

To appear in Charles Cairns & Eric Raimy, eds. *Handbook of the Syllable*. E.J. Brill.

## Chapter 12

### SYLLABLES IN SPEECH PROCESSING – EVIDENCE FROM PERCEPTUAL EPENTHESIS

#### 1. Syllables in speech production and processing

Much of phonological theory assumes that speech is both discrete and hierarchical. The speech stream can be divided into discrete units (segments/sounds), and these units are hierarchically organized into larger chunks such as syllables, feet, prosodic words, etc. (Nespor and Vogel 1986; Selkirk 1978, 1986, etc.). Phonological rules are often formalized in terms of these higher order groupings of sounds. This approach presents many problems to phonological theory. Not only is it difficult to define these higher order categories accurately, but it is probably even harder to find evidence of their reality.

Speech is transferred between speaker and listener as continuously varying air pressure fluctuations. The structure of the pressure fluctuation is sometimes such that discontinuities can be detected easily—for instance at the transition between a plosive and a vowel. This lends some support to the idea that speech is discrete—the speech stream can be divided into discrete chunks based on its physical properties. However, the boundaries between segments are not always demarcated clearly in the physical speech stream, as anyone who has ever tried to find the boundary between a vowel and glide on a spectrogram or waveform can attest. The fact that it is not always possible to find correlates to concepts like segments in the physical world has led some to doubt the existence of discrete segments (Port and Leary 2005).

Just as difficult as finding physical evidence for the division of the speech stream into discrete segments, is finding physical evidence for the grouping of these segments into larger

prosodically defined units such as feet, syllables, etc. Although there are temporal and intonational properties of the speech stream that have been shown to correlate with the boundaries of some of these larger prosodic units (Beckman and Edwards 1990; Byrd and Saltzman 2003; Coetzee and Wissing 2007; Selkirk 1981, 2001; Sugahara 2005; Turk and Shattuck-Hufnagel 2000, etc.), this correlation is not very strong. The observed variation in these temporal and intonational measures is often large, and the differences between different prosodic units are often very small (McQueen 1998, 21).

Given the difficulty of finding consistent physical correlates for these higher prosodic units, a question that needs answering is whether (all of) these units are real. If physical evidence of these units is lacking, is there evidence that language users parse the speech stream in terms of these units? If a positive answer can be given to this question, then it can be concluded that some of these prosodic units are at least psychologically real—that is, even if they do not exist in the physical properties of the speech stream, the mind of the language user imposes these structures onto the speech stream in the process of planning speech production and perceiving speech. These structures may then have mental reality even if they lack physical reality.

In this paper, I provide evidence that listeners do parse the speech stream into higher order prosodic groupings during the process of speech perception. Specifically, I show that listeners are sensitive to syllabically defined allophonic distributions. In English, aspirated voiceless stops only appear in syllable initial position. When presented with an acoustic signal such as [sp<sup>h</sup>ika] with an unambiguous aspirated [p<sup>h</sup>], listeners impose a percept on this stimulus where [p<sup>h</sup>] appears in syllable initial position. This leaves the [s] prosodically stranded—there is no vowel in the actual acoustic stimulus to which the [s] can affiliate. I show that the listeners “perceive” a vowel between the [s] and [p<sup>h</sup>] in this kind of situation, i.e. they perceive [səp<sup>h</sup>ika]. The [ə] that

forms part of the percept is not present in the physical stimulus, and is therefore imposed onto the percept by the perceptual system of the listener. I interpret this as evidence that the perceptual system does impose higher order structure onto the speech stream, and that this part of the speech processing process is even robust enough to result in percepts for which there is no physical evidence.

Before getting into the details of the current study, I briefly review some existing evidence for the mental reality of such higher order prosodic structures in both production and perception. It is not my goal to review the extensive literature about this, but rather to discuss just a few examples of the evidence that language users do use syllables both in production and perception. Similar evidence for other prosodic constituents also exists (Clifton et al. 2006; Schafer et al. 2000; Wightman et al. 1992, etc.).

### 1.1. Evidence for syllables in speech production

One of the arguments for the use of syllables in speech production comes from studies of speech errors. In a classic paper, Fromkin (1971, 38-40) shows that naturally occurring speech errors point to the syllable as a planning unit for speech production. Exchange errors nearly always respect the syllabic affiliation of the segments involved. Onsets exchange with onsets ('carp-si-hord' for 'harp-si-cord') and rhymes with rhymes ('hunk of jeap' for 'heap of junk'). Very rarely observed are errors where the onset of one syllable exchanges with the coda of another syllable. Fromkin claims that her corpus of errors contains only one such example ('whip-ser' for 'whis-per'). This suggests that the exchanges are not mere segment exchanges, but exchanges of segments in the same syllabic position, implying that the segments must be parsed into syllables at the point in the production planning when the exchange occurs. Similarly, Fromkin notes that speech errors that omit complete syllables ('tre-men-ly' for 'tre-men-dous-ly') are rather

common, while errors that omit a segment span that includes parts of two syllables are virtually non-existent (‘tre-men-dy’ for ‘tre-men-dous-ly’).

Speech production models, such as those of Levelt (2001), also often contain a stage where the phonological information is encoded into syllables. Levelt and his colleagues have conducted several experiments over the past two plus decades that show evidence for this syllabic stage of speech planning. I discuss the study of Cholin et al. (2004) as a recent example. They presented Dutch speakers visually with groups of four morphologically related words. There were two kinds of such groups: ‘constant’, where the first syllable of each word was always the same, and ‘variable’, where the initial syllable in one word was different. In both conditions, however, all four words had the same segmental make-up up to the affixes. The difference between the two lists is thus purely in terms of the syllable structure of one member. Examples are given in (1).

(1)	<i>Constant</i>		<i>Variable</i>	
	Meaning	‘lead’		‘smoke’
	Infinitive	[leɪ.dən] <i>leiden</i>		[ro:.kən] <i>roken</i>
	Gerund	[leɪ.dənt] <i>leidend</i>		[ro:.kənt] <i>rokend</i>
	Noun	[leɪ.dər] <i>leider</i>		[ro:.kər] <i>roker</i>
	Past	[leɪ.də] <i>leidde</i>		[ro:k.tə] <i>rookte</i>

Participants had to read each list out loud as accurately and fast as possible. Cholin et al. found the response latency to be significantly greater in the variable than the constant condition. They interpret this as evidence that the participants perform speech planning in terms of syllables. In the constant condition, only one initial syllable has to be planned while two have to be planned in the variable condition, accounting for the longer latency in the latter condition.

## 1.2. Evidence for syllables in speech perception

There is just as large a body of evidence that syllables play a role in speech perception, and I discuss only a few examples from the recent literature. Mattys and Melhorn (2005) conducted a dichotic listening experiment in which they presented participants simultaneously with two different disyllabic auditory stimuli in the left and right ear, respectively. Neither of the two stimuli were actual English words. However, exchanging corresponding parts of the left and right stimuli did create a word. In one condition, the parts that need to be exchanged to form a word corresponded to full syllables. For example, participants could be presented with [kir.fin] in the left ear and [dɔl.mæɪ] in the right ear, where exchanging the initial syllables results in the real word ‘dolphin’. In another condition, the parts that need to be exchanged consisted only of the nucleus of a syllable. Participants could be presented with [dil.fin] in the left and [kɔr.mæɪ] in the right ear. Exchanging the vowel of the first syllables now creates the word ‘dolphin’. In this experiment, participants would first hear a production of the actual word [dɔl.fin] played simultaneously in both ears. After that, they would hear one of the two dichotic stimuli described just above. Their task was to decide whether the second presentation contained the word they just heard. Mattys and Melhorn found that participants were significantly more likely to detect the word in the dichotic stimulus when full syllables have to be exchanged than when only syllabic nuclei have to be exchanged—i.e. more YES responses in [kir.fin]~[dɔl.mæɪ] than in [dil.fin]~[kɔr.mæɪ]. They interpret this as evidence that the acoustic signal from each ear is mapped onto a syllable sized representation before perceptual blending of the two stimuli happens. The blending is then performed by manipulating these syllable sized chunks, with the result that units smaller than a syllable are less accessible to the blending operation.

Another recent study that points to a role for the syllable in speech processing is the perceptual epenthesis study reported by Kabak and Idsardi (2007)—see also Berent et al. (2007), Dupoux et al. (1999, 2001) and Lennertz and Berent (2007). This study is of particular interest, since the experiments that I report later also investigate the phenomenon of perceptual epenthesis. Kabak and Idsardi tested the perception by Korean listeners of consonant sequences that are disallowed in Korean. They use two kinds of disallowed sequences. The first kind violates restrictions on the syllable structure of Korean. Korean allows only a small set of consonants in coda position—e.g. [k] is a possible coda while [tʃ] is not. The token [p<sup>h</sup>ak<sup>h</sup>t<sup>h</sup>a] is hence a possible Korean word, while [p<sup>h</sup>atʃ<sup>h</sup>a] is not. The second kind of disallowed sequence violates not restrictions on syllable structure, but rather restrictions on consonants that are allowed to follow each other. Although Korean allows [k] in coda position, it does not tolerate a sequence of an oral consonant followed by a nasal consonant. Like [p<sup>h</sup>atʃ<sup>h</sup>a], [p<sup>h</sup>ak<sup>h</sup>ma] is hence also not a possible word. However, the reason for the impossibility of these two tokens is quite different. [p<sup>h</sup>atʃ<sup>h</sup>a] is impossible because it has a disallowed coda [tʃ]. On the other hand, in [p<sup>h</sup>ak<sup>h</sup>ma], no single consonant appears in a syllabic position where it is not allowed. This form is impossible because of the oral plus nasal sequence.

Dupoux et al. (1999, 2001) demonstrated that Japanese listeners perform perceptual epenthesis when presented with stimuli that contain disallowed consonantal sequences—i.e. they perceive a vowel between the two disallowed consonants even if there is no actual vowel present. When presented with an acoustic stimulus like [ebzo], Japanese listeners perceive [ebuzo], since the sequence [bz] is disallowed in Japanese. Based on this, Kabak and Idsardi hypothesize that Korean listeners should do the same when presented with disallowed forms like [p<sup>h</sup>atʃ<sup>h</sup>a] and [p<sup>h</sup>ak<sup>h</sup>ma]. They perform a discrimination experiment where participants are presented with pairs

of tokens, and required to judge whether the two members in a pair were identical or not. The crucial pairs consisted of one of the disallowed forms, and a corresponding form with a vowel intervening between the two consonants—i.e. [p<sup>h</sup>akma]~[p<sup>h</sup>akɔma] and [p<sup>h</sup>atʃ<sup>h</sup>a]~[p<sup>h</sup>atʃit<sup>h</sup>a]. If Korean listeners perform perceptual epenthesis, they should not be able to distinguish the members of these pairs. Kabak and Idsardi found significantly better discrimination on pairs like [p<sup>h</sup>akma]~[p<sup>h</sup>akɔma] than on pairs like [p<sup>h</sup>act<sup>h</sup>a]~[p<sup>h</sup>atʃit<sup>h</sup>a]. Perceptual epenthesis was therefore significantly more likely to happen when a token was ill-formed for syllable structure reasons than if it was ill-formed for reasons of consonant sequencing. Kabak and Idsardi conclude that this gives evidence that the acoustic signal is processed in syllable-sized chunks during perception.

## 2. The use of allophonic distributions to determine syllable structure

In this paper, I follow in the footsteps of earlier research on perceptual epenthesis by showing that listeners perceive illusory vowels in order to arrive at a final percept that is syllabically well-formed. However, I also deviate from this research tradition in two ways. First, earlier research use acoustic stimuli that are phonemically ill-formed. Berent et al. (2007), for instance, present English listeners with stimuli like [lbɪf, bɪf, bɪf], and find that these listeners perceive the stimuli with an intrusive vowel—i.e. as [lɛbɪf, bɛɪf, bɛɪf]. A stimulus like [lbɪf] cannot be brought into agreement with English phonotactic grammar simply by mapping [l] onto some other allophone of /l/, [b] onto some other allophone of /b/, or even by a combination of these two options. No allophone of /l/ can appear in word-initial position followed by any other consonant in English. These earlier studies therefore do not answer the question of at which level of perceptual abstraction the acoustic stimulus is parsed into syllables (but see Lennertz and Berent 2007). Is each of the individual segments first mapped onto its corresponding phoneme,

and then these phonemic percepts are syllabically parsed—i.e. allophonic variation is factored out before syllabic parsing? Or is the syllabic parse performed on a percept that still contains allophonic information? In order to address this question, I use stimuli that are allophonically ill-formed but phonemically well-formed. Specifically, I use stimuli like [sp<sup>h</sup>ika]. In English, the aspirated allophone [p<sup>h</sup>] of the phoneme /p/ is not allowed in onset position after an [s]. But another allophone of /p/ is allowed in this position, namely [p]. If listeners first map the segments in an acoustic signal onto phonemes and then parse the phonemes into syllables, then [sp<sup>h</sup>ika] would have been transformed to /spika/ by the point that the syllabic parse is performed. The percept /spika/ can be parsed into well-formed English syllables. On the other hand, if the syllabic parse is performed on a percept that still contains allophonic information, then [sp<sup>h</sup>ika] has to be syllabified, and no well-formed syllabification of such a representation is possible. If listeners perform perceptual epenthesis when presented with a token like [sp<sup>h</sup>ika], it would be evidence that syllabification happens before allophonic information is abstracted away.

The second way in which this study differs from previous studies is on a methodological detail. In order to show that the vowels perceived by listeners are actually because of perceptual epenthesis and not just an artifact of some unforeseen acoustic property of the stimulus, it is necessary to have some control condition. In earlier studies, the control was always provided by listeners from a different language in which the stimuli were well-formed. Kabak and Idsardi (2007) use English listeners, since English allows both [k] and [tʃ] in coda position. They show that English listeners do not perceive an epenthetic vowel in stimuli like [p<sup>h</sup>atʃ<sup>h</sup>a]. The fact that the Korean listeners in their experiment do perceive such a vowel therefore cannot originate in the acoustic properties of their stimuli, but must arise during grammar mediated perceptual processing. Dupoux et al. (1999) similarly use French listeners as control (since stimuli like

[ebzo] are well-formed in French), and Berent et al. (2007) use Russian controls (since [bd, lb, bn] are possible onset clusters in Russian). In the study that I report here, I use the exact same listeners both in the perceptual epenthesis condition and in the control condition. This is possible because I use stimuli that are ill-formed in English simply because of the context in which they appear. The exact same stimuli are well-formed in a different acoustic context—that is, no well-formed syllabic parse of [sp<sup>h</sup>ika] is possible, but splicing [la-] onto this exact same stimulus gives [lasp<sup>h</sup>ika] for which a well-formed syllabic parse is possible. If English listeners perceive a vowel between [s] and [p<sup>h</sup>] in [sp<sup>h</sup>ika] but not in [lasp<sup>h</sup>ika], then we can conclude that the vowel percept in [sp<sup>h</sup>ika] does not originate in some unforeseen acoustic properties of the [s] and [p<sup>h</sup>]. This methodological difference between the study reported here and earlier studies is important for two reasons. First, it is logistically easier in that listeners of only one language need to be recruited. Secondly, by using the same listeners in both the experimental and the control conditions the effect of perceptual epenthesis can be tested within individuals rather than across different listeners with different native languages.

As already mentioned in the previous paragraph, I exploit the distribution of aspirated stops in English in the design of the experiments reported later in this paper. To the best of my knowledge, aspirated stops are only observed in absolute syllable initial position in all dialects of English. It is not the case that all voiceless stops in syllable initial position are aspirated—in American English, inter-vocalic /t/ is often flapped in the onset of an unstressed syllable, and in some British dialects such /t/'s are often glottalized. But what is true is that all aspirated stops appear in syllable initial position. If listeners pay attention to allophonic details such as aspiration, and if they use this allophonic information when they parse the acoustic signal into

syllables, then they should insert a syllable boundary before every aspirated stop in the syllabic parse imposed on the stimulus.

In the design of the experiments reported below, I assume that perceptual processing consists of at least two stages. During the first stage of *acoustic encoding* a faithful acoustic copy of the percept is created. No higher order prosodic structure, such as syllable structure, is present in the acoustic representation formed during this stage, implying that the illusory epenthetic vowels are also still absent at this stage of processing. In the second stage, *phonological interpretation*, the listener parses the acoustic representation into higher order prosodic units, and tests these structures for well-formedness against the grammar of his/her language. If the prosodic parse of the acoustic representation is well-formed, the final linguistic percept is equal to this parse. However, if the prosodic parse of the acoustic representation is ill-formed, the perceptual system performs an extra step in which the representation is altered in some fashion so that it is brought into agreement with the requirements of the grammar of the listener. This is the step during which, for instance, perceptual epenthesis occurs. The listener then settles on a linguistic percept that differs from the original acoustic representation created. This model is represented in Figure 1. See Kingston (2005) and Poeppel et al. (2008) for support of such a multi stage model.

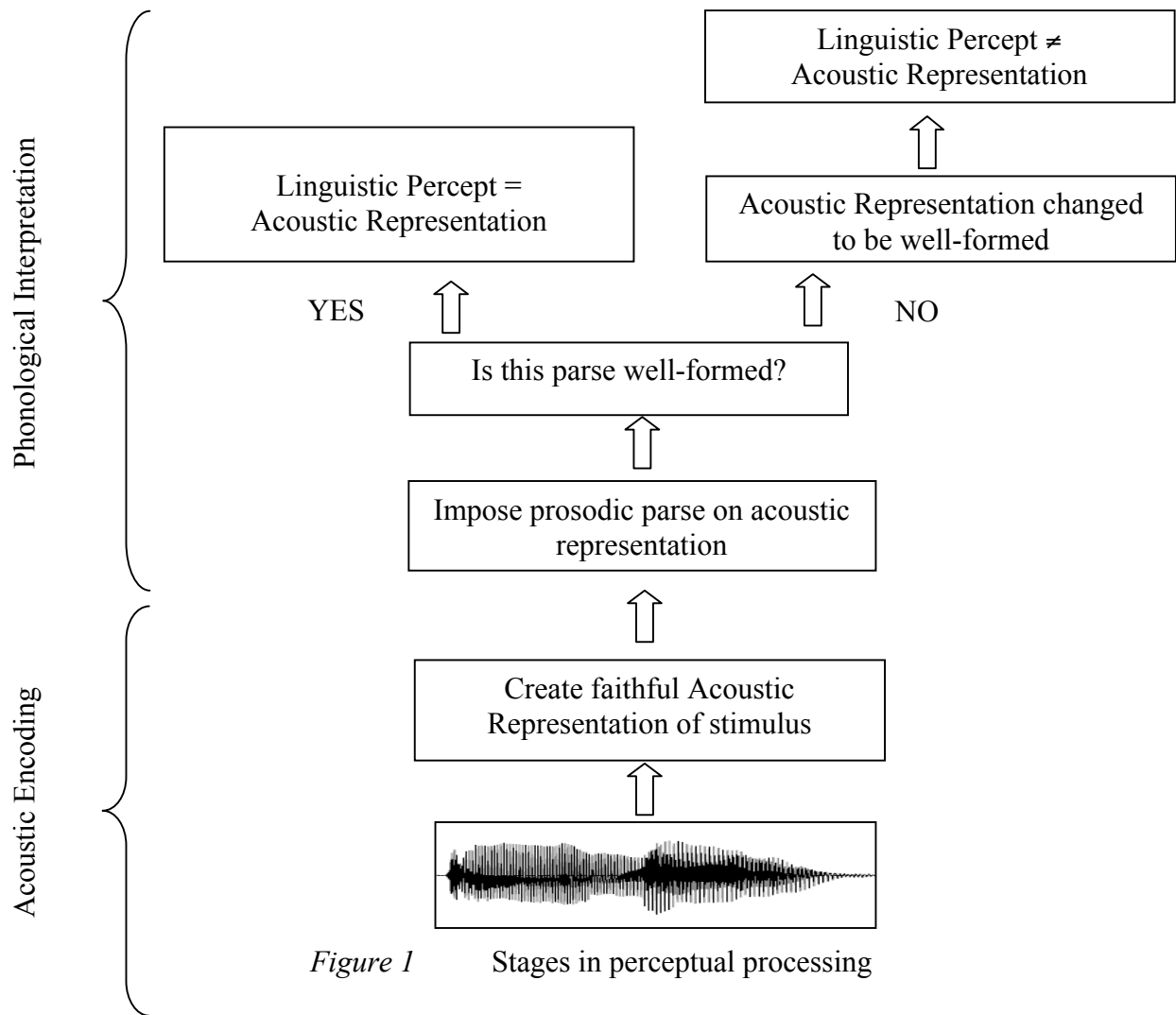


Figure 1 Stages in perceptual processing

Some of the implications of this model are tested in experiments discussed in the rest of this paper. Specifically: (i) In this model, prosodic parsing is performed on the acoustic representation—i.e. before allophonic information is abstracted away. Allophonic information will therefore influence the prosodic structure imposed on the stimulus, so that a token like [sp<sup>h</sup>ika] will receive a different prosodic parse from a token like [spika]. (ii) If a well-formed prosodic parse exists that is faithful to the actual acoustic properties of the stimulus, then this is the percept on which the listener will settle – [spika] will be perceived as [spi.ka]. On the other

hand, if no such a parse exists, the acoustic representation will be altered in some way to create a representation that can be parsed into a well-formed prosodic structure – [sp<sup>h</sup>ika] will be perceived [sə.p<sup>h</sup>i.ka]. (iii) Since the perception of a stimulus like [sp<sup>h</sup>ika] involves an extra processing step, listeners should be slower in arriving at a perceptual decision to this kind of stimulus. (iv) During the early stages of processing (acoustic encoding) stimuli like [səp<sup>h</sup>ika] and [sp<sup>h</sup>ika] will be represented differently, and this difference will only disappear in later processing stages (phonological interpretation). In the next three sections, I discuss three perception experiments designed to test these hypotheses.

### 3. Experiment 1: Same or Different?

In this experiment, participants are presented with pairs of stimuli, and their task is to decide whether the members of a pair is the same or a different “word”. This task taps into the later stages of perceptual processing. Not only do listeners have to listen to at least the first member of a stimulus pair in full, but the task requires of participants to treat the stimuli as “words”, hence increasing the likelihood that participants will rely on grammatical processing of the stimulus. The expectation is hence that the participants’ responses will reflect the output of phonological processing in the model represented in Figure 1. If the hypotheses stated above are correct, participants should have difficulty distinguishing the members of pairs such as [səp<sup>h</sup>ika]~[sp<sup>h</sup>ika]. Because no well-formed syllabic parse exists for the second member of this pair, participants are expected to perform perceptual epenthesis on this stimulus and therefore to perceive the stimulus as [sp<sup>h</sup>ika]—i.e. identical to the other member of the pair. On the other hand, when presented with a pair of stimuli that are identical to [səp<sup>h</sup>ika]~[sp<sup>h</sup>ika] except that each member is preceded by [la-], participants should more easily distinguish the members of the pair. Well-formed

syllabic parses are possible both for [lasəp<sup>h</sup>ika] and for [lasp<sup>h</sup>ika], and the participants should therefore arrive at different perceptual decisions for the members of such a pair.

### 3.1. Methods

*Participants.* Fifteen undergraduate students from the University of Michigan participated in this experiment. They were all native speakers of American English, reporting no hearing or speaking deficits. Participants were paid for their participation.

*Token selection and stimuli creation.* The stimuli used in this experiment were all based on the non-words in (2). All of these non-words have the structure [CV.sə.C<sup>h</sup>ʷ.CV] where [C<sup>h</sup>] is one of the voiceless aspirated stops of English, and [ʷ] is a stressed vowel. I recorded a native, phonetically trained, female speaker of American English reading each of these tokens ten times. From the ten recordings, I selected one to use in the creation of the stimuli. I selected as token to use a repetition in which all of the sounds, and specifically the [ə], were clearly articulated. All of the tokens were then equalized in intensity using the *Scale intensity...* function in *Praat* (Boersma and Weenink 2008).

(2)	<i>Labial</i>		<i>Coronal</i>		<i>Dorsal</i>	
	vəsəp <sup>h</sup> imu	masəp <sup>h</sup> áli	kusət <sup>h</sup> íla	lusət <sup>h</sup> ápi	basək <sup>h</sup> ífa	fisək <sup>h</sup> ána
	lasəp <sup>h</sup> íka	visəp <sup>h</sup> áno	fasət <sup>h</sup> ímo	misət <sup>h</sup> áku	masək <sup>h</sup> ílu	pisək <sup>h</sup> ámi

I discuss the further processing of [lusət<sup>h</sup>ápi] here, but each of the other tokens was processed in exactly the same way. All acoustic manipulation was done in *Praat*. From [lusət<sup>h</sup>ápi], I created two additional stimuli by splicing out first only the [ə], and then both the [ə] and the aspiration associated with [t<sup>h</sup>]. From each of these three stimuli (the original [lusət<sup>h</sup>ápi] and the two created

by splicing out [ə] and [ʰ]), I then created yet another stimulus by splicing off the initial [lu]. This gave the six stimuli shown in (3).

(3)	<i>[s] word-medial</i>	<i>[s] initial</i>
[ə] and [ʰ] present	lusət <sup>h</sup> ápi	sət <sup>h</sup> ápi
[ə] spliced out	lust <sup>h</sup> ápi	st <sup>h</sup> ápi
[ə] and [ʰ] both spliced out	lustápi	stápi

These six stimuli were used to create stimuli pairs. Four ‘identical’ pairs were created by matching up each of the four stimuli with [ʰ] with itself. Four ‘different’ pairs were created by matching a stimulus with both [ə] and [ʰ] with stimuli from which [ə], and both [ə] and [ʰ] have been deleted. Two different instantiations of each of the different pairs were created—with the stimuli appearing in each of the two possible orders. Stimuli in a pair were separated from each other by 500 ms of silence. These stimuli pairs were presented to participants in a same/different task. Given the hypotheses set out in the previous section, the expected responses to the different pairs are given in (4)—only one of the two orders between the stimuli is given, but the same response is expected for both orders. When the stop in Token 2 is unaspirated, [s]-initial and [s]-medial pairs are expected to be equally discriminable. Legal syllable parses faithful to the acoustic percept is possible for all four stimuli in these pairs, and it is expected that the listener will arrive at percepts that are acoustically identical to the stimuli. However, when Token 2 is aspirated, it is expected that the [s]-medial pair will be more discriminable than the [s]-initial pair. There is no legal syllabic parse of [st<sup>h</sup>ápi], and the listener is expected to perform perceptual epenthesis, and hence to arrive at the percept [sət<sup>h</sup>ápi] for this token, identical to the expected percept for Token 1.

(4) Expected responses to different pairs

Token 2 [+asp]?	Position of [s]	Token 1	Token 2	Expected response	Reason
No	Medial	lusət <sup>h</sup> ápi	lustápi	Different	Legal syllabic parse possible for both tokens.
	Initial	sət <sup>h</sup> ápi	stápi	Different	Legal syllabic parse possible for both tokens.
Yes	Medial	lusət <sup>h</sup> ápi	lust <sup>h</sup> ápi	Different	Legal syllabic parse possible for both tokens.
	Initial	sət <sup>h</sup> ápi	st <sup>h</sup> ápi	Same	No legal parse for Token 2 possible, perceptual epenthesis expected.

*Procedure.* Stimuli presentation and response collection were controlled with the *SuperLab* software package. Participants were run in groups of up to three. Data collection was done in a sound attenuated room. Participants were seated in front of individual laptop computers, each with a response button box attached. Stimuli were presented via headphones, and responses were collected via the button boxes. In each experimental block, each of the 12 pairs (4 identical, 4 different in two orders) was presented once. Since stimuli pairs were created from all 12 tokens in (2), each block contained 144 stimuli pairs. Before a stimulus pair was presented, a visual cue was presented in the middle of the computer screen. A token pair was then presented, and participants indicated their choice by pushing one of two buttons marked as “same” and “different”. Participants were instructed to respond as quickly and accurately as possible. The block was presented four times, with participants allowed a self paced break between repetitions. Stimuli pairs were differently randomized for each participant in each repetition. Before the first block, participants were given 10 practice trials on stimuli pairs created in the same manner as described above, but from tokens not used in the actual data collection part. Participants received no feedback during the experiment.

### 3.2. Results and discussion

The response patterns were transformed to  $d'$ -scores as in Signal Detection Theory (MacMillan and Creelman 2005) before they were statistically analyzed. Higher  $d'$ -scores correspond to higher discriminability. The  $d'$ -scores were submitted to  $2 \times 2$  ANOVA's, both by participants and by items, with factors aspiration (both members aspirated, one member unaspirated) and [s]-position (word-initial, word-medial). Both main effects were found to be significant: aspiration (by participant  $F(1, 14) = 139.0, p < .001$ ; by item  $F(1, 11) = 575.9, p < .001$ ), and [s]-position (by participant  $F(1, 14) = 15.7, p < .002$ ; by item  $F(1, 11) = 73.4, p < .001$ ). Also the interaction between the factors was significant (by participant  $F(1, 14) = 12.0, p < .005$ ; by item  $F(1, 11) = 12.8, p < .005$ ). The average by participant  $d'$ -scores are represented graphically in Figure 2.

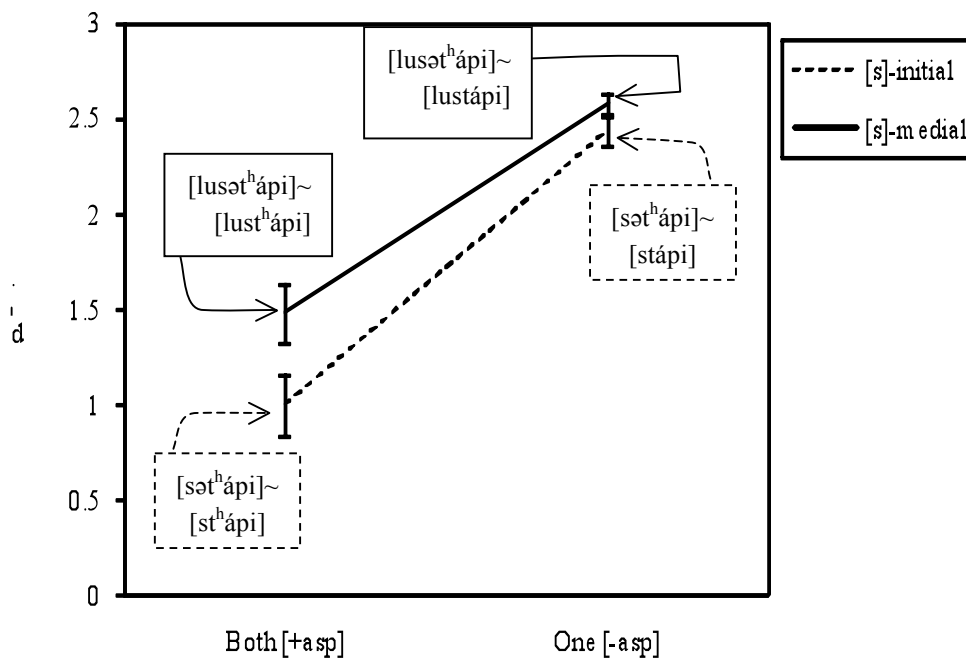


Figure 2 Mean  $d'$ -scores by participant for Experiment 1. Error bars show 95% confidence intervals.

The first thing to note is that the discrimination is better for the pairs where one of the tokens in the pair does not contain aspiration. This is a direct result of the actual size of the physical

acoustic difference between the members of a pair. Discrimination is better when the members differ both in terms of aspiration and the presence/absence of a schwa, than if they differ only in terms of the presence/absence of a schwa. Discrimination is also better for the pairs where [s] appeared in word-medial position than the pairs where [s] appeared in word-initial position. However, as indicated by the significant interaction between factors, and as is also clear from Figure 2, this difference comes primarily from the pairs where both tokens contain an aspirated stop. The d'-scores for [s]-initial and [s]-medial pairs with one unaspirated pair member do not differ significantly from each other, as indicated by the overlapping confidence intervals. On the other hand, d'-scores for pairs where both pair members are aspirated do differ significantly. Specifically, discrimination is significantly better for pairs like [lusət<sup>h</sup>ápi]~[lust<sup>h</sup>ápi] than for pairs like [sət<sup>h</sup>ápi]~[st<sup>h</sup>ápi]. Put differently, having aspiration on both pair members is more detrimental to discriminability in the [s]-initial condition than in the [s]-medial condition.

It is important to remember that the actual size of the acoustic difference between [lusət<sup>h</sup>ápi] and [lust<sup>h</sup>ápi], and between [sət<sup>h</sup>ápi] and [st<sup>h</sup>ápi] is exactly the same. In fact, since all these stimuli were created from the same token, these two pairs are acoustically identical except for the presence/absence of the initial syllable [lu-]. The difference in discriminability between these two pairs therefore cannot originate in the acoustic differences between pair members. The perceptual systems of the participants in this experiment are more likely to map [sət<sup>h</sup>ápi] and [st<sup>h</sup>ápi], than [lusət<sup>h</sup>ápi] and [lust<sup>h</sup>ápi], onto the same percept. I interpret this as evidence that the participants in this experiment tended to perceive the acoustic stimulus [st<sup>h</sup>ápi] with an intrusive schwa between the [s] and [t<sup>h</sup>].

I interpret the results of this experiment as evidence that listeners parse acoustic percepts into prosodic structure, and that these parses are evaluated against the grammar of the listener.

However, there is another possible explanation for these results that should be considered. It is true that in the mental lexicon of an English listener, the sequence [#sət<sup>h</sup>...] has higher probability than the sequence [#st<sup>h</sup>...]. In fact, [#st<sup>h</sup>...] probably has zero probability. It is hence also possible that the effect observed in this experiment follows from such frequency statistics calculated over the lexicon rather than from prosodic parsing and grammatical evaluation. Listeners could posit several perceptual hypotheses that are partially consistent with the acoustic stimulus, and then settle on the percept that has a higher statistical likelihood (Connine et al. 1993; Newman et al. 1997). When presented with a stimulus like [st<sup>h</sup>ápi], the listener could therefore posit both [st<sup>h</sup>ápi] and [sət<sup>h</sup>ápi] as perceptual hypotheses and consult his/her mental lexicon to determine the probability of each of these percepts. Since the probability of [st<sup>h</sup>...] is lower than that of [sət<sup>h</sup>...], the listener then settles on [sət<sup>h</sup>ápi].

There are reasons to doubt this interpretation of the results. In all of the previous studies that investigated the phenomenon of perceptual epenthesis, it has been shown that the results do not originate from such frequency statistics (Berent et al. 2007; Dupoux et al. 2001; Kabak and Idsardi 2007). However, to test this frequency based explanation more explicitly, I conducted a control experiment. I calculated the frequency of the sequences [#sək, #səp, #sət] and [#sik, #sip, #sit] in CELEX (Baayen et al. 1995). The log transformed frequencies are given in (5). The frequencies with the two vowels have the following relative ordering for each of the three places of articulation: (#sik > #sək), but (#səp > #sip) and (#sət > #sit). Under the frequency based account, the prediction is hence that for a stimulus like [sk<sup>h</sup>...] a percept with an intrusive [i] will be more likely than a percept with an intrusive [ə]. However, for [sk<sup>h</sup>...] and [st<sup>h</sup>...], a percept with an intrusive [ə] should be more likely. To test this, I ran a second experiment that was identical to the one described just above with one exception: the recordings from which the

stimuli were created had the structure [CV.si.C<sup>h</sup>v.CV] rather than [CV.sə.C<sup>h</sup>v.CV]. In order to make the two sets of stimuli as comparable as possible, tokens for the control experiment were selected such that the durations of [i] and [ə] did not differ significantly (two-tailed  $t(22) = 1.52$ ,  $p = .14$ ). Stimuli creation, and experimental design and procedure were exactly the same as for Experiment 1. Thirteen native speakers of American English participated in the control experiment. By comparing the response patterns between these experiments, the predictions of the frequency based account can be tested. The expected differences in the response patterns between the experiments under the frequency based account are given in 0.

The d'-scores for the crucial pairs in the two experiments are represented in (7). The d'-scores were higher in the control experiment (when the vowel was [i]) than in Experiment 1 (when the vowel was [ə]) for token pairs at all three places of articulation. This goes against the predictions of the frequency based account given in 0. The fact that discriminability is lower when  $V = [ə]$ , shows that an intrusive vowel percept is most likely to be [ə], irrespective of the frequency of the sequences involved. Given that epenthetic vowels in English production are also more [ə]-like (Davidson 2006, 2007), this is what is expected if this intrusive vowel percept is the result of the grammar of the listeners rather than an influence of frequency statistics calculated over the lexicon.

(5) Log transformed CELEX frequencies

	#sVk	#sVp	#sVt
V = [ə]	2.03	4.22	2.15
V = [i]	3.49	1.81	1.70

(6) Expected response patterns under the frequency based account

Token pair	Expected epenthetic	Expected discriminability
	vowel percept	
sVk <sup>h</sup> ámi sk <sup>h</sup> ámi	[i] (#sik > #sək)	Better when V = [ə]
sVp <sup>h</sup> áno sp <sup>h</sup> áno	[ə] (#səp > #sip)	Better when V = [i]
sVt <sup>h</sup> áki st <sup>h</sup> áki	[ə] (#sət > #sit)	Better when V = [i]

(7) Comparison between the results of Experiment 1 and the control experiment

Token pair	d'-score	
	V = [ə]	V = [i]
sVk <sup>h</sup> ámi sk <sup>h</sup> ámi	1.25	1.82
sVp <sup>h</sup> áno sp <sup>h</sup> áno	0.61	1.54
sVt <sup>h</sup> áki st <sup>h</sup> áki	1.14	1.84

Having shown that a frequency based account is unlikely, there still remains an alternative explanation for the effect that also assumes that the effect has its origins in the syllable structure grammar of English. All that the discrimination experiment shows is that participants were less successful at discriminating pairs like [sət<sup>h</sup>ápi]~[st<sup>h</sup>ápi]. However, based only on the results of this experiment, we do not know whether this is because they perceive [st<sup>h</sup>ápi] with an intrusive schwa, or whether they perform perceptual deletion on [sət<sup>h</sup>ápi] and perceive this stimulus without the schwa. Experiment 2 was designed to differentiate between these two explanations.

In this experiment, participants were presented with tokens such as [sət<sup>h</sup>ápi] and [st<sup>h</sup>ápi] with the task of counting the number of syllables in each token. If they perform perceptual epenthesis, they should count three syllables in [st<sup>h</sup>ápi]. On the other hand, if they perform perceptual deletion, they should count two syllables in [sət<sup>h</sup>ápi].

#### 4. Experiment 2: Syllable count

Berent et al. (2007) use a syllable count task to investigate the phenomenon of perceptual epenthesis. This task has two advantages over the discrimination task. First, as explained in the previous paragraph, it is better at diagnosing what the perceptual system does—perceptual epenthesis or perceptual deletion? Secondly, in a syllable count task, participants hear only one token before they have to respond. The time that elapses between stimulus presentation and response is hence considerably shorter, so that the syllable count task gives better insight into the time course of the processing involved in the perception of these stimuli. However, the syllable count task still taps into the later phonological interpretation stages of speech processing. Participants are asked to count “syllables” which implies that prosodic structure must be present by the time that they respond. The nature of the task also requires of participants to listen to the whole stimulus before a response can begin, thereby also slowing down the response and increasing the likelihood that the response will be given only after phonological interpretation has taken place.

##### 4.1. Methods

*Participants.* Twelve undergraduate students from the University of Michigan participated in this experiment. All participants were native speakers of American English with no known speech or

hearing deficits. There was no overlap in participants between Experiment 1 and 2. Participants were paid for their participation.

*Token selection and stimuli creation.* The stimuli used in this experiment were created from the 12 non-words in (8). These tokens have the form [CV.sə.C<sup>h</sup>VC] where [C<sup>h</sup>] is one of the voiceless aspirated stops of English, and [V] is a stressed vowel. The same female speaker as in Experiment 1 was recorded reading these non-words ten times. I again selected one repetition of each token to use for stimuli creation. As in Experiment 1, the token was selected so that all of the sounds in the token were clearly articulated, and these selected tokens were equalized for intensity in *Praat*.

(8)	<i>Labial</i>		<i>Coronal</i>		<i>Dorsal</i>	
	vəsəp <sup>h</sup> ɪm	masəp <sup>h</sup> ál	kusət <sup>h</sup> ɪf	lusət <sup>h</sup> ám	basək <sup>h</sup> íp	fisək <sup>h</sup> án
	lasəp <sup>h</sup> ɪf	visəp <sup>h</sup> áf	fasət <sup>h</sup> ɪk	misət <sup>h</sup> ál	masək <sup>h</sup> ɪf	pisək <sup>h</sup> áf

Each of these 12 non-words was processed in the same way to create stimuli. I discuss the processing of [fisək<sup>h</sup>án] here as an example. From [fisək<sup>h</sup>án], I created two additional stimuli by splicing out first only the schwa (i.e. [fisk<sup>h</sup>án]), and then both schwa and the aspiration associated with [k<sup>h</sup>] (i.e. [fiskán]). One additional stimulus was created from each of these three stimuli by also deleting the initial [fi-] (i.e. [sək<sup>h</sup>án], [sk<sup>h</sup>án], and [skán]). The other 11 non-words were processed in the same way, creating six stimuli from each. These stimuli were presented to participants in a syllable count task. Based on the hypotheses set out at the end of section 2, the expected response pattern for this experiment is given in (9). The response is expected to agree with the actual syllable count for all but one of the six token types. Well-formed syllable parses exist for all of the stimuli except for [sk<sup>h</sup>án]. In agreement with the

hypothesis that the perceptual system settles on a final percept that is identical to the acoustic input if a well-formed prosodic parse of the acoustic input exists, the expectation is hence that participants will accurately perceive the number of syllables for all but [sk<sup>h</sup>án]. For [sk<sup>h</sup>án], the aspirated [k<sup>h</sup>] has to be parsed into syllable onset position, leaving the word-initial [s] stranded. The perceptual system is then expected to perform perceptual epenthesis to supply [s] with a syllabic nucleus.

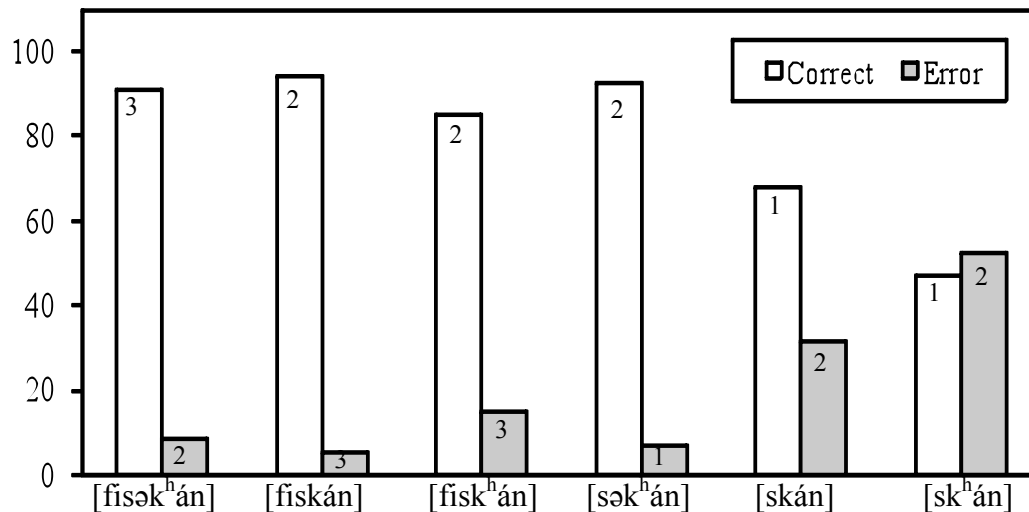
*Procedure.* Stimuli presentation and response collection were performed exactly as for Experiment 1. Six stimuli were created from each of the 12 non-words in (8), for a total of 72 stimuli. To these stimuli, 60 fillers were added. The fillers were created from non-words of the form [CV.C<sub>1</sub>ə.C<sub>2</sub>VC], where [C<sub>2</sub>] is a sonorant consonant, and [C<sub>1</sub>C<sub>2</sub>-] is a possible syllable onset in English. From the non-word [lusəmát], for instance, were created the fillers [lusəmát], [lusmát], [səmát], and [smát]. All test stimuli were presented 8 times each, and fillers 4 times each. Stimuli presentation was differently randomized for each participant. Participants were allowed a self-paced break halfway through the experiment. Before a stimulus was presented, a visual cue was presented in the middle of the computer screen. The stimulus was then presented, and participants indicated their responses by pressing one of three buttons on a response box marked as [1], [2] and [3], respectively. The response time was recorded from the onset of each stimulus up to the point where the participant registered a response. Participants received no feedback during the experiment. Before the experiment began, participants did 10 practice trials, randomly selected from the fillers. Participants were instructed to respond as accurately and quickly as possible.

(9) Expected responses in Experiment 2

Example	[s]-position	Schwa?	[+asp]?	Actual syllable count	Expected response
fɪsək <sup>h</sup> án	Medial	Yes	Yes	3	3
fɪsk <sup>h</sup> án		No	Yes	2	2
fɪskán		No	No	2	2
sək <sup>h</sup> án	Initial	Yes	Yes	2	2
sk <sup>h</sup> án		No	Yes	1	2
skán		No	No	1	1

#### 4.2. Results and discussion

Responses were coded as ‘correct’ or ‘error’ in terms of the actual syllable count of each token, as given in (9). The average response pattern per participant is represented in Figure 3. The percent correct responses on each stimulus type was subjected to a one sample *t*-test to determine whether it was significantly above chance. The results of these *t*-tests are given in (10). Except for the [sk<sup>h</sup>án]-type stimuli, the responses were indeed significantly above chance both by participants and by items. For the [sk<sup>h</sup>án]-type stimuli, neither the participants nor the items analysis returned a significant result. In fact, the average percent correct responses was actually below chance for these stimuli.



*Figure 3* Mean syllable counts for Experiment 2 by participant for each token type. Numbers written in the bars indicate the response represented by the bar. Each category on the x-axis represents all stimuli of that type.

The most informative comparison here is between [fiskʰán] and [skʰán]. These stimuli are acoustically identical except for the presence/absence of the initial [fi-]. Since participants were very accurate in their responses to [fiskʰán]-type stimuli, the extra syllable percept to [skʰán]-type stimuli cannot come from some unforeseen acoustic property of these stimuli. In the [s]-initial condition, participants perceive an extra vowel between [s] and [kʰ], but in the [s]-medial condition they do not perceive such a vowel between the exact same [s] and [kʰ]. This vowel percept therefore originates in the perceptual system and not in the acoustic properties of the stimulus. This result also answers the one question that was left open at the end of Experiment 1. The results of Experiment 1 showed that participants cannot accurately discriminate stimuli such as [sətʰápi] and [stʰápi]. What was not clear was whether this is because they incorrectly perceive [sətʰápi] as [stʰápi], or whether they incorrectly perceive [stʰápi] as [sətʰápi]. The fact that the

participants in Experiment 2 perceived [sk<sup>h</sup>án] with an intrusive vowel, indicates that it is [st<sup>h</sup>ápi] that was misperceived in Experiment 1.

(10) Results of *t*-tests for Experiment 2

<i>Stimulus type</i>	<i>Participant analysis</i>		<i>Item analysis</i>	
	<i>t</i> (11)	<i>p</i>	<i>t</i> (11)	<i>p</i>
[fisæk <sup>h</sup> án]	12.3	< 0.001	36.0	< 0.001
[fiskán]	21.5	< 0.001	88.0	< 0.001
[fisk <sup>h</sup> án]	12.2	< 0.001	6.1	< 0.001
[sæk <sup>h</sup> án]	30.7	< 0.001	45.9	< 0.001
[skán]	2.0	< 0.04	9.2	< 0.001
[sk <sup>h</sup> án]	-0.3	0.78	-0.9	0.37

The response times for Experiment 2 were also statistically analyzed. For each participant, the response times were first standardized. Response times for [s]-initial and [s]-medial stimuli were separately standardized, since the [s]-medial stimuli were all slightly longer because of the extra initial [CV-]. After the response times were thus transformed, all responses that were more than two standard deviations from the mean were excluded. This led to the exclusion of just over 4% of all responses.

The response times on the [fisæk<sup>h</sup>án]- and [fisk<sup>h</sup>án]-type stimuli, and the [sæk<sup>h</sup>án]- and [sk<sup>h</sup>án]-type stimuli were compared, using one-tailed paired-sample *t*-tests. [fisæk<sup>h</sup>án] contains an extra [ə] that is absent from [fisk<sup>h</sup>án]. Since response time was measured from the onset of a stimulus, the extra [ə] therefore contribute a little to the response time measured for [fisæk<sup>h</sup>án],

and similarly for the extra [ə] in [sək<sup>h</sup>án]. If only the actual duration of a stimulus determines the response time, then the response times to [fisək<sup>h</sup>án] and [sək<sup>h</sup>án] should on average be longer than those to [fisk<sup>h</sup>án] and [sk<sup>h</sup>án]. The analyses for [s]-medial tokens like [fisək<sup>h</sup>án] and [fisk<sup>h</sup>án] returned non-significant results both by items ( $t(11) < .01, p = .50$ ) and by participants ( $t(11) = 0.29, p = .39$ ). The response times to these two types of tokens therefore did not differ significantly. However, the results for the [s]-initial tokens like [sək<sup>h</sup>án] and [sk<sup>h</sup>án] were quite different. They were highly significant by items ( $t(11) = 8.3, p < .001$ ) and tended toward significance by participants ( $t(11) = 1.6, p = .07$ ). These results are represented graphically in Figure 4. Inspection of this figure shows that the response time for [fisək<sup>h</sup>án] and [fisk<sup>h</sup>án] is about equal, as is expected from the results of the *t*-tests reported just above. The response time for [sək<sup>h</sup>án] is also shorter than that for [fisək<sup>h</sup>án] and [fisk<sup>h</sup>án], which is again expected since [sək<sup>h</sup>án] is physically shorter, lacking the initial [fi-]. However, the response time for [sk<sup>h</sup>án]-type stimuli is longer than that for [fisək<sup>h</sup>án]- and [fisk<sup>h</sup>án]-type stimuli. This is true in spite of the fact that the [sk<sup>h</sup>án]-type stimuli are in reality the shortest of the four stimuli types compared in this figure.

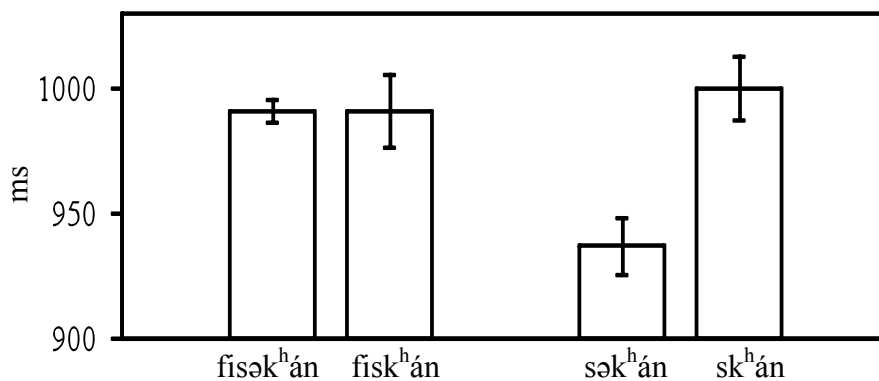


Figure 4 Mean response times by item for [fisək<sup>h</sup>án]-, [fisk<sup>h</sup>án]-, [sək<sup>h</sup>án]-, and [sk<sup>h</sup>án]-type stimuli. Error bars show 95% confidence intervals.

Most importantly, the response time for [sk<sup>h</sup>án]-type stimuli is significantly longer than that for [sək<sup>h</sup>án]-type stimuli, in spite of the fact that the [sək<sup>h</sup>án]-type stimuli are actually longer. The longer response time for [sk<sup>h</sup>án]-type stimuli therefore does not originate in the acoustic properties of these stimuli, but rather in the processing time dedicated to these stimuli. In the processing model represented above in Figure 1, stimuli like [sk<sup>h</sup>án] are subject to an extra stage of processing. The perceptual system first creates an acoustic representation of the stimulus, and then imposes a prosodic parse on this representation. This prosodic parse is then measured against the demands of the grammar, and if it is found to be well-formed a perceptual decision is taken. However, no well-formed prosodic parse of [sk<sup>h</sup>án] is possible. At this stage, the perceptual system therefore has to change the mental representation of the stimulus (via perceptual epenthesis) and re-parse the stimulus prosodically. This additional step takes up additional processing time, resulting in slower response times for stimuli like these.

##### 5. Experiment 3: Begins on ...

The results of Experiments 1 and 2 show that listeners represent stimuli like [st<sup>h</sup>api]~[sət<sup>h</sup>api] and [sək<sup>h</sup>án]~[sk<sup>h</sup>án] the same during late stages of perceptual processing, in agreement with the processing model represented in Figure 1. However, these results do not give evidence for the earlier stage of acoustic encoding proposed in Figure 1, where the members of these pairs are hypothesized to have different mental representations. Although the results of Experiments 1 and 2 are hence consistent with the model represented in Figure 1, they are also consistent with other processing models. Another possibility is that the percept [sək<sup>h</sup>án] is activated for both [sk<sup>h</sup>án] and [sək<sup>h</sup>án] stimuli right from the earliest stages of processing, but that activation of this percept reaches the critical decision level sooner for an input stimulus that is actually identical to this percept. Such an interpretation would be consistent with a perceptual processing model that does

not contain an initial stage of phonetic encoding that is informed only by the actual acoustic input stimulus. If this were correct, then both [sk<sup>h</sup>án] and [sək<sup>h</sup>án] will be mentally represented as [sək<sup>h</sup>án] from the earliest stages of processing. There should then be no stage of processing where the perceptual system can discriminate these stimuli consistently. On the other hand, if there is an initial stage of phonetic encoding, as represented in Figure 1, where the mental representations of [sk<sup>h</sup>án] and [sək<sup>h</sup>án] stimuli differ, then there must be a stage of processing where the perceptual system can discriminate these two types of stimuli.

The experimental tasks used in Experiments 1 and 2 by their very nature tap into the later stages of perceptual processing and therefore cannot give evidence about the earlier stages of processing. Experiment 3 is designed to address this issue. In this experiment, participants are presented with the same stimuli as in Experiment 2. However, rather than counting syllables, participants are instructed to decide whether a stimulus begins on, for instance, [sk...] or [sək...]. This task does not require prosodic parsing, nor does it require of participants to listen to the full token before responding. Participants can thus respond faster so that their responses may reflect earlier stages of processing.

### 5.1. Methods

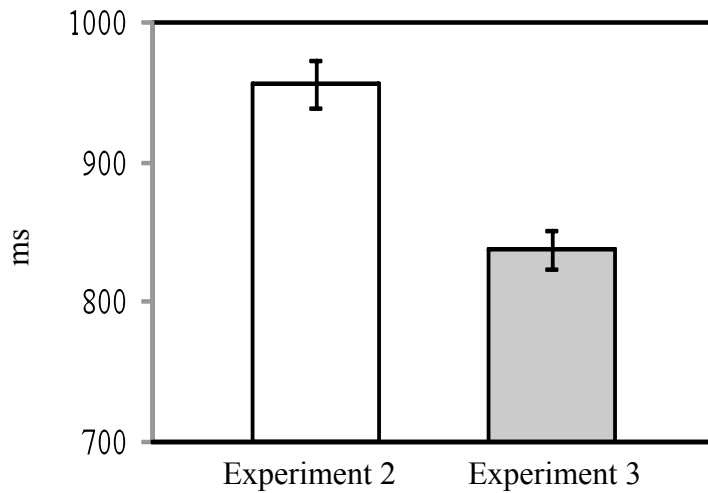
*Participants.* Eleven undergraduate students from the University of Michigan participated in this experiment. All participants were native speakers of American English with no known speech or hearing deficits. There was no overlap between participants in this and earlier experiments. Participants were paid for their participation.

*Token selection and stimuli.* The exact same stimuli as in Experiment 2 were used in Experiment 3.

*Procedure.* As with Experiments 1 and 2, stimulus presentation and response collection were controlled with *Superlab*. Presentation of each token was preceded by a visual cue on the computer monitor of two options represented in English orthography. The two options always corresponded to a percept with and without a schwa, with schwa represented with the letter *e*. For [s]-initial tokens, the options therefore consisted of *s* or *se* followed by one of {*p, t, k*}. For example, a token like [sk<sup>h</sup>án] would be preceded by the visual cue “sk... OR sek...”. For [s]-medial tokens, the visual cue consisted of an orthographic representation of the initial syllable, followed by *s* or *se* and one of {*p, t, k*}. For example, [fisək<sup>h</sup>án] would be preceded by “feesk... OR feesk...”. In the first syllable of these [s]-medial tokens, [i] was always represented by the letters *ee*, [u] by *oo*, [ɛ] by *e*, and [a] by *a*. The order between the cue with the schwa and without the schwa was balanced by token type. Participants were instructed to read the visual cue before the auditory token was presented, and to decide on which sound sequence the auditory stimulus begins by pushing either the leftmost or the rightmost button on the response box. Response times were recorded from the onset of the auditory stimulus to the moment that a participant pushed a button on the response box. In order to increase the likelihood that participants’ responses will be based on earlier stages of processing, they were encouraged to respond as quickly as possible, even before the full token has played if that was possible. Each token was included once per experimental block, for 72 total tokens. The block was presented 4 times, with participants allowed a short self paced break between blocks. Tokens were differently randomized for each participant and each block. Before the experimental trials began, participants received 20 practice trials, randomly selected from the experimental trials. A different selection was made for each participant. Participants received no feedback.

## 5.2. Results and discussion

In order to determine whether the different design succeeded in eliciting faster responses, and hence responses based on earlier stages of processing, the response times of Experiment 3 were analyzed. The response times were first standardized, exactly as in Experiment 2, and responses more than 2 standard deviations from the mean for each participant were excluded, resulting in the exclusion of just under 4% of all responses. As a rough measure of success, the response times to different tokens in Experiment 2 ( $\mu = 999$  ms) and Experiment 3 ( $\mu = 930$  ms) were compared using a one-tailed, paired sample  $t$ -tests, returning a highly significant difference ( $t(71) = 3.76, p < .001$ ). On average, response times in Experiment 3 were therefore shorter than in Experiment 2. However, the more important comparison is the response time to [sk<sup>h</sup>án] type tokens between the two experiments. The goal of Experiment 3 is to elicit responses for these tokens at an earlier stage of processing, and we therefore need the response time to these tokens to be faster in Experiment 3 than Experiment 2. The response times on these tokens for the two experiments are represented in Figure 5. These response were also compared using one-tailed, two sample  $t$ -tests. These tests returned significant results both by items ( $t(22) = 5.8, p < 0.001$ ) and by participants ( $t(21) = 2.0, p < 0.03$ ). I conclude that the difference in task in Experiment 3 has therefore resulted in speeding up the responses of the participants.



*Figure 5* Average response time by item to [sk<sup>h</sup>án] type stimuli in Experiments 2 and 3. Error bars show 95% confidence intervals.

Responses were coded as ‘correct’ or ‘error’ in terms of the actual acoustic properties of each token. The average response pattern per participant is represented in Figure 6. The percent correct responses on each stimulus type was subjected to a one sample *t*-test to determine whether it was significantly above chance. The results of these *t*-tests are given in (11)

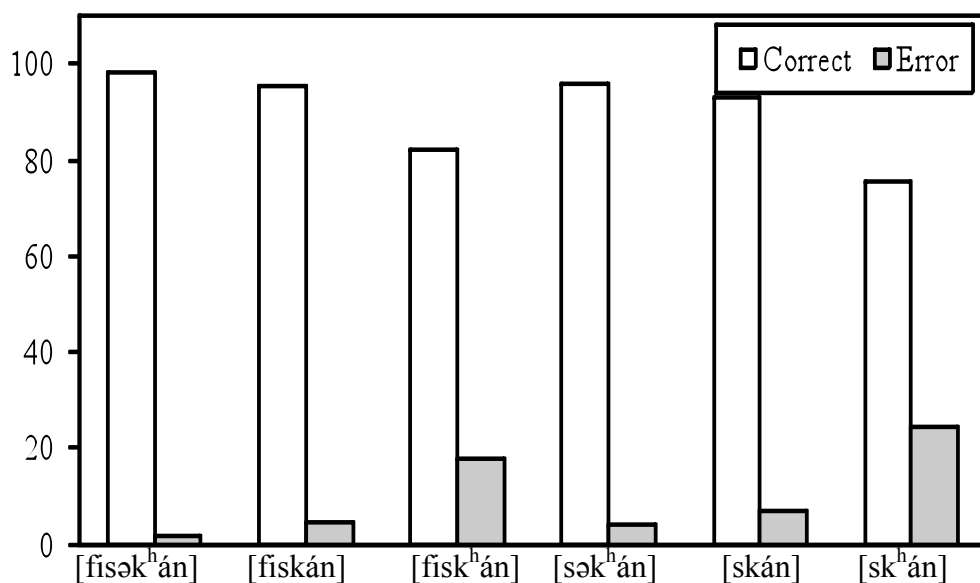


Figure 6 Mean percent correct for Experiment 3 by participant for each token type.

(11) Results of *t*-tests for Experiment 3

<i>Stimulus type</i>	<i>Participant analysis</i>		<i>Item analysis</i>	
	<i>t</i> (10)	<i>p</i>	<i>t</i> (11)	<i>p</i>
[fisækʰán]	76.5	< 0.001	94.9	< 0.001
[fiskán]	33.9	< 0.001	54.1	< 0.001
[fiskʰán]	8.3	< 0.001	5.9	< 0.001
[sækʰán]	45.3	< 0.001	52.2	< 0.001
[skán]	21.0	< 0.001	25.8	< 0.001
[skʰán]	5.4	< 0.001	7.2	< 0.001

The percent correct response was significantly above chance for all token types. Most importantly, the response pattern to [skʰán] token types in Experiment 3 is different from those in

Experiment 2. In Experiment 2, participants were about equally as likely to perceive these tokens as [sək<sup>h</sup>án] (with two syllables) or as [sk<sup>h</sup>án] (with only one syllable). However, in Experiment 3, participants overwhelmingly perceive these stimuli as starting on [sk...], and therefore distinguish these stimuli from [sək<sup>h</sup>án] type stimuli. Participants differentiate [sk<sup>h</sup>an] and [sək<sup>h</sup>an] at the faster response rates of Experiment 3 but they confuse [sk<sup>h</sup>an] with [sək<sup>h</sup>an] at slower response rates in Experiment 2. I interpret this as evidence that though [sk<sup>h</sup>an] and [sək<sup>h</sup>an] receive the same mental representation at later stages of processing, there is, in accordance with the model in Figure 1, an earlier stage of phonetic encoding where these types of tokens have separate representations.

## 6. General discussion

### 6.1. Mismatch between stimulus and percept as evidence about grammar

The human perceptual system does not function like a photocopier. It does not merely translate the input signal into a faithful mental representation. It rather takes the input signal and transforms it into a percept that bears resemblance to the input signal, but that also differs from it in lawful ways. This is true of all aspects of human perception—visual (Gordon 2004), olfactory (Wilson and Stevenson 2006), tactile (Schiff and Foulke 1982), auditory (Warren 2008), etc. One of the central goals of cognitive psychology is to discover the laws that dictate the ways in which the percept can deviate from the actual input. Part of the deviation is just the result of the physiological properties of the different organs involved in perception. But physiology does not explain all of perception. There are also operational or processing laws that influence the way in which the stimulus is processed once it has been translated by the primary perceptual organs into some kind of mental representation. One of the processing related factors that influence the auditory perception of speech is the grammar of the listener. Over the past several decades, a

large body of evidence has been amassed showing that the exact same acoustic speech stimulus can be perceived very differently by speakers of different languages. Under the assumption that there are no major differences in the physiology of the auditory perception organs between speakers of different languages, such differences must have their origin somewhere in the processing that happens in the mind after the auditory speech stimulus has been translated into a mental representation. In order to understand the human perceptual system and the cognitive processes that support perception, it is therefore necessary to understand which aspects of grammar can influence speech perception, and how these aspects influence speech perception.

In the same way that the study of speech perception is important for our general understanding of human cognition, it is also important for our understanding of the grammatical competence of the language user. Studying the mismatches between acoustic speech stimuli and the percepts that listeners form, can give us insight into the structure of grammar. In this paper, we have seen how such a mismatch can provide evidence that speech is parsed into higher order prosodic structure. The percept at which listeners arrive for the exact same acoustic stimulus depends partially on the position that the stimulus occupies in higher order prosodic structure. Specifically, an acoustic stimulus that corresponds to the phone sequence [st<sup>h</sup>] is perceived as such when it appears in a context where a syllable boundary can be lawfully inserted between the [s] and [t<sup>h</sup>]*—i.e. in a context such as [lu\_\_ápi]. But this same acoustic stimulus is perceived as [sət<sup>h</sup>] when it appears in a context where this syllable parse is not possible*—i.e. in a context such as [\_\_ápi].**

In English, aspirated stops are only tolerated in absolute syllable initial position. This presents a possible solution for the mismatch between the acoustic stimulus [st<sup>h</sup>] and the percept [sət<sup>h</sup>] in the [\_\_ápi] context. The listener accurately perceives the aspiration in the acoustic

stimulus. The aspiration is taken as the cue for the start of a new syllable, and the first three phones of the [st<sup>h</sup>ápi] stimulus are hence prosodically parsed as [s.t<sup>h</sup>á...]. This leaves the word-initial [s] stranded without a syllabic nucleus with which it can affiliate, and causes the perceptual system to perceive an illusory schwa between the [s] and [t<sup>h</sup>]. In the [lu\_\_ápi] context, this is not necessary. As before, the aspiration is taken as a cue to start a new syllable, resulting in a prosodic parse [lus.t<sup>h</sup>a...]. Since the [s] is now preceded by a vowel with which it can affiliate, this sequence is well-formed, and there is no need for an illusory vowel. Somewhere during the processing of the speech stimulus, higher order prosodic structure is therefore imposed on the stimulus. Even though it is impossible to detect a syllable boundary between [s] and [t<sup>h</sup>] in the acoustics of the stimulus, the way in which this stimulus is processed gives evidence for the reality of such a prosodic boundary.

It is important to note what the repair is that is imposed on the stimulus. When faced with a prosodically ill-formed representation such as [s.t<sup>h</sup>a...], there are several ways in which it could be changed to become well-formed. The aspiration can be perceptually deleted, so that the percept can be parsed syllabically as [stá...]. Alternatively, the initial [s] can be perceptually deleted, giving the well-formed parse [t<sup>h</sup>á...]. Another conceivable perceptual repair is to perceptually insert a vowel before [s], as in [əs.t<sup>h</sup>á...]. All of these would result in prosodically well-formed percepts. However, the English listeners in the experiments reported above preferred a different repair—perceptually inserting a vowel between the [s] and [t<sup>h</sup>]. A question that this paper does not address is whether the specific perceptual repair used to remedy such prosodically ill-formed stimuli is determined by the grammar of the listeners (i.e. it is language specific), or whether the same repair will be observed irrespective of the language of the listeners. Further research would be required to answer this question—though see Fleischhacker (2001) for

an exploration of the idea that the perceptual repair is selected that is acoustically most similar to the actual acoustic stimulus.

## 6.2. The time course of perceptual epenthesis

Behavioral data, such as those reported in this paper, are rather course-grained in their time resolution, and therefore not optimal for probing into the time course of mental processing where very small time differences can be at play. Even so, behavioral data can reveal much about processing. The model represented in Figure 1 makes two very specific predictions with regard to the time course of processing, and the experiments presented above provide evidence for both of these predictions.

First, the model predicts that there will be an additional processing step involved in tokens that are prosodically ill-formed, and hence that such tokens will take longer to process. The results of Experiment 2 are in agreement with this prediction. Although [st<sup>h</sup>áp]-type stimuli are physically of shorter duration than [sət<sup>h</sup>áp]-type stimuli, participants are slower at responding to [st<sup>h</sup>áp]-type stimuli. During acoustic encoding, faithful acoustic representations of both of these tokens are made. In the next step of processing, prosodic parses are imposed on these acoustic representations, and these parses are compared against the grammar for well-formedness. Since a well-formed prosodic parse of [sət<sup>h</sup>áp] is possible, namely [sə.t<sup>h</sup>áp], no additional processing is required for this token. On the other hand, no well-formed parse of [st<sup>h</sup>áp] exists—both [st<sup>h</sup>áp] and [s.t<sup>h</sup>áp] are not in agreement with English phonological grammar. An additional processing step is hence required to bring this token type into agreement with the phonology of English, resulting in slower processing.

However, although the longer processing time associated with [st<sup>h</sup>áp]-type stimuli are consistent with the extra processing step of the model represented in Figure 1, it does not

actually confirm the existence of such an extra processing step. It is also possible that multiple perceptual hypotheses are entertained from the earliest stages of perceptual processing, so that the percept [sə<sup>h</sup>áp] is activated for both [st<sup>h</sup>áp] and [sət<sup>h</sup>áp] stimuli even during early auditory processing, but that activation of this percept reaches the critical decision level sooner for an input stimulus that is actually identical to this percept. Such an interpretation would be consistent with a perceptual processing model that does not contain an initial stage of phonetic encoding that is informed only by the actual acoustic input stimulus. However, the results of Experiment 3 give evidence for the existence of such an early acoustic encoding stage in speech processing. In this experiment, participants responded faster to and clearly differentiated [st<sup>h</sup>áp] and [sət<sup>h</sup>áp] type stimuli.

Even the task in Experiment 3 is still behavioral in nature, and therefore not ideally suited to tap into the earliest stages of auditory processing. Non-behavioral methods that directly measure real time brain response to stimulus presentation, such as ERP (event related potentials) studies, are better suited to studying early stages and the fine-grained time course of processing (Praamstra et al. 1994; Praamstra & Stegeman 1993). In ERP studies, two stimuli are presented with a short inter stimulus interval, and brain activity is measured during the presentation of the second stimulus. Some of the stimuli pairs presented are identical, while others differ from each other. By comparing brain activity between the “same” and “different” conditions, it can be determined whether these two conditions are processed differently by the brain.

Wagner and Schafer (2008) conducted such an ERP study with stimuli very similar to those used in the experiments discussed in this paper. The stimuli in their “same” condition contained pairs like [ptaki]~[ptaki]/[pətaki]~[pətaki], and the stimuli in their “different” conditions pairs like [ptaki]~[pətaki]/[pətaki]~[ptaki]. The participants in their study performed a same/different

task similar to that used in Experiment 1 above. They found that English listeners could not consistently identify [ptaki]~[pətaki] pairs as “different”, giving more evidence of perceptual epenthesis. The crucial difference between their experiment and Experiment 1 is that they collected ERP data while their participants were performing the same/different task. They found evidence for early differences in brain activity between same and different pairs, around 500 ms after the onset of the second pair member. But this difference disappears at later processing stages, somewhere around the 1000 ms mark. They interpret this as evidence for a separate early stage of processing, consistent with the phonetic encoding stage proposed in Figure 1 above. The behavioral same/different response happens later than the 1000 ms time point, and is therefore based on the later processing stages where the [ptaki] and [pətaki] are no longer distinguished.

Dehaene-Lambertz et al. (2000) perform a similar study with Japanese listeners, investigating the results of Dupoux et al. (1999, 2001) further. Dupoux et al. showed that Japanese listeners cannot discriminate tokens like [ebzo] and [ebuzo] consistently because of perceptual epenthesis in [ebzo]. The ERP results of Dehaene-Lambertz et al., however, are in conflict with those of Wagner and Schafer described just above. They did not find evidence for consistent differences in the brain activity of Japanese listeners between the “same” ([ebzo]~[ebzo], [ebuzo]~[ebuzo]) and “different” ([ebzo]~[ebuzo], [ebuzo]~[ebzo]) conditions, not even during the earliest stages of processing. They conclude that their results “suggest that the impact of phonotactics takes place early in speech processing and support models of speech perception, which postulate that the input signal is directly parsed into native language phonological format” (Dehaene-Lambertz et al. 2000, 635). Their results thus speak against the initial stage of phonetic encoding proposed in Figure 1.

Given these conflicting results, it cannot be decided definitively at current whether there is indeed a separate level of phonetic encoding—however see Kingston (2005) and Poeppel et al. (2008) for more arguments for such a level. More research, and specifically more ERP type investigations of similar phenomena, is necessary.

## References

- Baayen, R. Harald, Richard Piepenbrock and Leon Gulikers L. Gulikers. 1995. *The CELEX Lexical Database (Release 2) [CD-ROM]*. Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania.
- Beckman, Mary E. and Jan Edwards. 1990. Lengthenings and shortenings and the nature of prosodic constituency. In John Kingston and Mary E. Beckman (eds.) *Papers in Laboratory Phonology I: Between the Grammar and the Physics of Speech*. Cambridge: Cambridge University Press, 152-178.
- Berent, Iris and Tracy Lennertz. 2007. What we know about what we have never heard: Beyond phonetics. Reply to Peperkamp. *Cognition*. 104:638-643.
- Berent, Iris, Donca Steriade, Tracy Lennertz and Vered Vaknin. 2007. What we know about what we have never heard: Evidence from perceptual illusions. *Cognition*. 104:591-630.
- Boersma, Paul and David Weenink. 2008. *Praat: Doing Phonetics by Computer (Version 5.0.24)*. [Computer program available at <http://www.praat.org/>].
- Byrd, Dani and Elliot Saltzman. 2003. The elastic phrase: modeling the dynamics of boundary-adjacent lengthening. *Journal of Phonetics*. 31:149-180.
- Cholin, Joana, Niels O. Schiller and Willem J.M. Levelt. 2004. The preparation of syllables in speech production. *Journal of Memory and Language*. 50:47-61.
- Clifton, Charles, Katy Carlson and Lyn Frazier. 2006. Tracking the what and why of speakers' choices: Prosodic boundaries and the length of constituents. *Psychonomic Bulletin & Review*. 13:854-861.
- Coetzee, Andries W. and Daan Wissing. 2007. Global and local durational properties in three varieties of South African English. *The Linguistic Review*. 24:263-289.

- Connine, Cynthia M., Debra Titone and Jian Wang. 1993. Auditory word recognition: extrinsic and intrinsic effects of word frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. 19:81-94.
- Davidson, Lisa. 2006. Phonology, phonetics, or frequency: Influences on the production of non-native sequences. *Journal of Phonetics*. 34: 104-137.
- . 2007. The relationship between the perception of non-native phonotactics and loanword adaptation. *Phonology*. 24:261-286.
- Dehaene-Lambertz, Ghislaine, Emmanuel Dupoux and Ariel Gout. 2000. Electrophysiological correlates of phonological processing: a cross-linguistic study. *Journal of Cognitive Neuroscience*. 12:635-647.
- Dupoux, Emmanuel, Kazuhiko Kakehi, Yuki Hirose, Christophe Pallier and Jacques Mehler. 1999. Epenthetic vowels in Japanese: A perceptual illusion? *Journal of Experimental Psychology: Human Perception and Performance* 25, 1568–1578.
- Dupoux, Emmanuel, Christophe Pallier, Kazuhiko Kakehi & Jacques Mehler (2001). New evidence for prelexical phonological processing in word recognition. *Language and Cognitive Processes*. 5:491–505.
- Fleischhacker, Heidi. 2001. Cluster-dependent epenthesis asymmetries. In Adam Albright and Taehong Cho (eds.) *UCLA Working Papers in Linguistics 7. Papers in Phonology 5*. Los Angeles: UCLA Linguistics Department, 71-116.
- Fromkin, Victoria. 1971. The non-anomalous nature of anomalous utterances. *Language*. 47:27-52.
- Gordon, Ian E. 2004. *Theories of Visual Perception*. New York, NY: Psychology Press.
- Kabak, Barış and William J. Idsardi. 2007. Perceptual distortions in the adaptation of English consonant clusters: Syllable structure or consonantal contact constraints? . *Language and Speech*. 50:23-52.
- Kingston, John. 2005. Ears to categories: New arguments for autonomy. In Sonia Frota, Marina Vigarío and Maria João Freitas (eds.) *Prosodies: With Special Reference to Iberian Languages*. New York: Mouton de Gruyter, 177-222.
- Lennertz, Tracy and Iris Berent. 2007. Markedness constraints on the perception of s/z-initial onset clusters. *Paper presented at the Workshop on Variation, Gradience and Frequency*

- in Phonology*. Stanford University. [Downloaded on June 23, 2008 from [www.stanford.edu/dept/linguistics/linginst/nsf-workshop/Lennertz&Berent\\_Poster.pdf](http://www.stanford.edu/dept/linguistics/linginst/nsf-workshop/Lennertz&Berent_Poster.pdf)].
- Levelt, Willem J.M. 2001. Spoken word production: A theory of lexical access. *Proceedings of the National Academy of Sciences of the United States of America*. 98:13464-13471.
- MacMillan, Neil A. and C. Douglas Creelman. 2005. *Detection Theory: A User's Guide*. Mahwah, NJ: Lawrence Erlbaum.
- Mattys, Sven L. and James F. Melhorn. 2005. How do syllables contribute to the perception of spoken English? Insight from the migration paradigm. *Language and Speech*. 48:223-253.
- McQueen, James M. 1998. Segmentation of continuous speech using phonotactics. *Journal of Memory and Language*. 39:21-46.
- Nespor, Marina and Irene Vogel. 1986. *Prosodic Phonology*. Dordrecht: Foris.
- Newman, Rochelle S., James R. Sawusch and Paul A. Luce. 1997. Lexical neighborhood effects in phonetic processing. *Journal of Experimental Psychology: Human Perception and Performance*. 23:873-889.
- Poeppel, David, William J. Idsardi and Virginie van Wassenhove. 2008. Speech perception at the interface of neurobiology and linguistics *Philosophical Transactions of the Royal Society. B* 363:1071-1086.
- Port, Robert F. and Adam P. Leary. 2005. Against formal phonology. *Language*. 81:927-964.
- Praamstra, Peter, Antje S. Meyer and Willem J.M. Levelt. 1994. Neurophysiological manifestations of phonological processing: Latency variation of a negative ERP component timelocked to phonological mismatch. *Journal of Cognitive Neuroscience*. 6:204-219.
- Praamstra, Peter and Dick F. Stegeman. 1993. Phonological effects on the auditory N400 event-related brain potential. *Cognitive Brain Research*. 1:73-86.
- Schafer, Amy, Katy Carlson, Charles Clifton and Lyn Frazier. 2000. Focus and the interpretation of pitch accent: Disambiguating embedded questions. *Language and Speech*. 43:75-105.
- Schiff, William and Emerson Foulke (eds.) 1982. *Tactual Perception: A Sourcebook*. Cambridge: Cambridge University Press.
- Selkirk, Elizabeth O. 1978. On prosodic structure and its relation to syntactic structure. In Thorstein Fretheim (ed.) *Nordic Prosody II*. Trondheim: Tapir, 111-140.

- 1981. On the nature of phonological representation. In Terry Myers, John Laver and John Mathieson Anderson (eds.) *The Cognitive Representation of Speech*. Amsterdam: North Holland, 379-388.
- 1986. On derived domains in sentence phonology. *Phonology Yearbook*. 3:371-405.
- 2001. On the phonologically driven non-realization of function words. In Charles Chang, Michael J. Houser, Yuni Kim, David Mortensen, Mischa Park-Doob & Maziar Toosarvandani (eds.) *BLS 27: Proceedings of the Twenty-Seventh Annual Meeting of the Berkeley Linguistics Society*. Berkeley: Berkeley Linguistics Society, 257-270.
- Sugahara, Mariko. 2005. Post-FOCUS prosodic boundaries in Tokyo Japanese: Asymmetric behavior of an F0 cue and domain-final lengthening. *Studia Linguistica*. 59:144-173.
- Turk, Alice. E. and Stefanie Shattuck-Hufnagel. 2000. Word-boundary-related duration patterns in English. *Journal of Phonetics*. 28:397-440.
- Wagner, Monica Palmieri and Valerie Shafer. 2008. Phonotactic influences on the perception of consonant clusters by English and Polish listeners. [Handout and audiorecording of presentation downloaded on June 26, 2008 from <http://www.cunyphonologyforum.net/SYLLPAPERS/>.]
- Warren, Richard M. 2008. *Auditory Perception: An Analysis and Synthesis*. Cambridge: Cambridge University Press.
- Wightman, Colin W., Stefanie Shattuck-Hufnagel, Mari Ostendorf and Patti J. Price. 1992. Segmental durations in the vicinity of prosodic phrase boundaries. *Journal of the Acoustical Society of America*. 91:1707-1717.
- Wilson, Donald A. and Richard J. Stevenson. 2006. *Learning to Smell: Olfactory Perception from Neurobiology to Behavior*. Baltimore, MD: Johns Hopkins University Press.