may influence consonant duration judgments more in Norwegian than in Italian, because the vowel duration ratios before singleton vs geminate consonants are greater in Norwegian than Italian, while the reverse is true for the geminate:singleton consonant duration ratios. English listeners could respond differently to both manipulations because consonants do not contrast for length in their language.

Although we predict that listeners from these different languages will show differences in how they respond to speech stimuli, we also predict that they will all respond similarly to non-speech analogues, as those responses depend on the linguistically-naïve, presumably universal auditory system.

Whether listeners from different languages respond differently in the discrimination task depends on whether the stimuli are discriminated only after they have been categorized. If listeners discriminate stimuli only after categorizing them, then they are predicted to respond differently, at least to the speech stimuli, depending on what their linguistic background is. However, if listeners instead discriminate stimuli on the basis of their auditory

qualities, we predict them to respond similarly regardless of their linguistic experience of the covariation between vowel and consonant duration.

In the next two sections, we present methods and results for the identification experiments, and then in the two following sections, do the same for the discrimination experiments.

3 Identification: Method

3.1 Stimuli

3.1.1 Speech

Consonant and vowel duration were varied orthogonally in our stimuli. Consonant duration varied incrementally from a value short enough for a singleton to one long enough for a geminate for Japanese listeners. The preceding vowels had three different durations, short, medium, and long. All speech stimuli were based on the Japanese minimal pair *hato* ‘dove’ and *hatto* ‘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 was chosen as the base for our continua. By choosing the geminate token with the shortest preceding vowel, we start with the one closest in duration to a vowel that would precede a singleton [t]. The duration of this vowel, 63 ms, is midway between the short vowel duration of roughly 40 ms observed before singletons and

the long vowel duration of roughly 80 ms observed before geminates. From this base stimulus, we extracted the four intervals that would compose the stimuli: [h], [a], [od], which included the preceding [t]-burst, and [esu] with the [u] devoiced as is usual in Japanese following a voiceless obstruent in utterance-final position.

The durations of the preceding [h] and the following [od] were held constant at 78 ms and 124 ms, respectively. The duration of the unedited [a] was 63 ms (we will refer to this V1 duration as the “medium V1”). To create a long and short preceding vowel for use as the two non-medium vowel contexts, we lengthened the original chosen token of *hatto* in its frame sentence by a factor of 1.4 and shortened it by a factor of 0.7, using the PSOLA function in Praat (Boersma & Weenink 2005). These factors yield a long V1 of 89 ms, and a short V1 of 49 ms. The [a] intervals were then extracted from these lengthened and shortened sentences, 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.

We then created a 7-step continuum of silent intervals ranging from 50 to 150 ms in 15 ms increments, which we inserted between the three [a] intervals and the following [od] interval, to simulate a continuum from a singleton to geminate stop, following short, medium, and long vowels. To compensate for the resulting differences in the total duration of the vowel-consonant interval, the duration of the [esu] interval was then manipulated in each stimulus via PSOLA to equalize the total stimulus durations. These adjustments should

prevent listeners from using total stimulus duration in either the categorization or discrimination tasks.⁴ The shortest combination of [a] and silent interval lasted 109 ms (49 ms V1 + 60 ms silence) and the longest 239 ms (89 ms V1 + 150 ms silence), so the longest [esu] interval differed from the shortest by 130 ms (239 – 108 ms). The duration of the longest [esu] was 550ms ([e] = 123 ms and [su] = 427 ms), and the duration of the shortest was 420 ms ([e] = 94 ms and [su] = 326 ms). Table 2 lists the durations of the constituent intervals that formed each stimulus:

*** Please put Table 2 here. ***

Error! Reference source not found, presents a spectrogram of one of our speech stimuli.

Deleted: Figure 1

*** Please put Figure 1 here. ***

After the intervals were concatenated, all stimuli were scaled to a peak intensity of 0.92 of the maximum intensity.

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 f_0 was 100 Hz, and it consisted of the 50 odd harmonics of this frequency. The ratios of their amplitudes were $1/\text{harmonic number}$, 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, i.e. 0.101503. Their amplitude ratios were $1/(2 \cdot \text{harmonic number}^2 + 1)$, 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 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). None of the listeners reported that these stimuli sounded like speech sounds. The characteristics of each portion of the non-speech stimuli are summarized in Table 3.

*** Please put Table 3 here. ***

The vowel analogues were all windowed by an off-and-on cosine ramp lasting 5 ms. The durations of all intervals were identical to those listed in Table 2. Figure 2 shows a spectrogram of one of these non-speech analogues.

*** Please put Figure 2 here. ***

3.2 Procedures

3.2.1 Listeners

Information about listeners is provided separately in the descriptions of the experiments for each language. No listener participated in more than one experiment. No listeners reported

any speech or hearing disorder, except one English listener in the speech identification experiment.

3.2.2 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—monitor, response box, 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 a few others were run off-campus in a quiet room; the remainder were run in quiet rooms at the Università degli Studi di Milan-Bicocca or the Scuola Normale Superiore di Pisa. 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.

3.2.3 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. For the English participants the response prompts were “Long” and “Short”, for Norwegian and Italian participants they were “t” and “tt”, and for Japanese participants, they were the *katakana* orthographic representations of “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, and were intended to teach the listeners what counted as “short” vs “long” (or the equivalent). After the 24 alternating trials, listeners worked through 6 more short training blocks, in which each endpoint in the three vowel contexts was presented once, in random order for, an additional 36 trials. The individual 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 1500ms response-period ended. The feedback appeared on the screen for 500 ms. The feedback was followed by the same 750-ms ITI and the presentation of the next stimulus. Listeners' responses to the training stimuli were not included in the subsequent analysis.

After finishing the training blocks, 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 section 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 run on-campus saw a different message instructing them to come out of the sound-attenuated room for a longer break. Listeners run off-campus were not instructed to leave the room, but were instructed to take a longer break before continuing. The longer breaks lasted roughly 5 minutes. The entire experimental session lasted under 60 minutes, including both written and verbal instructions, testing time, breaks, debriefing, and time for filling out consent forms and receipts for compensation.

3.2.4 Instructions

Before starting the experiment, listeners from languages other than Japanese were told they would hear a variety of syllable sequences and given the task of attending to the second consonant and deciding whether it sounded like a long or a short “t” (or the equivalent). The participants were told that during an initial training section, they would learn the two categories by first hearing them in alternation and then randomly, and by receiving feedback in the form of the correct answer after each response; they were then briefed about all the details of the trial structure. 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.

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, to allow them to respond without moving their arms or hands.

3.2.5 Conditions

Half the participants from each language used the left button for singletons and right button for geminates, with corresponding visual prompts, and vice versa for the remaining participants. Switching response assignment to buttons compensates for any predisposition

participants may have toward either the left or right response. The correspondence between color and hand remained constant across the two conditions: the left response was always red, and the right response always blue.

3.2.6 Non-Speech

The procedure used for categorizing the non-speech analogues was identical to that for speech outlined above, except for two details. First, all listeners used the arbitrary labels “A” and “B” for the left and right buttons, and learned these two categories during training. Half the listeners were taught during training that the analogue of the singleton 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). Second, listeners were told they would be hearing a variety of sound sequences, rather than syllables, and that their task was to label the sounds as “A” or “B” based on the categorization established during their initial training.

Although listeners from different languages were taught to associate different labels with the endpoints of the consonant duration continuum in the speech stimuli, and all listeners were taught to assign arbitrary labels in the non-speech experiments, for brevity’s sake we will henceforth refer to the number of “long” responses in discussing the results obtained in both the speech and non-speech experiments.

4 Identification results

4.1 Statistics

Repeated measures ANOVAs were run for each language, in which 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 vs medium vs long), and button (left vs right) was a between-subjects independent variable. *t*-tests were run to compare the “long” response proportions between long vs medium and medium vs short V1 durations. The alpha value was reduced to 0.025 to correct for multiple tests.

4.2 Japanese

4.2.1 Participants

20 Japanese listeners each participated in the speech and non-speech experiments for a total of 40 participants. They were recruited from International Christian University, the Tokyo University of Agriculture and Technology, Chuo University, and other, non-academic venues. The experiments were conducted in a quiet room. All the participants received payment of 1,000 yen for their participation. Although the listeners spoke a variety of different Japanese dialects, all varieties of Japanese contrast singleton and geminate consonants and use the same universal orthographic convention to represent them, so the

diversity should have no consequences for our experiments

4.2.2 Speech

Figure 3 plots the percentage of geminate responses across the continuum of silence duration after the three vowel contexts.

***** Please put Figure 3 here. *****

Listeners gave the geminate response more often when V1 was longer. The effect of V1 duration was significant ($F(2,36) = 93.040, p < .001$). The differences in the overall number of “t” responses between long V1 and medium V1 and between medium V1 and short V1 were also both statistically significant ($t(19) = 8.43, p < .001$; $t(19) = 7.71, p < .001$, respectively). The effect of button was not statistically significant ($F(1, 18) = 1.329, p = .264$), and it did not interact with V1 duration ($F(2, 36) = 1.326, p = .249$).

4.2.3 Non-speech

Figure 4 displays the percentage of long responses given in response to the non-speech stimuli.

***** Please put Figure 4 here. *****

Japanese listeners again gave the “long” response more often when the V1 analogue was

longer, and V1 duration was once again significant ($F(2, 36)=128.126, p<.001$). The differences between long V1 and medium V1 and between medium V1 and short V1 were both statistically significant ($t(19) = 14.73, p < .001$; $t(19) = 7.11, p < .001$, respectively). Button had no significant effect ($F < 1$) and did not interact with vowel duration ($F < 1$).

4.3 Norwegian

4.3.1 Participants

10 Norwegian listeners each took part in the speech and non-speech experiments. They were recruited from the University of Tromsø community.⁵ The listeners spoke different dialects, but singletons contrast with geminates in all dialects and the same orthographic convention is used to represent the contrast. Some participants additionally spoke Saami, a Finno-Ugric language. All participants were paid 10 Kroner for participating.

4.3.2 Procedure

The procedure used with the Norwegian participants was identical to that used with the Japanese participants, except that they were instructed to distinguish between single [t] and double [tt] during the instructions, and the alphabetic representations “t” and “tt” were used as response prompts.

4.3.3 Results: Speech

Figure 5 displays the results of the speech condition for Norwegian participants.

*** Please put Figure 5 here. ***

Norwegian listeners responded “t” more often when V1 was shorter, and the effect of V1’s duration was significant ($F(2,16) = 6.748, p = .007$). The number of “t” responses was significantly greater after the medium than the long V1 ($t(9) = 2.733, p = .023$) but there was only a non-significant trend toward more “t” responses in the short than the medium V1 ($t(9) = 1.582, p > .10$). The effect of button was statistically significant ($F(1, 8) = 6.062, p = .039$): listeners who used the left button to give geminate responses gave this response more often than those who used the right button (Left: 0.585 vs Right: 0.479). However, button did not interact with V1 duration ($F < 1$).

Unlike van Dommelen’s (1999) listeners, our listeners judged consonant duration rather than vowel duration, but like van Dommelen’s listeners they were more likely to treat one interval as long when the other was short than when it was long. The Norwegian listeners’ responses to the speech stimuli thus differed from the Japanese listeners, who responded “long” more often after a longer V1.

Figure 5 also shows that Norwegian listeners respond less categorically to differences in silence duration than Japanese listeners (cf. Figure 3 for Japanese). This finding is not surprising, given that singleton and geminate consonants differ less in duration in Norwegian than Japanese (see §2).

4.3.4 Results: Non-speech

Deleted: Figure

Error! Reference source not found,6 shows the Norwegian listeners' responses to the non-speech analogues.

*** Please put Figure 6 here. ***

Unlike their responses to the speech stimuli, the Norwegian listeners' responses to the non-speech stimuli resembled those of the Japanese listeners: more "long" responses after longer V1. V1 duration significantly affected geminate responses ($F(2,16) = 16.021, p < .001$). Geminate response proportions differed significantly after the three V1 durations (long V1 vs. medium V1: $t(9) = 3.843, p = .004$; medium V1 vs. short V1: $t(9) = 3.042, p = .014$). The effect of button was not statistically significant and did not interact with V1 duration (both F s < 1).

4.4 Italian

4.4.1 Participants

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. They were paid \$12/ hour for their participation. The remaining five for the non-speech experiment were students from the Università degli Studi di Milano-Bicocca and

the Scuola Normale Superiore di Pisa in Italy.⁶ The participants in the experiments carried out in Italy were paid at a rate of 10 euro/hour. No participants reported having any hearing or speaking disorder.

4.4.2 Procedures

The procedure used with Italian participants was identical to that used with Norwegian participants.

4.4.3 Results: Speech

Figure 7 displays the responses of the Italian listeners to the speech stimuli.

*** Please put Figure 7 here. ***

The Italian listeners responded “t” more often when the preceding vowel was short than when it was medium or long. The effect of vowel duration was significant ($F(2,16) = 26.244$, $p < .001$), but in the comparison between V1 durations only the difference between short and medium V1 was significant ($t(9) = 7.143$, $p < .001$; medium vs long V1: $t(9) = 1.362$, $p = .206$). Neither button nor its interaction with V1 duration was significant (both $F_s < 1$). These results resemble those obtained from Italian listeners by Esposito & Di Benedetto (1999).

4.4.4 Results: Non-speech

Figure 8 displays the responses of the Italian listeners to the non-speech stimuli.

*** Please put Figure 8 here. ***

Like the Norwegian listeners, Italian listeners' results on the non-speech experiment were opposite those of the speech experiment. Listeners gave more “long” responses when the V1 analogue was longer. V1 duration significantly affected geminate responses ($F(2,18) = 6.829$, $p = .006$). However, in the comparison between V1 durations, only the difference between medium and long V1 was significant ($t(10) = 2.982$, $p = .014$; medium vs short V1: $t(10) = 1.674$, $p > .10$). Neither button nor its interaction with V1 duration was significant (both F s < 1).

4.5 English

4.5.1 Participants

All participants were adult native English speakers recruited from the University of Massachusetts at Amherst community. They spoke a variety of American dialects. They earned either course credit or \$10 per hour for their participation. 17 listeners were run in the speech condition, and 16 in the non-speech condition. No participants reported having any hearing or speaking disorder, except for one in the speech experiment who reported high-frequency hearing impairment from exposure to loud music but who also responded

more accurately in training than any other listener.

4.5.2 Results: Speech

Figure 9 shows the the English listeners' responses to the speech stimuli.

*** Please put Figure 9 here. ***

These listeners gave more geminate responses when the vowel was longer. V1 duration significantly affected the frequency of geminate responses ($F(2,30) = 25.112, p < .001$). The differences between the three vowel conditions were significant (long V1 vs. medium V1: $t(16) = 4.532, p < .001$; medium V1 vs. short V1: $t(16) = 4.706, p < .001$). Button had no significant effects and did not interact with V1 duration (both $F_s < 1$).

Lengthening V1 increased geminate responses for English listeners just as it did for Japanese listeners and the English listeners in Fowler's (1992) study. Unlike the Japanese listeners, the English listeners in our experiment and Fowler's would have had no experience of the preceding vowel's duration varying directly with the duration of the consonant.

4.5.3 Results: Non-speech

Error! Reference source not found, displays the English listeners' responses to the non-speech analogues.

*** Please put Figure 10 here. ***

Deleted: Figure 10

As with the speech stimuli, the longer V1 analogues elicited more geminate responses. The duration of the V1 analogue significantly affected responses ($F(2,28) = 7.521, p = .002$). The differences between the three vowel conditions are all significant (long V1 vs medium V1, $t(15) = 2.671, p = .017$; medium V1 vs short V1, $t(15) = 2.445, p < .027$). Button had no effect and did not interact with V1 duration (both F s < 1). These non-speech results resemble those obtained by Fowler (1992) from English listeners in response to non-speech analogues.

4.6 Summary and interim discussion of identification results

Japanese and English listeners gave the geminate response to the speech stimuli significantly more often when V1 was longer. Norwegian and Italian listeners did the opposite: they gave the geminate response significantly more often when V1 was shorter. The difference in the effect of V1's duration between the Japanese listeners on one hand and the Norwegian and Italian listeners on the other can be readily attributed to differences in their linguistic experience: the vowel preceding a geminate in Japanese is longer than that preceding a singleton, while the opposite is true in Norwegian and Italian. The effect of V1's duration on the English listeners' responses cannot be attributed to linguistic experience in any obvious way. Since English listeners typically hear shorter vowels before the longer closures of voiceless stops, the English listeners in this experiment evidently did not treat the differences in closure duration as a correlate of the voicing contrast, and in this respect,

resemble the participants in Fowler's (1992) experiments who judged the duration of the stop closure and not its voicing. Unlike their responses to the speech stimuli, listeners from all four languages gave more geminate responses to the non-speech stimuli when the preceding V1 analogue was longer. The Japanese listeners' responses to the non-speech stimuli could reflect their linguistic experience or their categorization of the speech stimuli, but Norwegian and Italian listeners' cannot. However, all the non-speech responses could also be a product of a common mechanism that governs responses to the non-speech analogues for listeners from all four languages, while different mechanisms governed their responses to the speech stimuli.

One might be tempted to conclude that the common mechanism that governs all these listeners' responses to the non-speech analogues is an auditory transformation that causes the duration of one interval to be perceived as longer when the neighboring interval is longer (Kluender et al. 1988); this auditory transformation distorts the percept of duration such that a long target in a long context will sound even longer. An alternative explanation is that listeners simply add the duration of silence and preceding vowel. We will turn to the discussion of these two possibilities shortly below (see also §7 for more extensive discussion).

For categorization of the speech stimuli, it appears that more than one mechanism is at work. For the Norwegian, Italian, and Japanese listeners' responses to the speech stimuli,

that mechanism is a criterion shift reflecting linguistic experience. When speakers produce shorter vowels before a geminate, as Norwegian and Italian speakers do, listeners give more geminate responses after a short vowel, but when speakers produce longer vowels before a geminate, as Japanese speakers do, listeners instead give more geminate responses after a long vowel. That English listeners also give more geminate responses after a long vowel cannot be attributed to linguistic experience. As already noted, vowel duration in English does not covary directly with following consonant duration. Perhaps, in the absence of any pressure from linguistic experience to judge the durations of successive signal intervals contrastively, an assimilative auditory transformation influences responses. If so, then the Japanese listeners' responses could reflect this auditory transformation rather than their experience of the direct covariation between vowel and following consonant duration. Nonetheless, given that linguistic experience alone can explain the Norwegian and Italian listeners' responses, such experience cannot be ruled out as the force determining the Japanese listeners' responses.

Is either the direct or inverse response shift the product of a true perceptual distortion of the signal's acoustics? We have already shown that listeners only respond "long" more often after a shorter vowel if speakers of their language covary the vowel's duration inversely with the consonant's duration. That is, apparent cases of durational contrast in response to speech stimuli may be no more than the influence of linguistic experience. For

cases where the duration judgments are directly rather than inversely proportional, van Dommelen (1999) has already suggested that his Norwegian listeners, as well as Fowler's (1992) English listeners, may have simply *added* the durations of the vowel analogue and silent interval when categorizing non-speech stimuli.⁷ Addition can produce the appearance of assimilation, because a longer neighboring sound would lengthen the combined duration of the vowel and consonant intervals more and lead to a "long" judgment more often, but addition is not a genuine perceptual distortion of the signal's acoustics.

To investigate the effects of assimilation or its post-perceptual alternative, addition, further, we also had listeners discriminate stimuli in which vowel and consonantal silence duration differences were correlated. The discrimination results do not distinguish between these alternatives, but they do show us how general the byproducts of assimilation or addition are.

In one type of discrimination trial, the durations of these intervals covaried directly, i.e. the vowel and consonant durations were both short in one stimulus and both long in the other, while in the other type, these durations covaried inversely, i.e. in one stimulus the vowel was short and the consonant long and in the other the vowel was long and the consonant short. In both kinds of stimuli, the duration of the [esu] interval was again adjusted to compensate for the differences in combined vowel plus consonant duration caused by manipulating the vowel and consonant durations. If the durations of successive intervals

contrast perceptually, then the inversely covarying stimuli should be more discriminable, because contrast will particularly exaggerate the perceived differences in the durations of the two intervals.⁸ However, if these durations assimilate or add, then a short-short stimulus will sound especially short and a long-long one especially long, and the directly covarying stimuli should be more discriminable.

Running the discrimination task with both the speech stimuli and their non-speech analogues should also tell us whether differential linguistic experience of the covariation of these two intervals would influence 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 Discrimination: Method

5.1 Stimuli

The Japanese listeners' responses in the identification task were used to select three closure durations for use in the discrimination task: a short (S) closure of 75 ms, which was identified as geminate on less than 25% of trials with the medium V1, a long (L) closure of

105 ms, which was identified as geminate on more than 75% of trials with the medium V1, and a medium (M) closure of 90 ms midway between these values. The same three vowel durations from the identification experiment, 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).

*** Please put Table 4 here. ***

In the inversely covarying stimuli, vowel and silence duration differ in opposite directions in the two stimuli; in directly covarying stimuli, they differ in the same direction. The sizes of the differences in duration are the same in both sets of stimuli within a row.

5.2 Procedures

5.2.1 Locations and equipment

The discrimination data were collected in the same locations, using the same equipment as the identification data.

5.2.2 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” in English and Norwegian, and the equivalent words or phrases in Japanese or Italian. The visual prompts remained on the screen until the listener responded or for 1500 ms, whichever happened first. After the listeners responded or the 1500 ms 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. Finally, a 750 ms ITI elapsed before the next trial began.

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 were presented: 2 different trials, one for inversely covarying short-long vs long-short and the other for directly covarying short-short vs 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 the training section, listeners proceeded through 12 test blocks. Each test block included all the stimulus pairs listed in Table 4, with two presentations of each different pair, and an equal number of the corresponding same trials. Therefore, during the entire experiment, participants discriminated every different and same pair 24 times. 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 identification experiments described above.

5.2.3 Instructions

Before starting the experiment, listeners 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 that they would receive feedback in the form of the correct answer after they responded. They were told all the other details of the trial structure. Participants were also instructed to respond as quickly as possible. Listeners were again divided into two groups based on which response, “same” or “different,” was assigned to the left and right button.

5.2.4 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.

5.2.5 Participants

20 Norwegian participants were recruited from the University of Tromsø community (10 speech and 10 non-speech). 18 and 15 English speakers were recruited for the speech condition and non-speech conditions, respectively, and 10 Japanese speakers for each condition from the UMass Amherst community. The 20 Italian participants (10 speech and 10 non-speech) were recruited and run in Milano and Pisa, in Italy.

6 Discrimination: Results

6.1 Scoring the responses

The discrimination measure, d' , and bias measure, c , were calculated using the differencing rule, because our experiment used a roving same-different task (Macmillan & Creelman 2005). We report on the d' values from each language in detail below. The bias observed was largely consistent across languages: listeners were more biased to respond “same” to the inversely than the directly covarying stimuli and to differences involving the intermediate vowel duration. Biases generally mirrored discrimination success in that listeners responded “same” more often to stimuli that they discriminated less well. As bias does not contribute to our theoretical discussion, we do not report language-particular details below.

6.2 Statistical analysis

Repeated-measures ANOVA was run on the d' values, with correlation (directly vs inversely covarying) and difference (short vs medium, medium vs long, and short vs long) as within-subjects variables, and the button used for the “same” response (left vs. right) as a between-subjects variable. Values of d' obtained in response to inversely and directly covarying stimulus pairs for short vs medium, medium vs long, and short vs long differences were compared using t -tests, with the alpha value set at 0.0167 to correct for the effect of repeating these tests three times.

6.3 Results

6.3.1 Japanese: Speech

Figure 11 shows Japanese listeners' performance in discriminating the speech stimuli.

*** Please put Figure 11 here. ***

Japanese listeners discriminated directly covarying stimuli better than inversely covarying ones for short vs medium and short vs long differences but not for medium vs long differences. Correlation, difference, and their interaction were all significant ($F(1, 8) = 20.768, p = .002; F(2, 16) = 30.234, p < .001; F(2, 16) = 4.292, p = .032$). Directly covarying stimuli were easier to discriminate than inversely covarying ones when one of the intervals

was short (short vs medium: $t(9) = 5.424, p < .001$; short vs long: $t(9) = 5.947, p < .001$), but there was no difference between them when the difference was medium vs long ($t(9) = .057, p = .956$). Button and all interactions involving it were not significant (button: $F(1,8) = 3.158, p > .10$; button by correlation: $F(1,8) = 2.302, p > .10$; button by difference: $F(2,16) = 1.03, p > .10$; button by correlation by difference: $F < 1$).

6.3.2 Japanese: Non-speech

Figure 12 shows how well Japanese listeners were able to discriminate the non-speech analogues.

*** Please put Figure 12 here. ***

Japanese listeners discriminated the directly covarying non-speech stimuli better than the inversely covarying ones for all differences. The effects of correlation, difference, and their interaction are all significant ($F(1, 8) = 155.0, p < .001$; $F(2, 16) = 18.189, p < .001$; $F(2, 16) = 5.003, p = .021$). The difference in discriminability between directly and inversely covarying stimuli is only significant for differences involving the long vowel duration (short vs medium: $t(9) = 2.537, p = .032$ (marginal); medium vs long: $t(9) = 5.737, p < .001$; short vs long: $t(9) = 6.634, p < .001$). The interaction between button and correlation turned out to be significant ($F(1, 8) = 26.968, p = .001$): the advantage of directly over inversely covarying stimuli was larger for the participants who responded “same” with the right hand (average

positive vs negative difference in d' values: "same" right = 2.408, "same" left = 0.988). The main effect of button and all other interactions involving it were not significant ($F < 1$).

6.3.3 Norwegian: Speech

Figure 13 shows the success of the Norwegian listeners at discriminating the speech stimuli.

*** Please put Figure 13 here. ***

Norwegian listeners discriminated the directly covarying stimuli better than the inversely covarying ones for all differences between stimuli, although the difference was greater for the short vs medium than the medium vs long or short vs long difference. The effect of correlation was significant ($F(1, 8) = 9.998, p = .013$), as was that of difference ($F(2, 16) = 50.541, p < .001$), and these variables also interacted significantly ($F(2, 16) = 5.877, p = .012$): performance was only significantly better for the directly than the inversely covarying pairs for the short vs medium difference (short vs medium: $t(9) = 3.458, p = .007$; medium vs long: $t(9) = 1.024, p > .10$; short vs long: $t(9) = .427, p > .10$). Button and all interactions involving it were not significant (button by correlation: $F(1, 8) = 1.119, p > .10$; button by difference: $F < 1$; button by correlation by difference: $F(2, 16) = 1.348, p > .10$).

6.3.4 Norwegian: Non-speech

Figure 14 shows the Norwegian listeners' success at discriminating the non-speech

analogues.

*** Please put Figure 14 here. ***

Norwegian listeners were again more successful at discriminating the directly covarying stimuli than the inversely covarying ones for all differences, and their performance improved from the short vs medium to the medium vs long and to the short vs long difference.

Correlation and difference were again significant ($F(1, 8) = 10.022, p = .013$; $F(2, 16) = 7.023, p = .006$), but these variables did not interact ($F < 1$). The directly covarying stimuli were more easily discriminated than the negatively correlated ones but their advantage is only marginally significant for the short vs medium difference and not significant for the other differences (short vs medium: $t(9) = 2.744, p = .023$; medium vs long: $t(9) = 1.958, p = .082$; short vs long: $t(9) = 2.085, p = .067$). Button and all interactions involving it were not significant (all F s < 1 except button by correlation: $F(1,8) = 1.357, p > .10$).

6.3.5 Italian: Speech

Figure 15 shows the Italian listeners success at discriminating the speech stimuli:

*** Please put Figure 15 here. ***

Italian listeners differed from Japanese and Norwegian in that there was no significant effect of correlation: directly covarying and inversely covarying stimuli were discriminated more or else equally as well ($F < 1$). Average performance is actually better for the inversely than the

directly covarying pair when the difference is short vs medium. Difference, however, was significant ($F(2, 16) = 20.033, p < .001$). Despite performance being better for the inversely than directly covarying pair for the short vs medium difference but being better for the directly than inversely covarying pairs for the other two differences, correlation and difference did not interact with one another ($F(2,16) = 1.856, p > .10$). This lack of a significant interaction is a byproduct of the very large variances in these data. None of the pairwise comparisons of directly vs inversely correlated performance were even marginally significant (short vs medium: $t(9) = -.851, p > .10$; medium vs long: $t(9) = .322, p > .10$; short vs long: $t(9) = 1.003, p > .10$). Button had no significant effect on accuracy ($F(1,8) = 1.038, p > .10$), and did not interact significantly with any of the other variables (button by correlation: $F < 1$; button by difference: $F(2,16) = 1.888, p > .10$; button by correlation by difference: $F(2,16) = 1.128, p > .10$).

6.3.6 Italian: Non-speech

Figure 16 shows results of the non-speech experiment run on native Italian speakers.

*** Please put Figure 16 here. ***

Like listeners from Norwegian and Japanese, Italian listeners discriminated the directly covarying non-speech stimuli significantly better than the inversely covarying ones ($F(1,8) = 30.920, p < .001$). The advantage of the directly covarying pair over the inversely covarying

one was not significant for the short vs medium difference but was significant for other two differences (short vs medium: $t(9) = 2.106, p = .065$; medium vs long: $t(9) = 4.35, p = .001$; short vs long: $t(9) = 7.841, p < .001$). The effect of difference was also significant ($F(2,16) = 33.958, p < .001$). Difference was also independent from the direction of covariation ($F < 1$). Finally, button was neither significant on its own ($F(1,8) = 2.7, p > .10$), nor did it interact with any other variable (button by correlation: $F < 1$; button by difference: $F(1,8) = 1.999, p > .10$; and button by correlation by difference: $F < 1$).

6.3.7 English: Speech

Figure 17 shows how well English listeners were able to discriminate the speech stimuli.

*** Please put Figure 17 here. ***

For English listeners, too, the directly covarying stimuli were more discriminable than the inversely covarying ones. Discrimination also improved from the short vs medium to the medium vs long to the short vs long difference. The effects of correlation and difference were again significant ($F(1, 16) = 14.994, p = .001$; $F(2, 32) = 40.797, p < .001$). The interaction between these variables was nearly significant ($F(2, 32) = 3.297, p = .05$), because the directly covarying advantage was much smaller for the medium vs long and short vs long differences than the short vs medium difference. Directly covarying stimuli were more discriminable than inversely covarying ones for short vs medium and short vs long

differences, but not the medium vs long difference (short vs medium: $t(17) = 3.775, p = .002$; medium vs long: $t(17) = 1.618, p > .10$; short vs long: $t(17) = 2.952, p = .009$). Button had no significant effect and did not interact significantly with any other variable ($F < 1$ for all but button by correlation by difference: $F(2,32) = 2.622, p = .088$).

6.3.8 English: Non-speech

Figure 18 shows how well English listeners were able to discriminate the non-speech analogues.

*** Please put Figure 17 here. ***

Once again, the directly covarying stimuli were more discriminable than the inversely covarying ones, for all differences. Performance also improved from the short vs medium, to medium vs long, and short vs long differences as it did for the speech stimuli. The effects of both correlation and difference were significant ($F(1, 13) = 59.342, p < .001$; $F(2, 26) = 25.754, p < .001$), but they did not interact with one another ($F(2, 26) = 2.513, p > .10$).

Directly covarying stimuli were discriminated better than inversely covarying ones for all differences, but only those involving the long stimuli were significant (short vs medium: $t(14) = 2.205, p = .045$ (marginal); medium vs long: $t(14) = 4.027, p = .001$; short vs long: $t(14) = 6.801, p < .001$). Button was not significant ($F(1,13) = 1.254, p > .10$), but it did interact significantly with correlation ($F(1,13) = 5.963, p = .03$), because the difference between directly and inversely covarying stimuli was again greater for listeners who used to their right

hands to respond “same” (difference in d' values between directly and inversely covarying pairs is 1.241 for “same” right but only 0.644 for “same” left). No other interactions involving button were significant (button by difference: $F(2, 26) = 1.97, p > .10$; button by correlation by difference: $F < 1$).

6.4 Summary and discussion

Japanese, English, and Norwegian listeners all discriminated the directly covarying speech stimuli better than the inversely covarying ones, but the Italian listeners did not discriminate speech stimuli covarying in one direction any better than those covarying in the other. Listeners from all four languages discriminated directly covarying non-speech stimuli better than inversely covarying ones. Discrimination was generally better for the larger short vs long differences than the smaller short vs medium and medium vs long differences.

Discrimination of the speech stimuli matches categorization of those stimuli for the Japanese and English listeners but not for the Norwegian or Italian listeners. Listeners from the first two languages discriminated stimuli better than they were more likely to categorize differently, while listeners from the second two languages did not. The Norwegian listeners' actually discriminated stimuli less well than they were more likely to categorize as different, while the Italian listeners discriminated stimuli that they were more likely to categorize as different no better than those they were more likely to categorize as the same. Discrimination

of the non-speech matched categorization for listeners from all four languages.

7 Summary and general discussion

7.1 Summary

Table 5 summarizes the results obtained in the identification and discrimination experiments in terms of 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 interval.

*** Please put Table 5 here. ***

Japanese and English listeners' categorization of both speech and non-speech stimuli indicates that the likelihood of a geminate response is directly proportional to the duration of the preceding vowel. The same is 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 geminate response is instead inversely proportional to the preceding vowel's duration.

Listeners from the four language backgrounds do not differ in their discrimination of the non-speech stimuli: all discriminate the directly covarying stimuli better than the inversely covarying ones. The same pattern emerges in discrimination of speech stimuli among Japanese, Norwegian and English speakers. Only Italian listeners differ in that they were no

better at discriminating the inversely covarying speech stimuli than the directly covarying ones. This set of findings suggests that, except in special circumstances, the judgments of silence duration are directly proportional to vowel duration.

Given those results, four questions need to be answered: (1) why are Norwegian and Italian listeners' geminate responses inversely proportional to preceding vowel duration for the speech stimuli, but English and Japanese listeners' directly proportional, (2) why do speakers of these two groups of languages differ in their categorization of the speech but not the non-speech stimuli, (3) why do they also not differ in their discrimination of the stimuli, regardless of whether they are speech or non-speech (excepting those from Italian), and (4) why do all the results but the Norwegian and Italian listeners' categorization of the speech stimuli and the Italian listeners' discrimination of speech stimuli indicate that geminate responses are directly proportional to preceding vowel duration? Before attempting to answer any of these questions, we lay out the principal differences between the competing theories of speech perception from which we will craft answers.

7.2 The predictions of competing models

The auditorist theory of speech perception outlined at the outset of this paper distinguishes three stages within a trial in a categorization or discrimination experiment: hearing, perceiving, and responding. As indicated in the introduction, our theoretical goal was

to determine whether the unobservable stage in this model, perceiving, can be distinguished from hearing — contrary to direct realism — or from responding — contrary to feedback models.

In the auditorist theory, hearing provides the acoustic input to the auditory system, whose continuous outputs are percepts. The acoustic properties that a listener hears may be transformed in their passage through the auditory system. The potential auditory transformations of the acoustic properties that have interested us here are durational contrast and assimilation, that is, *distortions* of an interval's acoustic duration that cause it to be perceived as different from or similar to a neighboring interval's duration. Showing that a property of an interval's neighbor can distort the perception of one of its properties in this way would confirm the separation between perceiving and hearing that distinguishes the auditorist from the direct realist theory of speech perception.

After hearing and perceiving a stimulus, the percepts are mapped onto categories. Listeners ultimately make judgments that can be based on these categories or on the prior percepts — which one depends on the stimuli and task. If the responses are based on the categories, then they may be influenced by linguistic knowledge, which includes listeners' experience of the details of speakers' pronunciations of the categories. This knowledge can change the location of the criterion that the listener uses to categorize the continuum of percepts put out by the auditory system. If the responses are instead based on percepts, then

they may show no influence of linguistic experience, as that experience cannot alter the percepts. Showing that linguistic knowledge only changes response biases and not the percepts themselves would confirm the separation of responding from perceiving that distinguishes the feedforward auditory model we advocate from one that permits feedback.

Our results show that the percept of one interval's duration influences the judgment of its neighbor's in a way that's neither a criterion shift nor a perceptual distortion of its acoustics. In particular, other than the Norwegian and Italian listeners' categorization of the speech stimuli and, potentially, Italian listeners' discrimination of the speech stimuli, all our results can be explained as a consequence of listeners *adding* the duration of the vowel or vowel analogue to the duration of the following silent interval. Addition cannot plausibly be the product of any kind of auditory transformation if it is units of time that are being added together. Addition is also not a consequence of applying linguistic knowledge of speakers' pronunciations to categorizing or discriminating stimuli whose intervals differ in duration like these. Instead, listeners evidently add the two intervals together in performing the tasks we set them because it is often tactically advantageous to do so. We postpone further discussion of that tactical advantage until after we have presented the case that listeners do indeed add the durations of the two intervals, and only note here that if addition really occurs, it is a post-perceptual but linguistically naïve (or indifferent) process.

The first set of results we will interpret is the listeners' categorization of the speech

stimuli. We turn next to their discrimination of the speech and non-speech stimuli. After those results are discussed, we consider the evidence that listeners add the durations of the two intervals. Listener's categorization of the non-speech stimuli is taken up in the course of discussing this possibility. The discussion concludes with brief remarks on how these results might help us choose between the auditorist, direct realist, and feedback models of speech perception.

7.3 Categorization of speech stimuli

How listeners categorized the speech stimuli depended on how vowel and following consonant duration covaried in the speech of other speakers of their language. If these durations covary inversely, as they do Italian and Norwegian, then listeners were more likely to respond "tt" when the preceding vowel was short than when it was long, but if they instead covary directly, as in Japanese, then they were more likely to respond "tt" when the preceding vowel was long. These results agree entirely with previous studies of the perception of the geminate 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). The categorization of the speech stimuli by the English listeners, which most closely resembled that by the Japanese listeners, cannot be attributed to their having experienced these durations

covarying directly in the speech of other English speakers, but we postpone for the moment dealing with this problem in order to show how we model the effects of experience on the categorization of speech stimuli.

Experience with speakers' pronunciations could shift a listener's criterion (or response bias) for responding "t" depending on the preceding vowel's duration. In a language such as Norwegian or Italian, where that vowel is shorter before a long consonant, the criterion shifts toward the short end of the silence duration continuum when the preceding vowel is shorter, but toward the long end when that vowel is longer. The result is more "t" responses after short than long vowels. Short and long preceding vowels shift the criterion in the opposite directions in a language such as Japanese, where the preceding vowel is longer before a long consonant. If experience can only shift criteria, then the top-down application of linguistic knowledge has not actually altered the perception of the stimulus in any way, but only the likelihood with which a listener responds that it belongs to a particular category.

The English listeners' bias toward "t" responses following longer vowels appears not to reflect a top-down, experience-driven criterion shift. Unlike Norwegian, Italian, or Japanese, English does not contrast consonants for length, so English listeners cannot have any experience of vowel and following consonant duration covarying directly or inversely in the pronunciation of such a contrast. Indeed, English listeners have extensive experience of the durations of these two intervals covarying *inversely*, in sequences where the consonants

are obstruents contrasting for voicing. Our English listeners' bias toward more geminate responses after longer vowels is, however, not an aberration, as Fowler (1992) also obtained it from English listeners who categorized speech stimuli as containing a "short" or "long" consonant (cf. Kluender, et al. 1988).

Though English listeners' experience of the *inverse* variation of the duration of vowel and consonant in their voicing contrast did not cause them to respond like Norwegian or Italian listeners, perhaps experience with other patterns of durational variation might have influenced their categorization. We do not appeal to the only example of such experience that we know of, word-final lengthening, because that pattern of variation in duration has nothing to do with the pronunciation of segmental contrasts, there is no necessary connection between the facts, and because taking it seriously would unduly broaden the scope of explanations based on linguistic experience.⁹ We opt instead for an alternative mechanism that can supplant linguistic experience, if the stimuli or instructions do not lead listeners to conceive of their judgment in terms of a voicing contrast. This mechanism is addition, which, as we will show below accounts for speech discrimination by listeners of three of the four languages and non-speech categorization and discrimination by all listeners, and is therefore the obvious alternative for explaining English listeners' categorization of speech as well. Before discussing addition in detail, we discuss the discrimination results.

7.4 Discrimination

7.4.1 Speech

The categorization of speech stimuli by Japanese, Italian, and Norwegian listeners was very likely influenced by these listeners' experience of how singleton and geminate consonants are pronounced by native speakers of these languages. Japanese listeners' discrimination of these same stimuli could also have been influenced by such experience, but the Norwegian and Italian listeners' discrimination clearly was not – directly covarying stimuli were discriminated better by Norwegian listeners, and Italian listeners discriminated directly covarying stimuli as easily as inversely covarying ones even though vowel and silence duration covary inversely in Norwegian and Italian speakers' pronunciations. English listeners also discriminated the directly covarying stimuli better than the inversely covarying ones, a result which accords with their categorization of these stimuli, but which does not reflect any experience they might have of English pronunciations. The Japanese and English listeners' greater success at discriminating the directly covarying speech stimuli could be a product of the same criterion shifts that appear to have determined their categorization of these stimuli, but the Norwegian listeners' greater success with the directly covarying stimuli cannot. Because the performance of listeners from the three language backgrounds is so similar in this task, it's worth entertaining the notion that a single mechanism is responsible

for all three rather than a different mechanism for Norwegian than Japanese or English. We propose such a mechanism in §7.5, after showing how the similarity extends to non-speech discrimination in the next section.

Italian listeners did not discriminate the speech stimuli in the same way as listeners from the other three languages. They were no better at discriminating the directly than the inversely covarying pairs. Crucially, they were also *not* better at discriminating the inversely covarying pairs, a result that is also unexpected if linguistic experience of how vowel and silence duration covary in speakers' pronunciations influences perception.¹⁰

Because discrimination of the speech stimuli by the Norwegian and Italian listeners differs from what would be predicted from their categorization of them, the results obtained from these listeners are also incompatible with both the strong claim that categorization predicts discrimination (Liberman, Cooper, Shankweiler, & Studdert-Kennedy 1967) and the weaker one that acoustic differences provide a secondary, weaker basis than category membership for discriminating stimuli (Fujisaki & Kawashima 1970). The results from Norwegian at least, and Japanese and English as well, if they are determined by the same mechanism, suggest listeners discriminated perceptual qualities rather than categories. This possibility arises naturally in an auditory model in which perceiving is a distinct process from responding (Kingston 2005). This model predicts that listeners could discriminate the stimuli on the basis of the continuous output from the auditory system, that is, on the basis of the

percepts of the stimuli, before they are categorized. Discrimination can therefore be unaffected by any criterion shifts induced by linguistic experience.

7.4.2 Non-speech

Listeners whose native languages differ also did not differ in their discrimination of the non-speech analogues: they uniformly discriminated the directly covarying stimuli better than the inversely covarying ones. Linguistic experience is not expected to influence discrimination of non-speech stimuli; they are instead discriminated on the basis of the pre-categorical percepts of the continuous output of the auditory system.

7.5 Addition?

If linguistic experience is not responsible for listeners' discrimination of these stimuli, what is? Two alternatives merit consideration: a bottom-up auditory mechanism and a parallel post-perceptual process that is indifferent to linguistic experience. An auditory transformation could literally make the silent interval sound longer after a longer vowel; this is the perceptual distortion we have called "assimilation" above. Alternatively, listeners may have added the durations of the two intervals together when asked to discriminate them, as suggested by van Dommelen (1999). Adding the durations of the two intervals together

would make the directly covarying stimuli more different from one another, and make the inversely covarying ones more similar. We first discuss the possibility that listeners added the durations of the two intervals together in the next section before taking up again the possibility that V1's duration distorts the percept of the consonant's duration. We discuss these alternatives in this order because if addition explains our results, it is to be preferred over an auditory transformation because it would avoid adding an extra mechanism to psychological events that take place between hearing and responding.

7.5.1 Simple addition

It is easy to dismiss the simplest version of the addition story, 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 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 vs long-short pair and 11.4 ms for the medium-long vs long-medium pair, but these stimuli were also discriminated well enough above chance that listeners are unlikely to have relied on this

small duration difference alone.

The categorization data also show that listeners could not have been responding 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 appear 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 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.

*** Please put Table 6 here. ***

Table 6 shows that listeners from all four languages show fewer "long" responses as silence duration decreases, even though that decrease is nearly completely made up for by the increase in preceding vowel duration. Their responses are therefore not determined by the simple sum of the vowel or vowel analogue and silence durations.

7.5.2 Weighted addition

Instead of adding the vowel and silence durations together in this simple way, listeners might add just some of the vowel's duration to the silence's. A more general way to portray this alternative is that listeners combine the vowel's and the silence's durations in

such a way that the two durations contribute unequally to the judgment of their combined duration. That is, one more ms of vowel may not contribute as much (or as little) to changing the likelihood of a geminate response as one more ms of silence. If this is what listeners were doing, then the contributions to the combined judgment of the vowel's and silence's duration are independent of one another, and there is no basis for assuming that vowel's duration actually alters the percept of the silence's.

This weighted addition explanation was tested 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 in these analyses was the frequency with which listeners categorized the stimuli as "short" vs "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 stage. Because the effect of vowel duration was typically 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 addition for auditory distortion of the vowel's and silence's durations.

Two criteria were used to determine whether adding the interaction between

consonant and vowel duration significantly improved the fit of the model to the data and thereby disconfirmed the weighted addition model. First, was the log likelihood ratio at the level including this interaction significantly smaller than that at the previous level, in which the vowel and silence duration variables are independent? Second, is the coefficient (beta) that represents the interaction's contribution to predicting the observed frequency of "long" responses different from zero?

Modeling began with 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 it is very likely that the responses of one listener in a group will be correlated with those of another, we then constructed a set of jackknifed models in which each listener's data was left out in turn. Each of these models provides partial estimates of both the difference between the log likelihood ratios at the two levels in the hierarchy and the beta values for vowel duration, silence duration, and their interaction. These partial estimates can then be used to calculate less biased estimates of the difference between the log likelihood ratios and the betas than in the model which includes all the listeners' data, and equally importantly confidence intervals for these estimates. In Table 7 below, we report the estimates obtained from the model that included all the listeners' data and those obtained from the jackknife procedure; 95% confidence intervals are included for the latter.

If adding the vowel by silence duration interaction significantly improves the fit of the

model to the data, then the difference between the log likelihood ratios should exceed 3.842, the criterion value of X^2 statistic for $\alpha = .05$ in a model with 1 degree of freedom. Moreover, the confidence intervals of the jackknife estimates of the change in log likelihood ratio and the beta corresponding to this interaction should also not include 0.¹¹

*** Please put Table 7 here. ***

Among the models of all the data, the change in the log likelihood ratio exceeds the criterion 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 interaction does not reliably improve the fit over the model in which vowel and silence duration are independent. The beta 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 of the jackknife estimate for this beta also includes 0. Otherwise, no beta value for this interaction differs significantly from 0 in either the models of all the data or the jackknife models. Together, these findings provide little reason to reject the weighted addition model in favor of one in which the vowel duration alters the percept of silence duration.

It is worthwhile commenting briefly on a number of other outcomes of this modeling exercise. To begin with, beta values for vowel and silence duration differ markedly between

the models of the Norwegian listeners' responses to speech vs non-speech stimuli. The effect of vowel duration is twice as large as that of silence duration for the speech stimuli, as well as being negative, while it is not significant and positive for the non-speech stimuli. This difference is yet further evidence that listeners use a quite different mechanism to categorize the speech than the non-speech stimuli when moved to do so by linguistic experience.

In this light, the very similar beta values for these variables in the models of the English listeners' responses to the speech and non-speech stimuli are even more striking than they would be otherwise. Their similarity suggests that these listeners relied instead on the same mechanism in categorizing both kinds of stimuli, and that this mechanism is not driven by linguistic experience.

In the models of all the Italian listeners' responses to the speech stimuli, the beta value for silence duration is at least three times larger than the beta values obtained for this variable in any other model. Moreover, the beta 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 geminates and singletons in this language, although its contribution is moderated as the preceding vowel gets longer. The difference between these beta values and those obtained in modeling the Norwegian listeners' responses to the speech stimuli reinforces the conclusion that listeners' categorization of such stimuli are profoundly shaped by their linguistic experience of speakers' pronunciation of the contrasts.

Finally, we must distinguish between the specific proposal that listeners add vowel and silence durations in categorizing the stimuli and the more general claim that the effects of the vowel and silence durations on the frequency of geminate responses are statistically additive. Finding that including the vowel by silence interaction does not significantly improve the fit and that the beta value for this interaction is also not significantly different from 0 shows that vowel and silence duration are statistically independent of one another, and thus additive in a statistical sense.¹² These findings can furthermore be interpreted as evidence that Japanese and English listeners added the durations of the two intervals together, in unequal proportions, in categorizing the speech stimuli, and that listeners from all four languages did so in categorizing the non-speech stimuli. But what of the Norwegian and Italian listeners' categorization of the speech stimuli? The modeling indicates that the vowel and silence duration were certainly statistically additive for the Norwegian listeners and perhaps also the Italian listeners. However, the betas for vowel duration are negative in these two cases. Arithmetically, this poses no problem: each additional ms of vowel could *subtract* some amount from the combined duration of the vowel and silence, though subtraction would be an entirely different psychological operation than addition. However, we already have an alternative account of the effects of vowel duration on silence duration judgments of speech stimuli by listeners from these two languages: their experience of the inverse covariation between vowel and consonant duration in the speech of speakers of their languages shifts

their criterion for judging a consonant as geminate toward a shorter value after a shorter vowel. 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.

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 including the interaction between vowel and silence duration were not reliably different from 0, so this indication is weak. Finally, these tests of the weighted addition explanation provided further confirmation of how closely listeners' responses correspond to speakers' productions.

7.6 Choosing among models

We have not found evidence that one interval's duration distorts the perception of an adjacent interval's duration and thus that perceiving is distinct from hearing, as in the auditorist model. Our results are compatible with direct realism's claim that the acoustic properties of speech or other sounds are not transformed in their passage through the auditory system, but they do not provide positive confirmation of any of direct realism's central claims.

The first task in our future research on these phenomena is therefore to devise means of obtaining evidence that would confirm the predictions of one or the other theories of speech perception.

We have found evidence that categorization of speech stimuli is profoundly and subtly influenced by linguistic experience. We have described this influence as being implemented by criterion shifts: in Norwegian and Italian, a shorter preceding vowel shifted the criterion for categorizing the following silent interval as “long” toward shorter silence durations, while in Japanese, it instead shifted the criterion toward longer ones. Categorization of the speech stimuli by the Norwegian and Italian listeners can be accounted for by these criterion shifts, but their discrimination of these stimuli cannot be, because they incorrectly predict better discrimination of the negatively than the positively correlated stimuli. Some other mechanism is required to account for these listeners’ discrimination of the speech stimuli. The Japanese listeners’ discrimination of the speech stimuli could be accounted for by the same criterion shift as accounts for the categorization of these stimuli, but as they discriminate the positively correlated stimuli better than the negatively correlated ones just as the Norwegian and Italian listeners do, parsimony requires we attribute their behavior in the discrimination task to the same mechanism as drives the Norwegian and Italian listeners’ behavior in this task. The dissociation between discrimination and categorization in Norwegian and Italian listeners’ responses indicates that feedback from linguistic experience does not change the percepts of a

stimulus's acoustics but only changes the likelihood of a particular response by shifting the criterion for deciding whether a stimulus belongs to a particular category. If we also assume that Japanese listeners' discrimination is no more influenced by their linguistic experience, despite the apparent correspondence of discrimination to categorization in their responses, we must assume that the other mechanism can produce effects that look like criterion shifts.

We needed another mechanism anyway to explain English listeners' categorization of the speech stimuli, as they lack any relevant experience of a vowel's duration covarying directly with the following consonant's that would shift the criterion for giving a geminate response. Although their categorization of these stimuli resembles that by Japanese listeners', the resemblance is fortuitous because it is produced by a distinct mechanism, addition. Addition is also responsible for the English listeners' discrimination of these stimuli just as it is for the listeners from the other three languages.

Adding the durations of vowel and silence together readily explains the greater discriminability of the positively than the negatively correlated speech stimuli by listeners from all four languages: the summed durations of the short-short and long-long stimuli will always differ more than those of the short-long and long-short stimuli. Addition also explains all listeners' categorization and discrimination of the non-speech stimuli. Linguistic experience of covariation in speaker's productions determines whether categorizing the speech stimuli is governed by a criterion shift, but in the absence of such experience or in a

task like discrimination which is done on the basis of percepts rather than categories, they routinely add the two durations together.

A closer look at how the directly covarying stimuli might be discriminated more easily than the inversely covarying ones provides even more reason to think that the discrimination results are not determined by criterion shifts. Consider Figure 19, which 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):

*** Please put Figure 19 here. ***

The horizontal axis in the figure represents the perceptual values corresponding to a portion of the continuum of silent interval durations. This portion spans a relatively short silence duration and a relatively long one. The two distributions represent the probability that the short stimulus (solid line) and the long stimulus (dotted line) will be perceived as having short and long durations, respectively. The solid vertical line labeled “0” bisecting the horizontal axis is the criterion for the singleton:geminate 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 identify a stimulus with the short silent interval (solid line) as a singleton because most of the distribution corresponding to the short silent interval was already well to the left of the criterion. Similarly,

a shift of the criterion toward the singleton endpoint after a long vowel (to the dotted line labeled “L”) would have equally little effect on the likelihood that a stimulus with a long silent interval (dashed line) is identified as a geminate because most of the distribution corresponding to that stimulus is already well to the right of the criterion. The increases correspond to the areas under the solid curve between the medium (0) and short (S) criteria and under the dotted curve between the medium and long (L) criteria. For these distributions, the “singleton” and “geminate” percepts increase by a little more than 6% as a result of these shifts.

Let’s now consider the inversely covarying stimuli, where the short silent interval (solid line) follows a long vowel and the long silent interval (dotted line) follows a short vowel. The shift of the criterion from 0 to L after a long preceding vowel substantially reduces the likelihood that the short silent interval (solid line) would be identified as a singleton, and that from 0 to S after a short preceding vowel likewise substantially reduces the likelihood that the long silent interval (dotted line) would be identified as a geminate. These reductions are equal to the area under the dotted curve between 0 and S and that under solid curve between 0 and L. For these distributions, the likelihood of these responses decreases by a little more than 24%. These criterion shifts are thus likely to change the perceived category of the silence interval in the inversely covarying stimuli rather than 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 singleton vs geminates 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, “singleton” and “geminate” 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. Although criterion shifts do change the likelihood that the listener will assign a stimulus with a particular silence duration to a one category or the other, and they do so differently for directly than inversely covarying stimuli, changes in category assignment neither help nor hinder discriminability.

Let us now reconsider the alternative, where the vowel and silent durations are added together. Even if the vowel and silence durations contribute unequally to the sum, the summed durations of the two intervals 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 another

aspect of our results. Specifically, 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 before being categorized, and that the basis for performance is differences in the percepts of the continuous outputs of the auditory system. 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. But if listeners use the categories “singleton” vs “geminate” to discriminate the stimuli, then they should instead find both pairs of stimuli equally discriminable, regardless of how the vowel’s duration shifts the criterion for assigning stimuli to these categories. Finally, if the basis for discrimination is the percepts of the continuous outputs of the auditory system, as the discrimination results suggest, then perceiving can be distinguished from responding, contrary to the central claim of the feedback theory.

Remarks are called for before concluding about just what addition is. Above, we indicated that it is a post-perceptual process. By this we mean that it operates on the percepts put out by the auditory system. As we noted, these are continuous values rather than categories, and their form is thus appropriate for an operation such as addition.

We also suggested above that addition is tactical, that is, it is a way of responding to

the stimuli that lets the listener succeed at the tasks we set them. One indication that listeners responded tactically in this way is that a number of them who categorized the speech stimuli reported during debriefing that they used either the duration of the portion of the stimulus spanning the vowel and silence or the stimulus's apparent speed in deciding what response to give. Such reports were rarer from listeners who categorized the non-speech stimuli, but a number reported that the rhythm of the stimuli differed in briskness or in whether the beats corresponding to the vowel analogues seemed to have equal timing. These reports, too, indicate these listeners at any rate may also have decided how to respond by using the duration of the interval spanning the vowel analogue and silence. Responding in this way would have been strongly encouraged for the directly covarying stimuli by the feedback in the discrimination tasks. A final piece of evidence that addition is tactical is that linguistic experience can prevent listeners from adding in a speech categorization task, but apparently not in a discrimination task, where it is strongly encouraged. We are presently working on other means of turning addition on and off in an effort to better understand it.

In summary, the dissociation between speech categorization and discrimination separates perceiving from responding, 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 it is tactically advantageous to

do so.

References

- Arakawa, M., & Kawagoe, I. (1998). 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*, 2, 87-92.
- Argiolas, F., Federico M., & Di Benedetto, M.G (1995). Acoustic analysis of Italian [r] and [l]. *Journal of the Acoustical Society of America*, 97, 3418, 1995.
- Behne, D. M., & Moxness, B. (1995). Syllable- and rhyme-internal timing: postvocalic voicing and distinctive word length in Norwegian, *PHONUM*, 3, 65-72.
- Boersma, P., & Weenink, D. (2005). Praat: Doing phonetics by computer. Version 4.3.18.
- Brenner-Alsop, E. (2006). Parsing time: Rate normalization vs. durational contrast. *The Journal of the Acoustical Society of America*, 119, p. 3241 (Abstract).
- Diehl, R.L., & Walsh, M.A. (1989). An auditory basis for the stimulus-length effect in the perception of stops and glides. *The Journal of the Acoustical Society of America*, 85, 2154-2164.
- Diehl, R.L., Walsh, M.A. & Kluender, K.R. (1991). On the interpretability of speech/nonspeech comparisons: A reply to Fowler. *The Journal of the Acoustical Society of America*, 89, 2905-2909.
- Van Dommelen, W. (1999) Auditory accounts of temporal factors in the perception of

- Norwegian disyllables and speech analogs. *Journal of Phonetics*, 27, 107-123.
- Elman, J. L., & McClelland, J. L. (1988). Cognitive penetration of the mechanisms of perception: Compensation for coarticulation of lexically restored phonemes. *Journal of Memory and Language*, 27, 143-165.
- Esposito, A., & Di Benedetto, M. G. (1999). Acoustical and perceptual study of gemination in Italian stops. *Journal of the Acoustical Society of America*, 106.4, 2051-2062.
- Faluschi, S., & Di Benedetto, M.G. (2001) Acoustic analysis of singleton and geminate affricates in Italian. *The European Journal of Language and Speech*, Feb. 2001, available at <http://www.essex.ac.uk/web-sls/>
- Fintoft, K. (1961). The duration of some Norwegian speech sounds. *Phonetica*, 7, 19-39.
- Fowler, C. A. (1986). "An event approach to the study of speech perception from a direct realist perspective," *Journal of Phonetics*, 14, 3-28.
- Fowler, C. A. (1990). Sound-producing sources as the objects of perception: Rate normalization and nonspeech perception. *Journal of the Acoustical Society of America*, 88, 1236-1249.
- Fowler, C. A. (1991). Auditory perception is not special: We see the world, we feel the world, we hear the world. *Journal of the Acoustical Society of America*, 89, 2910-2915.
- Fowler, C. A. (1992). Vowel duration and closure duration in voiced and unvoiced stops: there are no contrast effects here. *Journal of Phonetics*, 20, 143-165.

- Fujisaki, H., & Kawashima, T. (1970). Some experiments on speech perception and a model for the perceptual mechanism. Annual report of the Engineering Research Institute, 29, Faculty of Engineering, University of Tokyo, Tokyo.
- Fukui, S. (1978). Perception for the Japanese stop consonants with reduced and extended durations. *Onsei Gakkai Kaihou*, 59, 9-12.
- Giovanardi, M., & Di Benedetto M.G. (1998). Acoustic analysis of singleton and geminate fricatives in Italian. *WEB-SLS The European Journal of Language and Speech*, 1-13, available at <http://wrangler.essex.ac.uk/web-sls>.
- Han, M. (1994). Acoustic manifestations of mora timing in Japanese. *Journal of the Acoustical Society of America*, 96, 73-82.
- Hirata, Y. (1990). Tango/bun reberu ni okeru sokuon no kikitōri [Perception of geminate consonants at word and sentence level. *Onsei Gakkai Kaihou*, 194, 23-28.
- Kawahara, S. (2005). Voicing and geminacy in Japanese: An acoustic and perceptual study. *University of Massachusetts Occasional Papers in Linguistics*, 31, 87-120.
- Kawahara, S. (2006a). Contextual effects on the perception of duration. *Journal of Acoustical Society of America*, 119, 3243. (Abstract).
- Kawahara, S. (2006b). A faithfulness ranking projected from a perceptibility scale. *Language*, 82, 536-574.
- Kingston, J. (2005) From ears to categories: New arguments for autonomy. In S. Frota, M.

- Vigario, & M. J. Freitas (Eds.), Proceedings of the first conference on phonetics and phonology in Iberia. Berlin: Mouton de Gruyter.
- Klatt, D. H. (1973). Interaction between two factors that influence vowel duration. *Journal of the Acoustical Society of America*, 54, 1102-1104.
- Kluender, K. R., Diehl, R. L, and Wright, B. (1988). Vowel-length differences before voiced and voiceless consonants: An auditory explanation. *Journal of Phonetics*, 16, 153-169.
- Kristoffersen, G (2000). *The phonology of Norwegian*. Oxford: Oxford University Press.
- Kubozono, H. (1999). Mora and syllable. In N. Tsujimura (ed.), *The handbook of Japanese linguistics* (pp. 31-61). Oxford: Blackwell.
- Kusumoto, K., and Moreton E. (1997). Native language determines parsing of nonlinguistic rhythmic stimuli. *Journal of Acoustical Society of America*, 102, 3204. (Abstract.)
- Lieberman, A. M., Cooper, F.S., Shankweiler, D.P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, 74, 431-461.
- Lisker, L. (1957). Closure duration and the intervocalic voiced-voiceless distinction in English. *Language*, 33, 42-49.
- Lisker, L. (1986). "Voicing" in English: A catalog of acoustic features signaling /b/ versus /p/ in trochees. *Language and Speech*, 29, 3-11.
- Macmillan, N., & Creelman, D. (2005). *Detection Theory: A User's Guide*. (2nd Edition). Mahwah: Lawrence Erlbaum Associates Publishers.

- Mattei, M. and Di Benedetto, M.G. (2000) Acoustic analysis of singleton and geminate nasals in Italian. *The European Journal of Language and Speech*, <http://www.essex.ac.uk/web-sls/>
- McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, 18, 1-86.
- McClelland, J.L., Mirman, D., & Holt, L.L. (2006). Are there interactive processes in speech perception? *TRENDS in Cognitive Sciences*, 10, 363-369.
- Miller, J.L., & Liberman, A.M. (1979). Some effects of later-occurring information on the perception of stop consonant and semivowel. *Perception & Psychophysics*, 25, 457-465.
- Norris, D., McQueen, J. M., & Cutler, A. (2000). Merging information in speech recognition: Feedback is never necessary. *Behavioral and Brain Sciences*, 23, 299-370.
- Ofuka, E., Mori Y., & Kiritani, S. (2005). 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*, 9, 59-65.
- Parker, E., Diehl, R., & Kluender, K. (1986). Trading relations in speech and non-speech. *Perception and Psychophysics*, 39, 129-142.
- Patel, A. D., Iverson, J. R., & Ohgushi, K. (2004). Native language influences the perception of non-linguistic rhythm. In J. Slifka, S. Manuel, & M. Matthies (Eds.), *From Sound to Sense*. (Abstract).

- Pickett, J. M., & Decker, L. R. (1960). Time factors in the perception of a double consonant. *Language and Speech*, 3, 11-17.
- Port, R., & Dalby, J. (1982). Consonant/vowel ratio as a cue for voicing in English. *Perception and Psychophysics*, 32, 141-152.
- Turk, A., & Shattuck-Hufnagel, S. (2000). Word-boundary-related durational patterns in English. *Journal of Phonetics*, 28, 397-440.
- Watanabe, S., & Hirato, N. (1985). The relation between the perceptual boundary of voiceless plosives and their moraic counterparts and the duration of the preceding vowels. *Onsei Gengo*, 1, 1-8.
- White, L. S. (2002). English speech timing: a domain and locus approach. University of Edinburgh PhD dissertation.

Figure legends.

Figure 1: Spectrogram of a speech stimulus: [hatodesu].

Figure 2: Spectrogram of a non-speech analogue.

Figure 3: Percentages of “long” responses out of 30 trials/stimulus/listener for speech stimuli (20 Japanese participants). Error bars represent 95% confidence intervals.

Figure 4: Percentages of “long” responses for non-speech stimuli (20 Japanese participants).

Figure 5: Percentages of geminate responses out of 30 trials for speech stimuli (10 Norwegian participants).

Figure 6: Percentages of “long” responses out of 30 trials for non-speech stimuli (10 Norwegian participants).

Figure 7. Percentages of “long” responses out of 30 trials for speech stimuli (10 Italian participants).

Figure 8. Percentages of “long” responses out of 30 trials for speech stimuli (11 Italian participants).

Figure 9: Percentages of “long” responses out of 30 trials for speech stimuli (17 English participants).

Figure 10: Percentages of “long” responses out of 30 trials for non-speech stimuli (16 English participants).

Figure 11: Japanese listeners’ discrimination of the speech stimuli (10 participants). Mean d'

values with 95% confidence intervals.

Figure 12: Japanese listeners' discrimination of the non-speech analogues (10 participants).

Mean d' values with 95% confidence intervals.

Figure 13: Norwegian listeners' discrimination of the speech stimuli (10 participants). Mean

d' values with 95% confidence intervals.

Figure 14: Norwegian listeners' discrimination of the non-speech analogues (10 participants).

Mean d' values with 95% confidence intervals.

Figure 15: Italian listeners' discrimination of the speech stimuli (10 participants). Mean d'

values with 95% confidence intervals.

Figure 16: Italian listeners' discrimination of the non-speech analogues (10 participants).

Mean d' values with 95% confidence intervals.

Figure 17: English listeners' discrimination of the speech stimuli (18 participants). Mean d'

values with 95% confidence intervals.

Figure 18: English listeners' discrimination of the non-speech stimuli (15 participants). Mean

d' values with 95% confidence intervals.

Figure 19: 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.

Notes.

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.

² Other studies have presented results that could be interpreted as evidence for durational contrast effects, notably Diehl & Walsh (1989, cf Miller & Liberman 1979), but this evidence, too, remains controversial (Fowler 1990, 1991; Diehl, Walsh, & Kluender 1991). We do not have the space to discuss these data and the dispute over their interpretation here.

Brenner-Alsop (2006) presents relevant results and discussion, which are being developed in a forthcoming paper.

³ The independent inverse covariation between vowel and consonant duration induced by the voicing contrast in Japanese, Norwegian, and Italian plays no role in the experiments reported here because the intervals used to simulate stop closures were entirely silent.

⁴ Neither Fowler (1992) nor van Dommelen (1999) adjusted the other intervals in their stimuli to prevent listeners from using total stimulus duration.

⁵ We are very grateful to Curt Rice and Carina Reinholtsen for their hospitality and assistance in conducting the experiment in Tromsø.

⁶ 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.

⁷ Linguistic experience of inverse covariation between vowel and consonant duration presumably prevented Norwegian listeners from adding the durations of the two intervals when responding to speech stimuli in our experiment and van Dommelen's, as well as for Fowler's listeners when they were asked to judge the voicing rather than the duration of the consonant in the speech stimuli.

⁸ 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.

⁹ In English, vowels in word-final syllables are longer than in comparable non-final syllables, e.g. in *tune* compared to *tuna* (Klatt 1973; Turk & Shattuck-Hufnagel 2000, White 2002). Second, long or geminate consonant constrictions only arise across word boundaries, eg in *top pick* (Pickett & Decker 1960) but not *happy* or perhaps even the compound *hip-pocket*. But even if listeners treat a longer vowel in the first syllable as word-final, that does not require them to also treat the following consonant as long.

¹⁰ Correlation and difference did not interact significantly in the analysis of the Italian listeners' discrimination of the speech stimuli, but nonetheless they were slightly better at

discriminating the inversely covarying stimuli than on the directly covarying stimuli for the short-long pair. Perhaps, only in this pair do the closure duration differences in the tokens (singleton vs. geminate) come close enough to resembling those actually existing in Italian, thus making them easier to discriminate because they are representative of the pronunciation of consonant length contrast in Italian. Even if this apparent effect of linguistic experience is real, one still has to explain why it was not obtained from Norwegian listeners, who have very similar experience.

¹¹ Both vowel and silence duration are expressed in ms in the modeling, so the sizes of the corresponding beta values can be compared directly. The interaction of vowel and silence duration is expressed as the product of their individual durations, so it's not surprising that the corresponding beta values are several orders of magnitude smaller.

¹² These findings also rule out an explanation of our results in which listeners rely more on vowel duration when silence duration is ambiguous. This would produce a pattern of responding that looks just like that which addition would produce because they would be more likely to give a "long" response when the preceding vowel is longer. Such a shift in which interval listeners rely on to respond 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. However, that would imply that all the stimuli between the endpoints are equally ambiguous, and a look back at the identification functions in Figures 3-10 shows that this is not so.

Language	Vowel	Consonant	Consonant:Vowel	
	Singleton:Geminate	Geminate:Singleton	Singleton	Geminate
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	1.32-1.36	1.21-1.31	0.46-0.66	0.83-1.04

Table 1. Duration ratios for vowels preceding singleton vs geminate consonants, for geminate vs singleton consonants, and for consonants to vowels when the consonant is singleton or geminate. For English, the values are for voiced and voiceless stops, rather than singletons and geminates, 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).

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

Table 2: Durations of the constituent intervals in the stimuli.

- [h]: Square wave band-pass filtered between 1000Hz and 3000Hz, i.e. roughly the range of F2 and F3 in a vowel. Peak intensity at 0.1 of maximum.
- [a]: Anharmonic complex (component frequencies separated by equal log steps)
- [t]: Silence
- [o]: Anharmonic complex (component frequencies separated by equal log steps)
- [d]: Square wave band-pass filtered between 50 Hz and 150 Hz. Peak intensity at 0.1 of maximum.
- [e]: Anharmonic complex (component frequencies separated by equal log steps)
- [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 cosine window.

Table 3: Characteristics of each portion of the non-speech stimuli.

Difference	Vowel-Consonant Covariation	
	Inverse	Direct
Short-Medium	SM vs. MS	SS vs. MM
Medium-Long	ML vs. LM	MM vs. LL
Short-Long	SL vs. LS	SS vs. LL

Table 4. Stimulus pairs used in the discrimination tasks.

	Identification		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 5: Summary table of the overall results.

language	short+long 48.7+105= 153.7	medium+medium 62.8+90= 152.8	long+short 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 48.7+120= 168.7	medium+medium 62.8+105= 167.8	long+short 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 48.7+135= 183.7	medium+medium 62.8+120= 182.8	long+short 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 6. Percentages of “geminate” 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.

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

Table 7. Two values are entered in each cell in this table. The top values were obtained from a model in which all the listeners data are included, 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 betas for the vowel by silence duration interaction and the independent variables representing silence duration and vowel duration.

Figure

[h a t o d e s u]

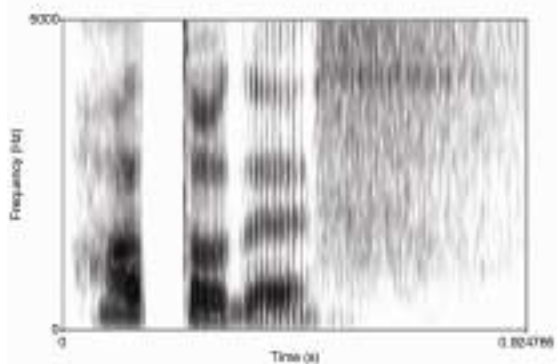


Figure 1: Spectrogram of a speech stimulus: [hatodesu].

[h a t o d e s u]

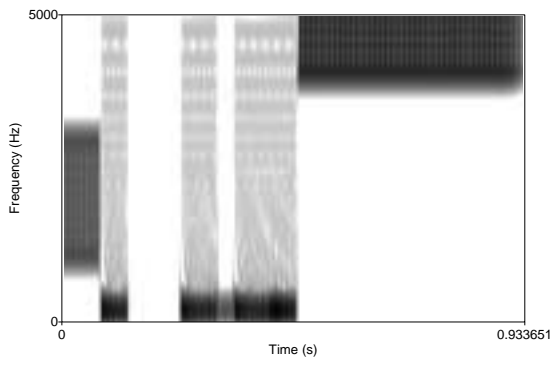


Figure 2: Spectrogram of a non-speech analogue.

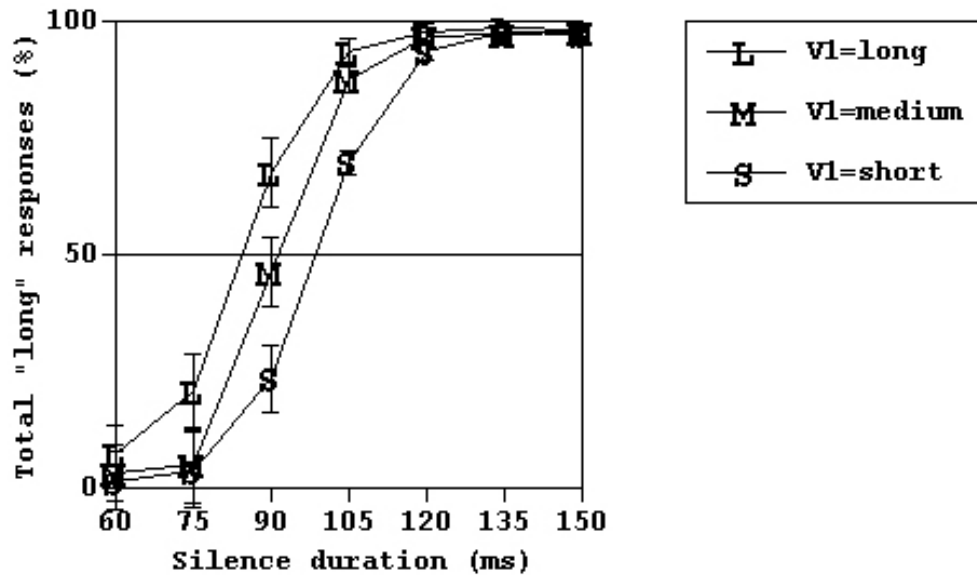


Figure 3: Percentages of “long” responses out of 30 trials/stimulus/listener for speech stimuli (20 Japanese participants). Error bars represent 95% confidence intervals.

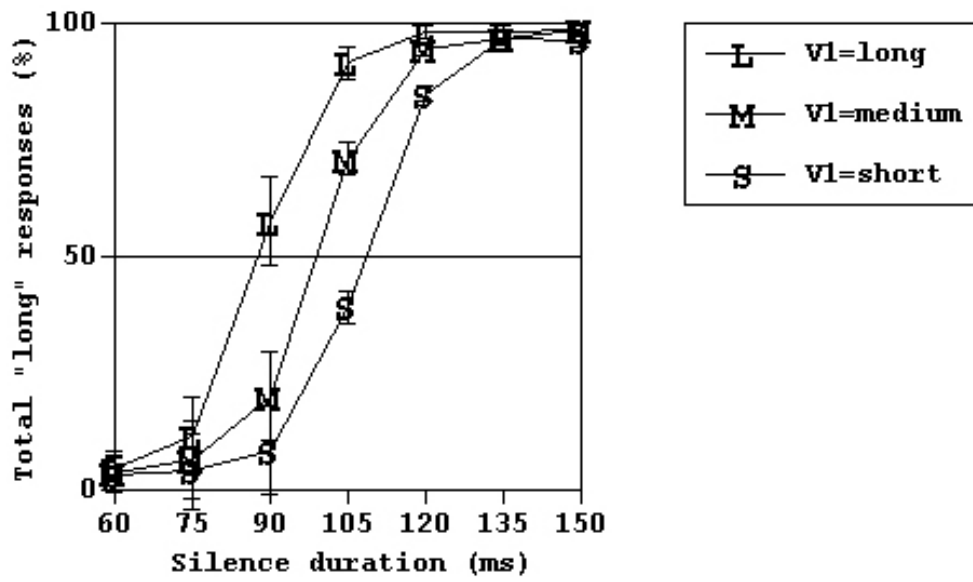


Figure 4: Percentages of “long” responses for non-speech stimuli (20 Japanese participants).

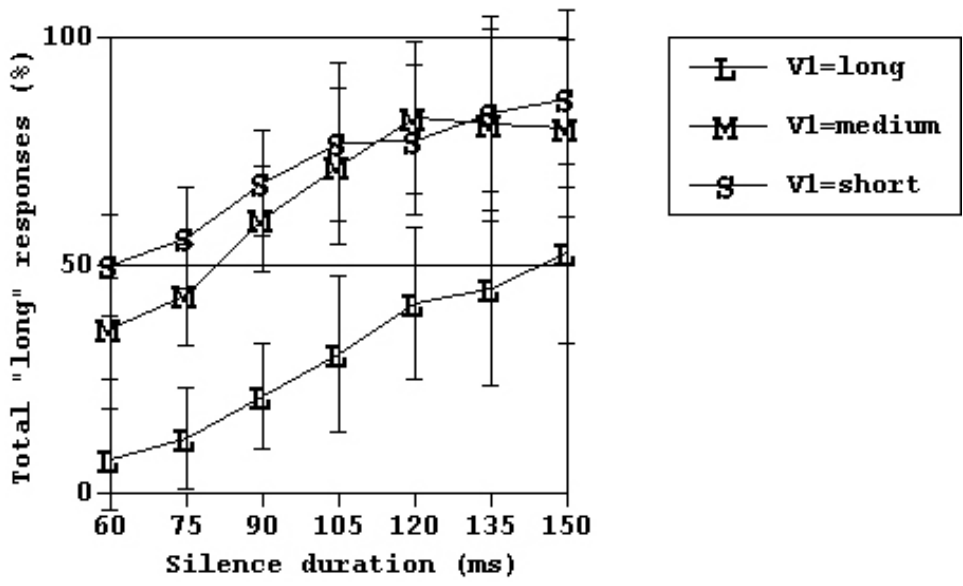


Figure 5: Percentages of geminate responses out of 30 trials for speech stimuli (10 Norwegian participants).

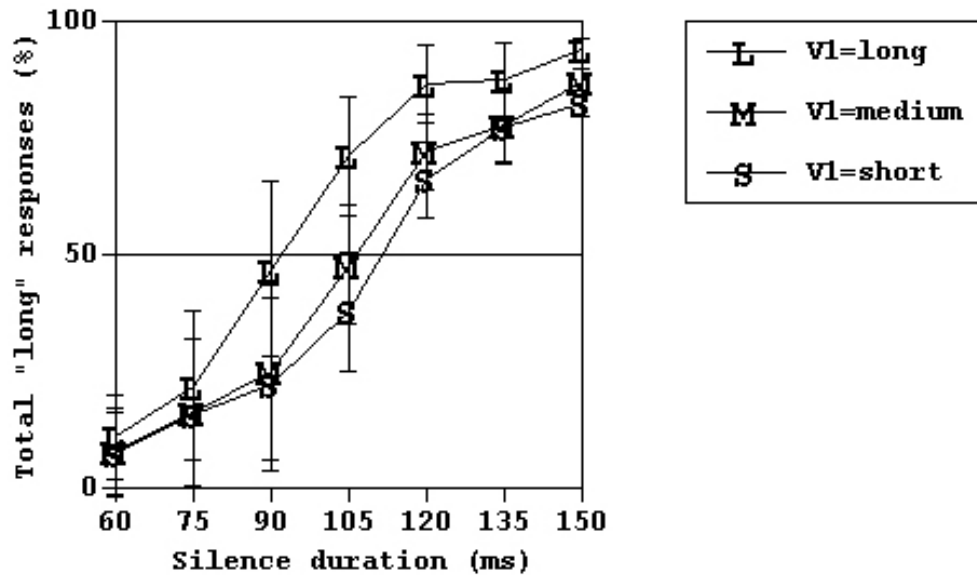


Figure 6: Percentages of “long” responses out of 30 trials for non-speech stimuli (10 Norwegian participants).

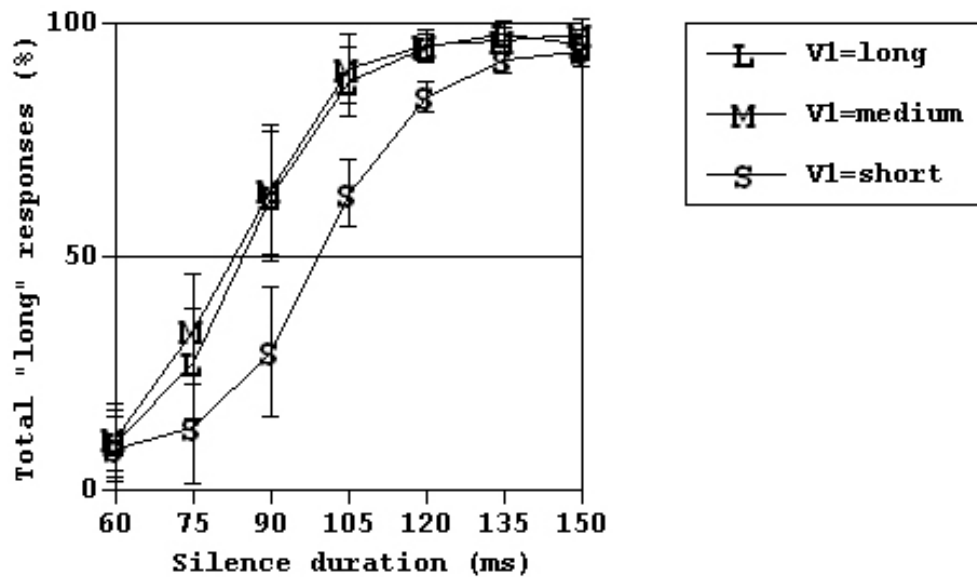


Figure 7. Percentages of “long” responses out of 30 trials for speech stimuli (10 Italian participants).

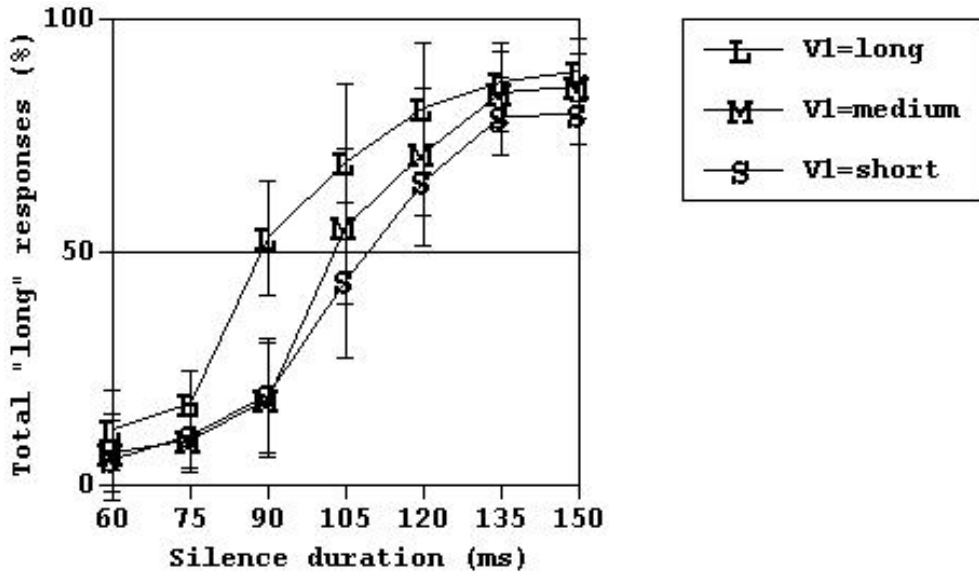


Figure 8. Percentages of “long” responses out of 30 trials for speech stimuli (11 Italian participants).

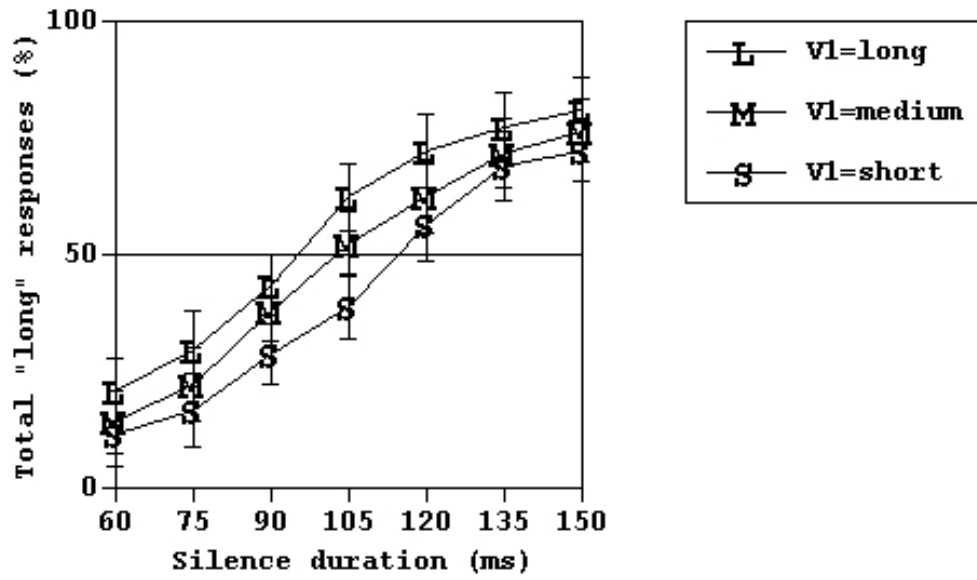


Figure 9: Percentages of “long” responses out of 30 trials for speech stimuli (17 English participants).

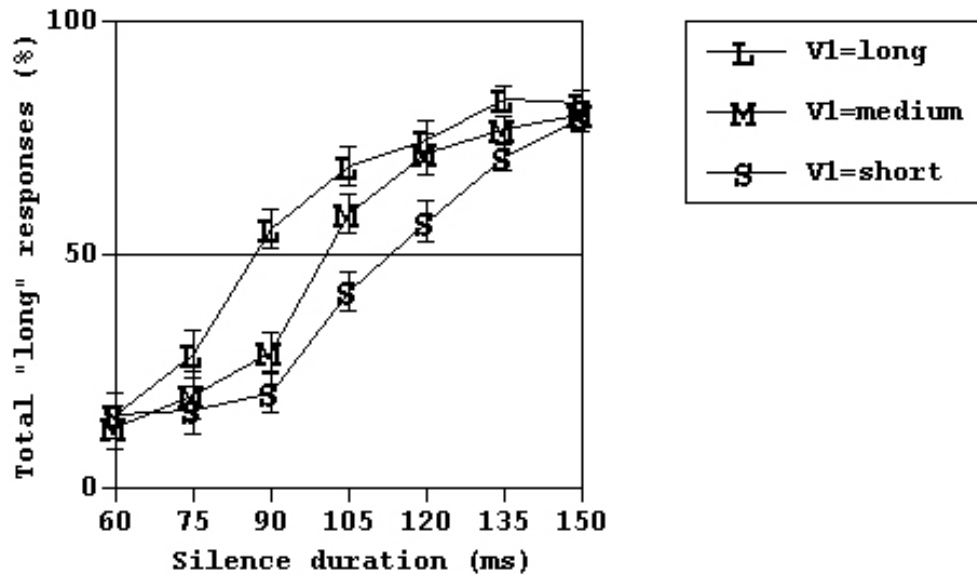


Figure 10: Percentages of “long” responses out of 30 trials for non-speech stimuli (16 English participants).

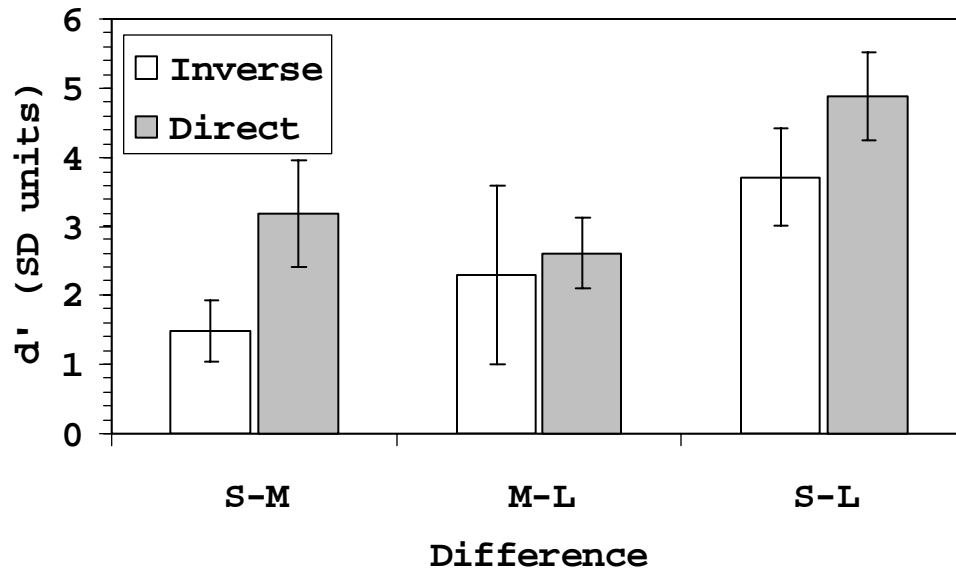


Figure 11: Japanese listeners' discrimination of the speech stimuli (10 participants). Mean d' values with 95% confidence intervals.

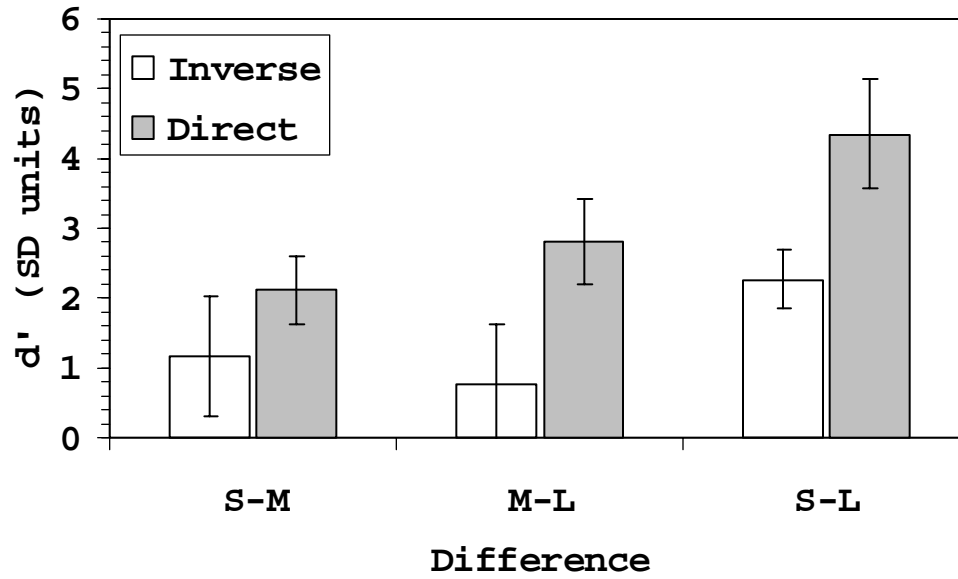


Figure 12: Japanese listeners' discrimination of the non-speech analogues (10 participants). Mean d' values with 95% confidence intervals.

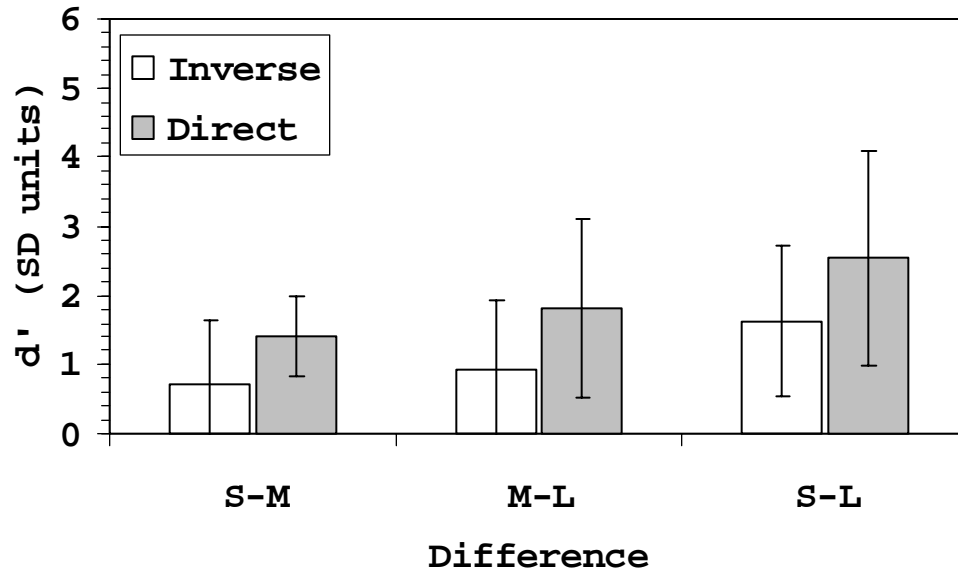


Figure 13: Norwegian listeners' discrimination of the speech stimuli (10 participants).

Mean d' values with 95% confidence intervals.

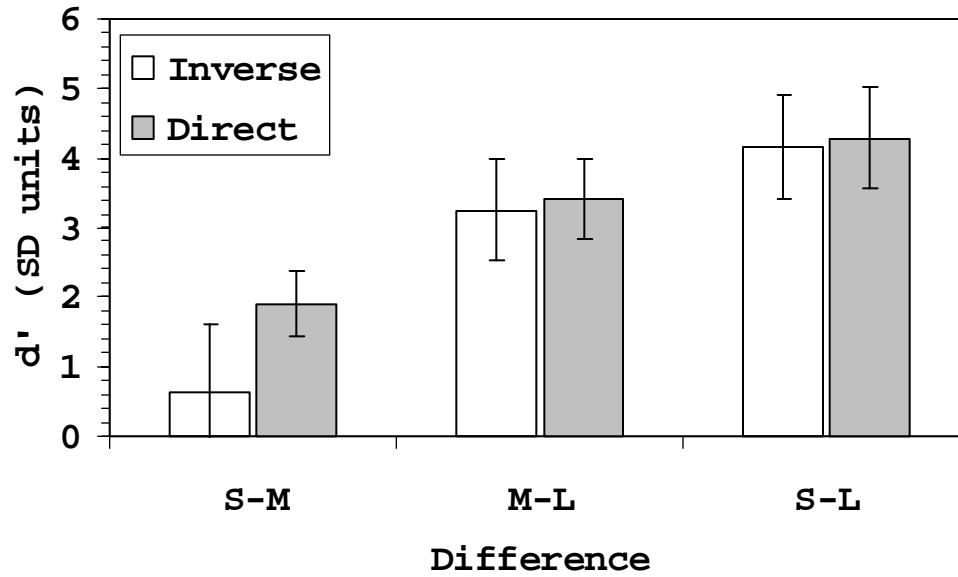


Figure 14: Norwegian listeners' discrimination of the non-speech analogues (10 participants). Mean d' values with 95% confidence intervals.

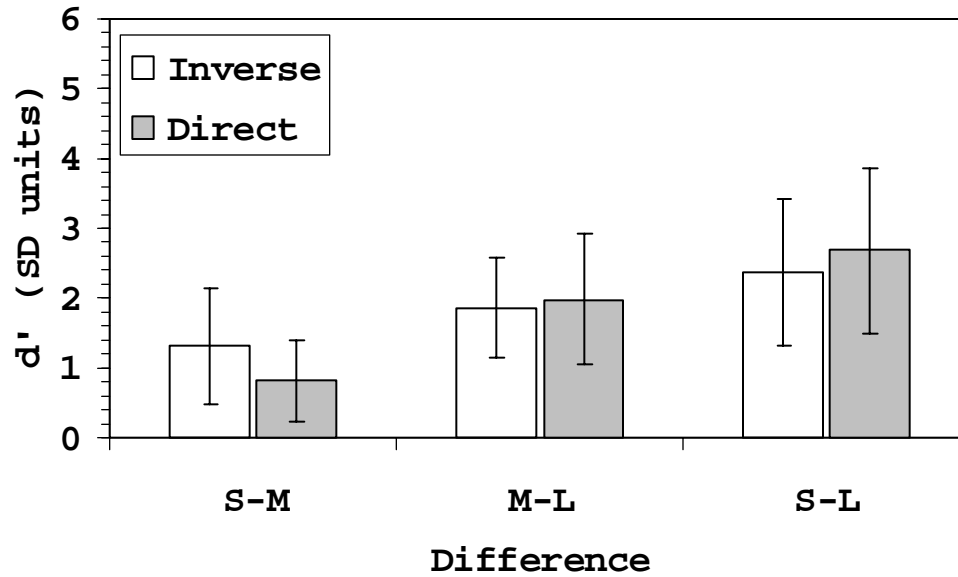


Figure 15: Italian listeners' discrimination of the speech stimuli (10 participants). Mean d' values with 95% confidence intervals.

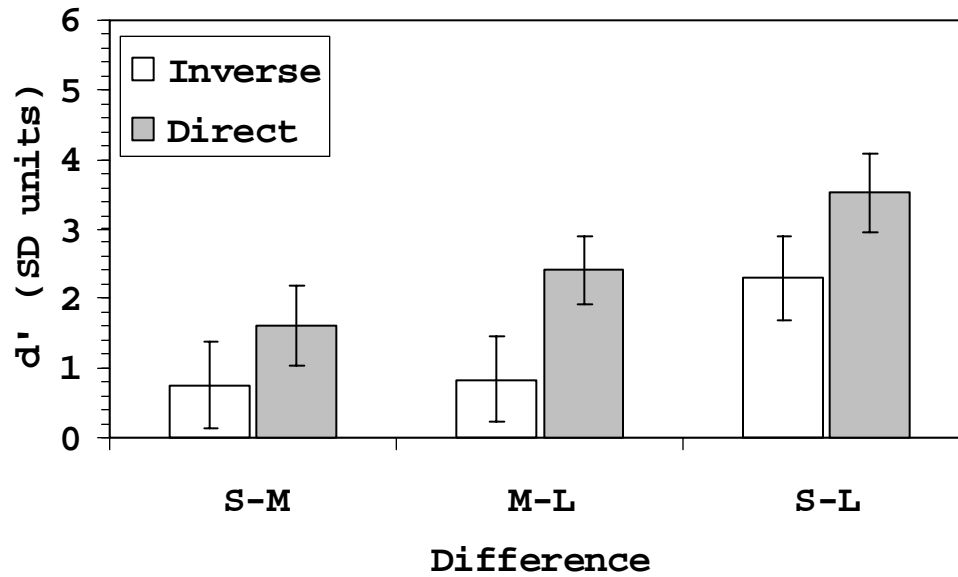


Figure 16: Italian listeners' discrimination of the non-speech analogues (10 participants).

Mean d' values with 95% confidence intervals.

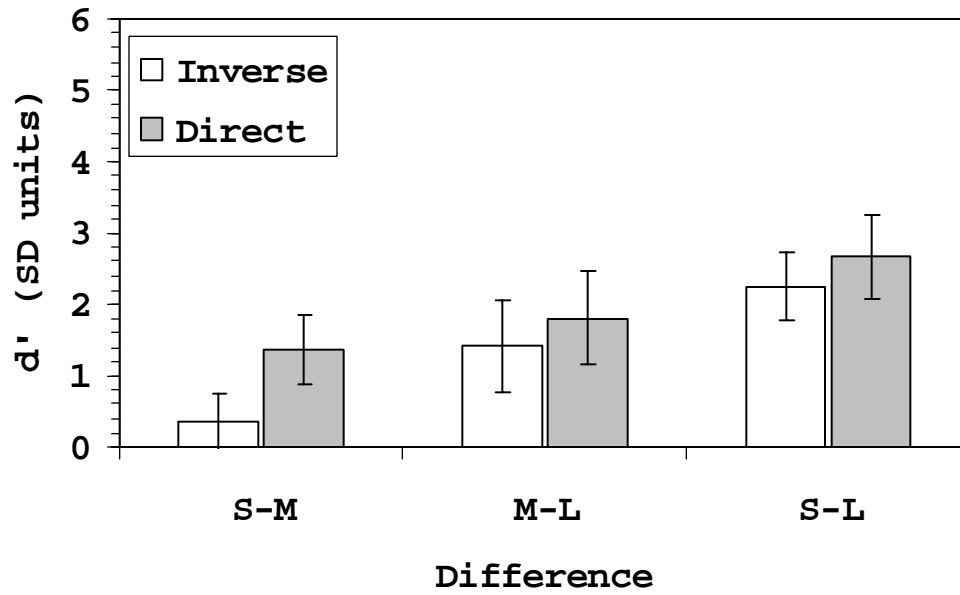


Figure 17: English listeners' discrimination of the speech stimuli (18 participants). Mean d' values with 95% confidence intervals.

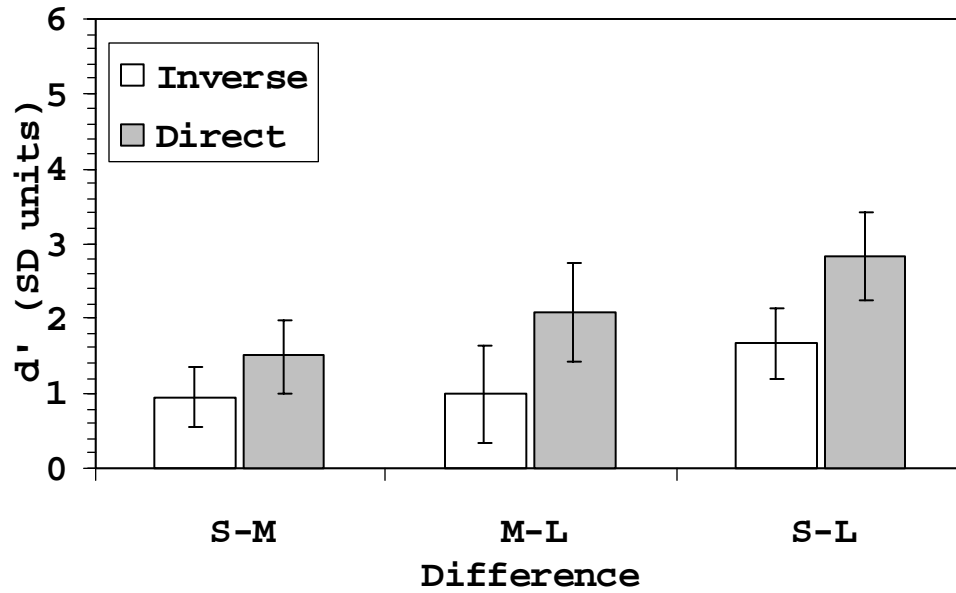


Figure 18: English listeners' discrimination of the non-speech stimuli (15 participants).

Mean d' values with 95% confidence intervals.

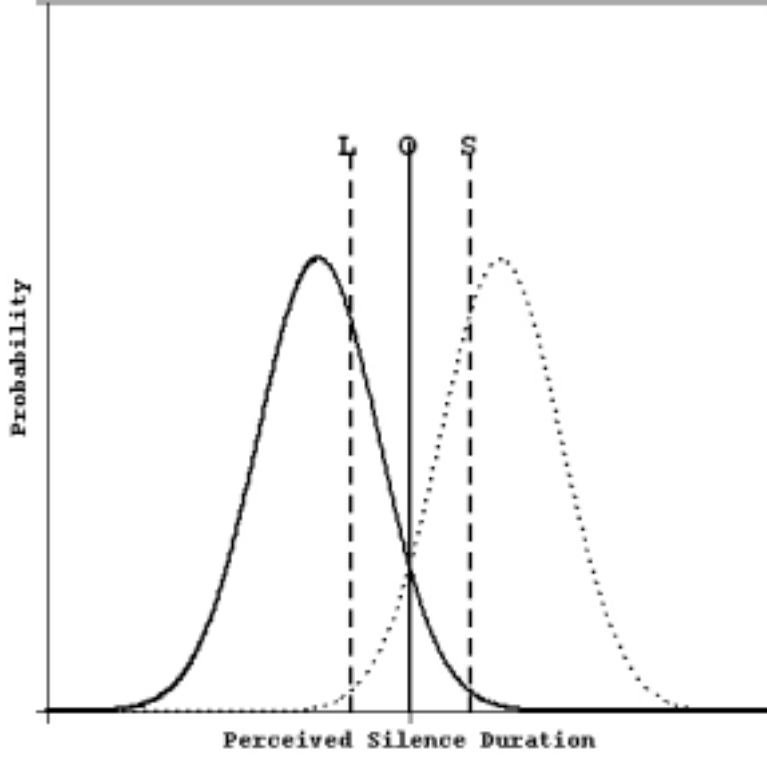


Figure 19: Response distributions corresponding to a shorter and a longer closure duration (solid and dashed lines, respectively) and criteria (decision boundaries) corresponding to a medium (0), short (S), and long (L) duration of the preceding vowel.