Niebuhr *EURASIP Journal on Audio, Speech, and Music Processing* (2017) 2017:19
DOI 10.1186/s13636-017-0115-3

EURASIP Journal on Audio, Speech, and Music Processing

RESEARCH Open Access

On the perception of “segmental intonation”: F0 context effects on sibilant identification in German

Oliver Niebuhr[1,2]

Abstract

In normal modally voiced utterances, voiceless fricatives like [s], [ʃ], [f], and [x] vary such that their aperiodic pitch impressions mirror the pitch level of the adjacent F0 contour. For instance, if the F0 contour creates a high or low pitch context, then the aperiodic pitch impression of the fricative in this context will also be high or low. This context-matching effect has been termed “segmental intonation”. While there is accumulating evidence for segmental intonation in speech production, less is known about if and how segmental intonation is actually integrated in the perception of utterance tunes. This question is addressed here in a perception experiment in which listeners identified target words ending in either [ʃ] or [s]. The two sibilants inherently create low or high aperiodic pitch impressions in listeners due to their characteristically different spectral energy distributions. The sibilants were preceded by high or low F0 contexts in the target words. Results show a clear F0-context effect. The context effect triggered more [ʃ] identifications in high-F0 and/or more [s] identifications in low-F0 contexts. The effect was larger for sibilants that were less clearly identifiable as either /ʃ/ or /s/. The effect represents strong supporting evidence that listeners in fact perceive the segmental intonation of fricatives and integrate its aperiodic pitch with the F0-based pitch when perceiving utterance intonation. Thus, the term “segmental intonation” is perceptually appropriate. Furthermore, the results are discussed with respect to reaction-time measurements and an additional effect of the quality of the adjacent vowel phoneme on sibilant identification.

Keywords: Speech, Prosody, Intonation, F0, Pitch, Sibilant, Perception, German, Segmental intonation

1 Introduction

1.1 Background: the notion of “segmental intonation”

Studying intonation, be it for phonological or technical purposes, focuses on the primary acoustic correlate of perceived pitch: the fundamental frequency (F0). This focus on F0 is, of course, well justified and allowed speech scientists of the past decades to accumulate considerable knowledge of the syntagmatic structure of intonational tunes and the forms and functions of melodic elements at each structural position in the syntagma (for German, cf. [1–5]).

However, focussing on F0 alone ignores that noise segments in speech can in principle also convey pitch, more specifically, aperiodic pitch impressions. The speech signal that we transmit is, roughly speaking, a combination (or, more specifically, a multiplication) of source and filter characteristics [6]. Voiceless segments in speech lack a periodic phonatory source. But, many of these voiceless segments have a noise source instead. In speech, we call these noise-excitated sound segments “voiceless fricatives”. Their noise is created somewhere within the speech production apparatus and hence also subject to vocal-tract filtering, primarily based on the resonance cavities that follow in the direction of the airflow after the noise excitation [6]. Just like for vowels or sonorants, this noise-source filtering gives voiceless fricatives at each place of articulation, a specific resonance pattern. It shapes the corresponding fricative’s spectral energy distribution and creates a characteristic formant structure. As is stressed by Johnson [6], “The spectrum of turbulent noise normally seen in fricatives is not completely flat”. Particularly sibilants, i.e. the research subject of the present study, “have very pronounced spectral peaks” [7].

Correspondence: olni@sdu.dk
[1] Innovation Research Cluster Alsion, Mads Clausen Institute, University of Southern Denmark, Sonderborg, Denmark
[2] Ketelsenweg 6b, D-24983 Handewitt, Germany

The crucial point with these non-flat spectra or formant structures is now that they also contribute to our perception of pitch, as was shown, for example, in studies on intrinsic pitch in speech and other studies with more psychoacoustic stimuli [8]. For voiceless fricatives, the formant pattern determines our aperiodic pitch impression. Traunmüller [9], for example, termed the aperiodic pitch impression of fricatives "sibilant pitch". Based on a series of psychoacoustic experiments, he developed a method for quantifying sibilant pitch in terms of a weighted combination of acoustic energy in different frequency bands. Roughly speaking, the sibilant pitch perceived for voiceless front vowels (i.e. variants of the voiceless fricative [h]) is determined by the vowels' second formant frequency (F2) and subsequent higher formants in the spectral energy pattern. For voiceless back vowels, it is mainly the first formant frequency (F1) that determines the sibilant pitch evoked by the vowel. These findings are in agreement with Thomas [10] who already pointed out, on an auditory basis, the close correlation between formant frequencies (F2 in particular) of voiceless vowels and their aperiodic pitch impression. Later, Higashikawa [11] and Higashikawa and Minifie [12] added to this picture that sibilant pitch perception—or "whisper pitch" in the terminology of Higashikawa and colleagues—is independent of the listener's linguistic or phonological background and that simultaneous changes in F1 and higher formants like F2 are more effective in changing sibilant pitch than changes in F1 or F2 alone.

The notions of "sibilant pitch" or "whisper pitch" raise the question of whether speakers make use of this perceptual phenomenon in speech communication by systematically changing the spectral energy distribution of the speech sounds they produce. The obvious starting point for addressing this question is whispered speech, i.e. voiceless speech with a continuous noise excitation. Both production and perception studies on whispered speech presented largely converging evidence for systematic changes in the spectral energy distribution such that the resulting aperiodic pitch contour still allows listeners to reliably identify all those communicative meanings and functions that would otherwise be encoded in F0-based intonation and/or tone patterns (cf. [13–19, 93]; see [20, 21] for exceptions).

While the studies above leave hardly any doubt that speakers can in principle control and vary aperiodic pitch impressions of voiceless sound sections in speech production and that they actually do so in whispered speech, it seems to be a common—either implicit or explicit—assumption (not only in studies on whispered speech) that controlling aperiodic pitch impressions becomes irrelevant for speakers as soon as F0 information is available. Also, the study of Coleman et al. [93] whose hypothesis (H1) and results stress the similarities in the pitch control of normal and whispered speech did not include and analyze a mixed condition in which unvoiced sounds are produced in combination with voiced sounds. In other words, it is typically assumed or at least not explicitly questioned that voiceless fricative sounds like [f], [s], [ʃ], and [x] make no separate contribution to the pitch contour in all regular everyday utterances with modally voiced vowels and sonorants. This would mean that the aperiodic pitch impressions of voiceless fricatives either remain approximately constant for each fricative and speaker or that they vary randomly across different contexts, particularly those contexts of intonation and tone. This is quite an assumption given that (a) about 25% of the speech signal in, for example, English everyday conversation is voiceless, that (b) almost half of this percentage is due to voiceless fricatives [22], and that (c) languages have more voiceless fricatives than voiceless stops (according to the UPSID database, [23]).

Challenging the assumption that the aperiodic pitch impressions of voiceless fricatives are irrelevant in all but whispered utterances, Niebuhr [24–26] analyzed postaspiration noises, i.e. [_^{h}] after /t/, as well as the voiceless fricatives [f], [ʃ], and [x] in modally voiced utterances of German read speech. He took spectral measurements like the average center-of-gravity (CoG) that are more holistically oriented and hence easier and more reliable to apply than measurements of individual formants or frequency ranges, but still reflect the known parameter changes of "sibilant pitch" or "whisper pitch". The results of the acoustic analyses led to the clear conclusion that fricative segments do show systematic variation as to their aperiodic pitch impression, also in regular, modally voiced utterances. This conclusion was recently supported by Heeren [27] who found for VCV syllable productions of Dutch speakers that "the acoustic correlates of fricatives […] are systematically varied" in both "whispered and normal speech" (p. 3427).

In addition, Niebuhr [26] found this variation to be related to the adjacent F0 context: "the spectral properties of [voiceless] segments vary in different F0 contexts in such a way that the pitch impressions caused by these segments can support the signalling of intonational forms and functions" (p. 21). An example of this phenomenon is provided in Fig. 1a, b. They show the German utterance "nur hundert Schuss" (only hundred shot), realized with (a) statement and (b) question tunes. The utterance-final pitch-accented target word "Schuss" (shot) is framed by two voiceless fricatives, [ʃ] and [s]. In the right panel (b), the word "Schuss" is associated with a low pitch accent, followed by a high edge tone. Thus, F0 falls into the accented short vowel [ʊ] and then changes into a steep rise that continues until the end of voicing. The opposite intonational pattern spans the word "Schuss" in panel (a).

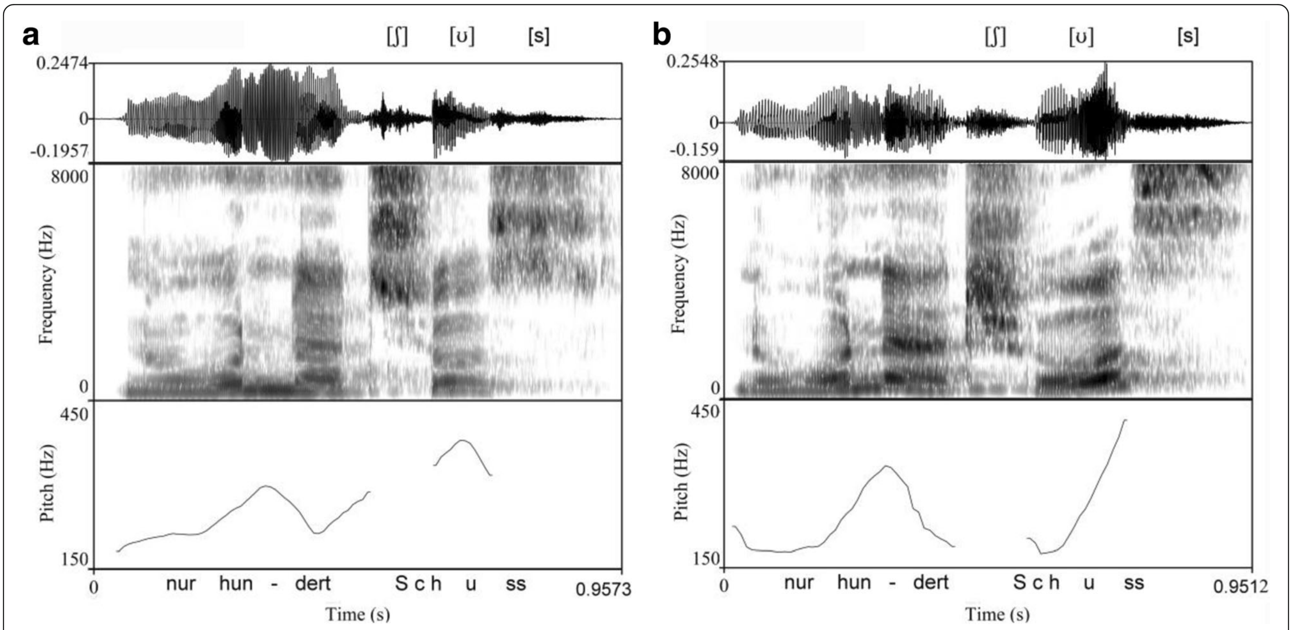


Fig. 1 Waveform, spectrogram, and F0 contour of "nur hundert Schuss" (only hundred shots), produced naturally by a female speakers (DT) as a statement (**a**) and as a question (**b**). The production differences in the final word "Schuss" concern both the nuclear F0 pattern and its fricative frame constituted by the sibilants [ʃ] and [s]

The voiceless fricatives surrounding the [ʊ] interrupt F0. But, at the same time, they mirror the F0 context in which they occur: Compared to the high pitch-accent context in Fig. 1a, the word-initial [ʃ] in the low pitch-accent context of Fig. 1b has a lower spectral energy boundary (i.e. that frequency along the ascending *y*-axis at which the first significant increase in acoustic noise energy can be observed) and much more noise energy at lower spectral frequencies. In consequence, [ʃ] creates a lower aperiodic pitch impression in the low pitch-accent than in the high pitch-accent context. Likewise, the final [s] sounds higher pitched in the high edge-tone condition (b) than in the low edge-tone condition (a), due to the fact that the [s] in (b) has a higher lower spectral energy boundary and shows more noise energy at higher spectral frequencies than the [s] in (a).

Niebuhr [25] introduced the term *"segmental intonation"* to refer to this parallel between the pitch impressions caused by fricative segments and their adjacent F0 contexts. While the studies of Niebuhr [24] and [25] addressed segmental intonation only utterance-finally, Niebuhr et al. [28] found segmental intonation also in utterance-medial F0 contexts.

The phenomenon of segmental intonation has recently been replicated for other phonological F0 contexts of German [29] as well as for other languages like Polish [30] and Dutch [27]. However, for some languages and intonational or tonal embeddings, studies failed to find segmental intonation, which is one of the reasons why a general articulatory mechanism of pitch change, such as a shift in larynx height (cf. Coleman et al. [93]), is not a likely explanation for the parallel between the pitch impressions of fricatives and their adjacent F0 contexts. At least, it cannot be the only explanation. Further evidence that argues against larynx height as the underlying source of segmental intonation is the mere order of magnitude of the phenomenon. The acoustic changes associated with segmental intonation are much larger (formants move upward or downward by up to 66%) than those that would be expected from F0-related shifts in larynx height. These shifts contract or expand a speaker's vocal tract and the corresponding resonance cavities by about 2% (Coleman et al. 2002). The acoustic changes triggered by the F0 context also do not affect all formants at all points in time in the same way (e.g. in the same direction) as it would be the case if segmental intonation was caused by a F0-related shift in larynx height. Furthermore, Niebuhr [25] and Niebuhr et al. [28] report that the articulatory changes associated with segmental intonation also involve changes in a fricative's place of articulation, rounding or spreading (de-rounding) of lips, and even changes in post-lexical assimilation patterns in favour of those fricatives whose aperiodic pitch impressions match with the adjacent F0 context. These findings suggest that segmental intonation is based on or at least strongly shaped by an extrinsically controlled articulatory change that adjusts of a fricative's aperiodic pitch to the F0 context.

While the phenomenon of segmental intonation is more and more substantiated at the level of production, we still know little about how it affects speech communication at the level of *perception*. The present study is to shed more light on this issue.

1.2 Question and aim: the perception of "segmental intonation"

Niebuhr [24] and Kohler [31] found that segmental intonation interacts with pitch-accent meanings. They used two sets of stimuli. One set consisted of naturally produced utterances whose final high-rising or low-falling intonation contours were followed by either voiceless alveolar sibilants [s] or postaspirated stops [t^h]. For the other stimulus set, they interchanged the sibilants or postaspiration noises of the high-F0 and low-F0 contexts. Listeners judged the meanings of the nuclear (phrase-final) intonation contours in the two stimulus sets by means of a semantic differential. The results of both experiments showed that the semantic profiles of intonational meanings were less clear-cut in the interchanged conditions in which the segmental intonations mismatched with their F0 context.

Mixdorff et al. [32, 33] conducted a series of perception experiments in which listeners compared the prominence levels of disyllabic words whose prominence-causing F0 contours were interrupted by voiceless sound segments. The result of these discrimination tests was that listeners perceived higher prominence levels in those words in which the interruption concerned F0 slopes rather than F0 peaks and in which the interruptor was a voiceless fricative rather than voiceless stop. Mixdorff et al. concluded on this basis that, unlike the silence of stops, the noise of fricatives enables listeners to perceptually fill-in or restore missing F0, except when the missing F0 requires listeners to extrapolate phonologically relevant tonal targets like peaks and valleys. Furthermore, Mixdorff et al. speculate that the segmental intonation of fricatives could be one reason why utterance tunes appear "*subjectively continuous*" ([34]; 275), despite the considerable number and duration of F0 gaps in the acoustic signal.

Additional empirical support for the findings and conclusions of Mixdorff et al. was recently provided by Welby and Niebuhr [35]. Their study made use of the fact that the location of low-pitched F0 elbows, which may or may not be masked by voiceless sound segments, indicates word-boundary locations and is hence relevant for lexical-identification processes in French. Thus, Welby and Niebuhr were able to investigate how listeners treat F0 gaps of stops and fricatives simply by asking them to identify word sequences instead of letting them rate relative word prominences. Based on this straightforward word-identification task, Welby and Niebuhr found that the identification rate of French words decreased in comparison to a reference set of fully sonorous words when voiceless obstruents masked the F0 elbow that disambiguates the word-boundary position in pairs like "l'appel" and "la pelle". However, and this is the crucial point here, the decrease in word-identification rate was significantly greater for voiceless plosives than for voiceless fricatives (e.g., "l'affiche" vs. "la fiche"). It was ruled out by the selection of similarly common target words that these differences, particularly that between voiceless plosives and voiceless fricatives, could have emerged as a mere experimental artefact of lexical frequency. On this basis, Welby and Niebuhr concluded that French native listeners were to some degree able to reconstruct and perceive the low-pitched F0 elbow through the fricative noise, but not through the silence of a plosive.

In summary, previous studies gathered converging cross-linguistic evidence that listeners can perceive the variation in the sound quality of fricatives that is caused by the F0 context and for which Niebuhr [25] introduced the term "segmentation intonation". There is also evidence that listeners in some way relate the sound qualities of fricatives to the F0 context and that creating this relation allows them to (better) perceive intonational meanings or process and perceive noise gaps differently from silent gaps.

The crucial point is that what has *not* been shown so far is that listeners perceive the variation in the sound quality of fricatives directly as a variation in aperiodic pitch. The only evidence that listeners *can* perceive the variation in the sound quality of fricatives literally as "segmental intonation" comes from the study of Heeren [27]. She found in a discrimination task that listeners were reliably able to identify whether a fricative ([s] or [f]) was produced by speakers in combination with a high or a low tonal target, no matter if the fricative was extracted from whispered speech or a normally voiced utterance. Although this is a valuable initial insight, it is hard to evaluate to what extent Heeren's setup, i.e. isolated fricatives in a discrimination task, allows conclusions about real speech perception; more importantly, she directly asked her listeners to rate the fricative stimuli as either higher-pitched or lower-pitched. Thus, it is unclear if listeners would have also perceived the different fricative qualities in terms of aperiodic pitch impressions if they were not explicitly asked to do so.

For this reason, it is the aim of the present study to build upon the research of Heeren [27] and address the question whether the relation that is obviously established by listeners between a voiceless fricative and its F0 context is actually one of perceived pitch. In other words, do listeners perceive the F0-related variation in the sound quality of fricatives literally as "segmental intonation", i.e. as a variation in aperiodic pitch, even when they are not explicitly instructed to do so?

1.3 Approach and hypotheses

The experimental approach used in the present study is inspired by the feature-parsing model of Gow [36] and the work of, for example, Ohala and Feder [37], Fowler et al. [38], Harrington et al. [39], and Kleber et al. [40] on the perceptual compensation of coarticulation. The basic idea behind this approach is the following: If listeners interpret a certain aspect of a sound segment as being due to an external factor, then they "subtract" this aspect from the sound segment when identifying its phoneme category. For example, if listeners interpret a certain proportion of the formant-frequency level of a vowel as being due to coarticulation with the adjacent consonantal or vocalic context, then they "subtract" this proportion from the formant-frequency level before they use the F1-F2 pattern to identify the vowel's phoneme category. The subtracted proportion is then, however, not treated as cognitive waste. Rather, it is reassigned and included as a further acoustic cue or element in the perception of the sound or event that caused the coarticulation. For this reason, "parsing" is another frequent terminological metaphor that is used next to "subtraction".

Harrington et al. [39] found that increased F2 values can still trigger /u/ (rather than /i/) identifications if these increases were attributable by listeners to the high F2 level of an adjacent alveolar consonant. Similarly, but in the opposite direction, Kleber et al. [40] found lowered F2 values to still trigger /ʊ/ (rather than /ɪ/) identifications, if this lowering was attributable by listeners to the low F2 level of a subsequent back vowel. To sum up in the words of Ohala and Feder [37]: "There is, in fact, abundant evidence, both from perceptual experiments as well as phonology […], that listeners identify speech sounds in part by normalizing them with respect to their phonetic context".

The perceptual compensation of coarticulation is one of the many manifestations of an evolutionary essential top-down mechanism in our brain (cf. [41]; note that 'mechanism' is used here in a generic sense, i.e. there may well be different subtypes of this mechanism for different stimuli and modalities): The world consists of meaningful objects or categories. These objects or categories do not manifest themselves in the same way all the time. They vary constantly in different contexts. If we want to survive and succeed in such a world, then we need some kind of cleaning procedure that is able free meaningful objects or categories from their contextual contamination and thus stabilize our perception of these objects or categories. For example, red is often a warning color in nature. Red berries and mushrooms can be toxic. However, the spectral energy pattern of the color red, which reaches our eye as a visual stimulus, is different in the light of the midday sun than it is in the shadow or in the light of the rising or setting sun. This context effect on the spectral energy pattern of the color red must be removed if we want to protect ourselves consistently from eating potentially dangerous food. Likewise, in loud environments, the Lombard effect [42] changes the speech pattern of a person and makes it in many acoustic parameters (F0, speaking rate, spectral tilt, voice quality) very similar to the characteristic settings of anger or rage [43]. We are geared to identify speakers who are angry or in a rage as a potential threat. So, if we do not want to accidentally kill someone who actually could be an ally in our struggle for survival or if do not want ourselves to be killed by someone who we talk to in a loud environment, then it is better for us and our interlocutor to be able to "subtract" the noise-induced Lombard changes again from the speech signal before we interpret it under adverse listening conditions. The subtracted information is then free to be reassigned and used otherwise in perception, for example, for phoneme identification. A well-established example for this reassignment is the F0 micro-perturbation by consonants. This micro-perturbation is perceptually subtracted from the F0 course and, thus, not perceived at the level of intonation. Instead, it is reassigned and used for distinguishing voiced and voiceless consonants at the segmental level [94].

The subtraction principle as a manifestation of a general (type of) object or category stabilizer in our brain is also made use of in the present study and applied to the two German voiceless sibilant fricatives [ʃ] and [s]. The acoustic characteristics of the two sibilant fricatives and their relative differences have been analyzed in great detail in many previous studies (e.g., [28, 44–51]). Due to its lower spectral energy boundary and its high acoustic-energy concentration at mid frequencies of about 3–4 kHz, [ʃ] creates an inherently "dark" aperiodic pitch impression. This low-pitched impression is further enhanced by the fact that [ʃ] is produced with pronounced lip rounding in German [52]. In contrast, [s] is produced without lip rounding. Its lower spectral energy boundary is at least 1 kHz higher than that of [ʃ], and its acoustic energy is mainly bundled at frequencies higher than 5 kHz, typically at about 6–9 kHz. Thus, [s] creates an inherently "bright", i.e. high-pitched impression.

So, if [s] occurs in a high-F0 context, for example, after a high-rising phrase-final intonation, and if listeners interpret a part of the high aperiodic pitch impression of [s] as belonging to the high-F0 context, then they will "subtract" this part from the fricative. In consequence, [s] segments in high-F0 context will become more [ʃ]-like, see the illustration in Fig. 2. The same perceptual process should also occur in the opposite direction. That is, if [ʃ] occurs in a low-F0 context, for example, after a low-falling phrase-final intonation, and if listeners interpret a part of the low aperiodic pitch impression of [ʃ] as

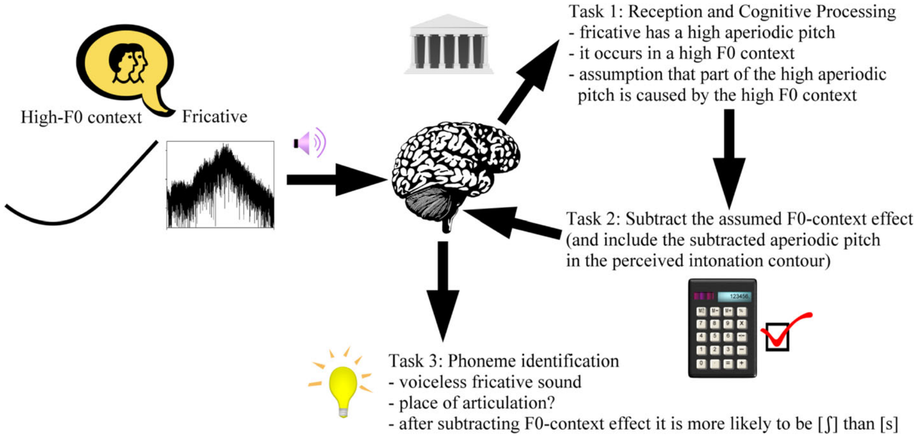

High-F0 context
Fricative
Task 1: Reception and Cognitive Processing
- fricative has a high aperiodic pitch
- it occurs in a high F0 context
- assumption that part of the high aperiodic pitch is caused by the high F0 context
Task 2: Subtract the assumed F0-context effect (and include the subtracted aperiodic pitch in the perceived intonation contour)
Task 3: Phoneme identification
- voiceless fricative sound
- place of articulation?
- after subtracting F0-context effect it is more likely to be [ʃ] than [s]

Fig. 2 Schematic illustration of the "subtraction" idea and the corresponding top-down cognitive speech processing mechanism that underlie the hypotheses 2 (a) and (b) and 3 of the present study. See, e.g. Ohala and Feder [37], Harrington et al. [39], and Kleber et al. [40] for related work and findings on the perceptual compensation of coarticulation

belonging to the low-F0 context, then they will "subtract" this part from the fricative. In consequence, [ʃ] segments in low-F0 context will become more [s]-like. In quantitative terms, a high F0 context could make listeners interpret a larger number of [s]-like sibilants as higher-pitched realizations of [ʃ]. Therefore, [s] responses could decrease in favour of [ʃ] responses. A low F0 context would have the opposite effect and cause [ʃ] responses to decrease in favour of [s] responses.

An acoustic [ʃ]-to-[s] continuum—attached to the ends of minimal-pair target words—is used to test this assumption. The decisive advantage of the applied experimental method (which is described in detail in the "Method" section) is that it allows us to gain insights into the perception of segmental intonation by means of a simple and natural word-identification task, i.e. without forcing the listeners into any laboratory exercises or giving them any metalinguistic or explicitly pitch-related instructions. Moreover, evidence in favour of the described subtractions would additionally suggest (in terms of the reassignment of subtracted information) that listeners include the subtracted parts of the sibilants' aperiodic pitch impressions in their perception of that domain of the speech signal that caused the sibilants' contextual variation: the F0 or intonation contour.

On this basis, the following hypotheses are tested:

- 1. *Sibilant effect*: The [ʃ]-[s]-continuum leads to a change in phoneme identification from /ʃ/ to /s/, which, in turn, causes a corresponding change in target word identification.
- 2. *F0 Context effect*:
- (a) If listeners perceive aperiodic pitch impressions of fricatives and relate them to the F0 context, then sibilant identification is altered by the F0 context: In a high F0 context, a larger number [s]-like sibilants are interpreted as higher-pitched realizations of [ʃ] so that [s] responses decrease in favor of [ʃ] responses. In a low F0 context, a larger number [ʃ]-like sibilants are interpreted as lower-pitched realizations of [s], which makes [ʃ] responses decrease in favor of [s] responses.
- (b) As a further consequence of (a), the high F0 context causes a phoneme boundary shift within the [ʃ]-[s]-continuum in favor of [ʃ], i.e. the phoneme boundary moves closer to the [s]-end of the continuum. In a low F0 context, the phoneme boundary within the [ʃ]-[s]-continuum is shifted in favor of [s], i.e. it moves closer to the [ʃ]-end of the continuum.
- 3. *Interaction of sibilant effect and F0 context effect*: In view of the general "subtraction" mechanism in Fig. 2, and in parallel to the compensatory effect in the perception of coarticulatory variation, it is expected that the F0-context effect is stronger for more ambiguous sibilant tokens at the center of the sibilant continuum than for clear instances of [ʃ] and [s] at the periphery of the continuum.

There are a number of previous studies whose results have shown that the identification of fricative phonemes like /ʃ/ and /s/ is basically affected by context or signal-external factors like adjacent vowels, speaker gender, visual cues, and even perceived sexual orientation [53–55]. The present study contributes to this research area insofar as it addresses a further context factor: intonation. However, the present study is *not* within the same line of research. It is not the primary aim of the study to advance our understanding of the context-dependent perceptual distinction of phonemic fricative contrasts. Rather, the identification of fricative phonemes is merely used here as a means towards testing whether speakers are sensitive to the aperiodic pitch impressions of fricatives and relate them to the F0 context in normally voiced utterances.

2 Method

2.1 The sibilant continuum

The basis of the perception experiment was a sibilant continuum. Its endpoints as well as all intermediate realizations were produced naturally by a trained female phonetician, SB. She produced several hundred isolated sibilants, while varying the following:

- The place of articulation and tongue shape from apical narrow-grooved alveolar to wide-grooved laminal postalveolar
- The degree of lip rounding from protruded and rounded to strongly spread lips (cf. [52], for a description of sibilant articulation in German).

SB's productions were recorded digitally at 48 kHz, 24-bit, in a sound-treated recording booth at Kiel University.

In a subsequent step, SB and the author together selected 14 sibilants on an auditory basis and arranged them to a continuum from a clear [ʃ] starting point through 10 auditorily similarly sized steps to a clear [s] endpoint. Soundness and integrity of the auditorily assembled sibilant continuum were checked by two independent measures.

First, we measured the average center-of-gravity (CoG) values of each of the sibilants within a frequency range from 1 to 20 kHz. As is shown in Table 1, the acoustic measurements support the auditory selection. CoG values consistently increase throughout the continuum. Moreover, the individual CoG increases between adjacent sibilants in the continuum are all in the same order of magnitude, i.e. between about 600 to 800 Hz. Overall, the CoG values cover a frequency range of 8–9 kHz, which is in the same order to magnitude as for the sibilant continuum described in Boersma and Hamann [56]. Furthermore, all sibilants had about the same overall duration of about 350 ms. There were small differences in overall duration of <10%, i.e. between 10 and 20 ms. However, such small differences fall below the perceptual difference limen, which is, for stimulus durations of 300 ms and more, at about 25–30 ms according to empirically based estimations of Lehiste [57] or Klatt and Cooper [58]. Other studies refer to the perceptual difference limen as the just noticeable difference (JND), cf. Huggins [59]. The threshold value reported by Huggins is in the same order of magnitude as those of Lehiste [57] and Klatt and Cooper [58]. In addition, Huggins stresses that the threshold (JND) is even higher for changes in consonant than in vowel duration. This lends further support to the assumption that participants of the present experiment were unable to perceive the small duration differences between the individual sibilants of the arranged sibilant continuum.

The acoustic-energy levels of all 14 sibilants were normalized to 30% of the maximum signal amplitude using Abode Audition (http://www.adobe.com/de/products/audition.html). This was to level out naturally produced acoustic-energy differences (that could have an uncontrolled effect on aperiodic pitch perception, cf. [60]). In addition, it ensured that all sibilants could be seamlessly integrated into the final stimulus utterances.

Second, soundness and integrity of the sibilant continuum were additionally tested in a perception experiment, using the established combination of 2AFC phoneme identification and AXB discrimination tasks that are well known from categorical speech perception (cf. [61], and [62], for summarizes of this experimental paradigm). The AXB test was preferred over the "prototypical ABX discrimination test" ([63]; 364) as studies comparing various discrimination tasks showed that the AXB variant represents an "economic variant of 4IAX" in which listeners are particularly sensitive to stimulus differences [63, 64]. In addition, unlike the ABX design, the AXB design allows listeners to perceive and judge X in immediate adjacency to both A and B. This avoids the well known and empirically supported issue of ABX tests that listeners compare "only B and X" ([65]; 42) and largely disregard A.

Ten untrained native speakers of Northern Standard German (seven females and three males, between 20 and 30 years old) took part in the perception experiment. The phoneme identification test presented each of the 14 sibilants five times in an overall randomized order. The listeners were to classify the 70 sibilants with reference to their corresponding German graphemes as either

Table 1 Mean center-of-gravity (CoG) values measured in the 14 steps of the naturally produced [ʃ]-to-[s] sibilant continuum

Sibilant number	Mean CoG (Hz)	Sibilant number	Mean CoG (Hz)	Sibilant number	Mean CoG (Hz)
1	4439	6 (4)	8042	10 (8)	10,633
2	5108	7 (5)	8686	11 (9)	11,215
3 (1)	5820	8 (6)	9423	12 (10)	11,844
4 (2)	6691	9 (7)	10,029	13	12,633
5 (3)	7368			14	13,337

The numbers in brackets refer to the sibilant and stimulus numbers in the actual experiment

<s> or <sch>, the latter being the grapheme sequence of [ʃ]. In the AXB task, A and B were represented by the two endpoints of the continuum, i.e. sibilants 1 and 14 in Table 1. The X slot was filled by all 14 sibilants. Thus, we also included AAB and ABB triplets. Listeners were to decide whether X was more similar to A or B. All AXB triplets were presented ten times, five times with A = sibilant 1 and B = sibilant 14, and five times with the inverted reference frame, i.e. A = sibilant 14 and B = sibilant 1. The total 140 triplets were presented—with one half-time break—also in an overall randomized order.

Half of the ten listeners started with the phoneme identification task, the other half with the AXB discrimination task. Judgments as well as reaction times were recorded. The experiments were conducted by means of PRAAT-MFC scripts (cf. [66]) on individual desktop PCs (using PRAAT v5.4, cf. [67], see also http://www.fon.hum.uva.nl/praat/). Listeners wore HiFi headphones and judged the stimuli at a constant, pre-adjusted loudness level in a silent, sound-treated lecture room at Kiel University. Note that conducting the experiment on different desktop PCs did not introduce any variability into the reaction-time measurements, not least because the same PRAAT version (5.4) was used on all computers.

The results of the experiment are summarized in Fig. 3a, b. The Figures show percentages pooled across listeners, stimulus repetitions, and, in the case of the discrimination test, also the two A-B orders. The key findings are as follows: The gray identification bars in Fig. 3a take an ogee-like shape across the ascending sibilant numbers. Sibilants 1–5 yielded 98–100% [ʃ] responses. Sibilants 10–14 yielded 100% [s] responses. Sibilants 6–9 created a *unidirectional* judgment change from [ʃ] to [s]. The finding that [s] and [ʃ] can be identified by listeners without any vowel context is consistent with the early perception experiment on American English fricative identification by Harris [68]. Reaction times are higher in the phoneme transition section than in the two sections of clear phoneme identification. The discrimination results in Fig. 3b are consistent with the identification results. That is, the change in the discriminative comparison of X with A and B occurs for the same center stimuli (6–9) as the change in identification; and like the latter, the change in discrimination is also *unidirectional* and coincides with a clear increase in reaction times.

In summary, the experiments clearly agree with the acoustic CoG measurements and provide strong additional evidence for the soundness and integrity of the sibilant continuum. That is, the two endpoints of the continuum are unambiguous instances of [ʃ] (stimulus 1) and [s] (stimulus 14). The stimuli in between produced a *unidirectional* change from [ʃ] to [s]. Unidirectionality is important, as it proves that the consistent acoustic continuum embodies consistent auditory and behavioral changes. The change from [ʃ] to [s] is gradual in the acoustic domain and to a certain extent more abrupt in the perceptual domain. Note that it is irrelevant for the present study and beyond the question addressed here, whether or not the perceptual [ʃ]-to-[s] transition in the sibilant continuum can actually be considered categorical. Hypothesis 2 (b) only required determining a phoneme boundary. We defined this boundary as the 50% crossing in the identification results, which is the standard criterion in psychophonetic studies (cf. [40, 62, 69–71]). Based on this criterion, and for the sibilants presented in isolation, the phoneme boundary from [ʃ] to [s] was located somewhere between stimuli 8 and 9 in the continuum.

The section with the more ambiguously identified sibilants of the 14-step continuum extends approximately from stimuli 6–9. With reference to hypothesis 3, we expected that this section of the continuum would be primarily relevant for an effect of the F0 context on sibilant

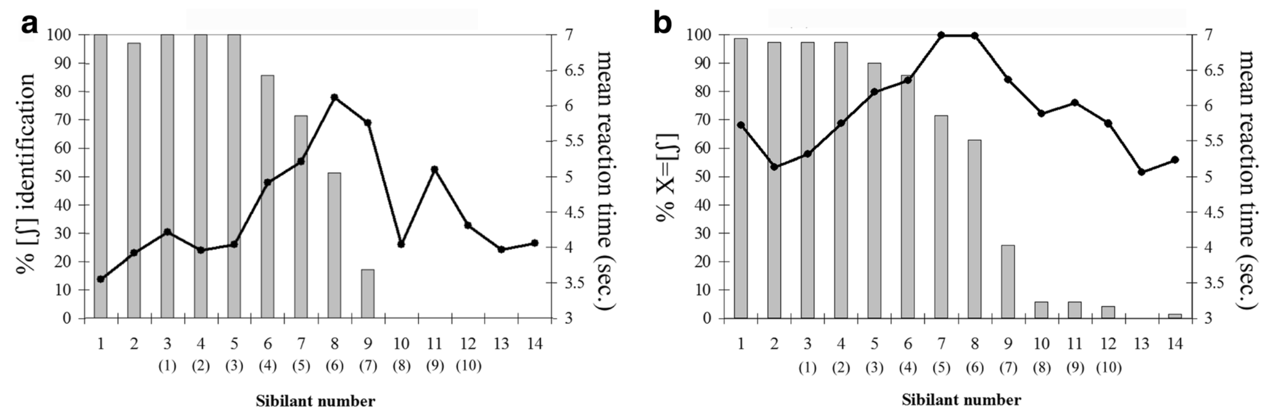


Fig. 3 Results of the identification test (**a**) and discrimination test (**b**) pooled across all ten listeners and five stimulus repetitions. *Bars* indicate percentages, *solid lines* show mean reaction times. Each value in **a** represents 50 responses and each value in **b** 100 responses (due to five repetitions in the AXB and BXA order). The numbers in brackets refer to the sibilant and stimulus numbers in the actual experiment

identification. For this reason, we decided to maintain this critical section, but to downsize the continuum outside this section by two steps/sibilants, in order to reduce the number of stimuli in the main experiment. The first and last two sibilants of the continuum, i.e. sibilants 1–2 and 13–14, were removed, as sibilants 3 and 12 proved to be equally clear representatives of [s] and [ʃ] in the identification test. Figure 4a displays the final 10-step sibilant continuum in terms of the individual sibilants' spectrograms. The spectral patterns of the new endpoints 1 and 10 of the continuum (i.e. the original sibilants 3 and 12, see Table 1) are depicted in more detail in Fig. 4b in the form of spectrograms (upper panel) and average spectra with a frequency range from 0 to 20,000 Hz. The acoustic-energy distributions and ranges in Fig. 4b show the "very pronounced spectral energy peaks" of [s] and [ʃ] highlighted by Wagner et al. [7]. The two spectra additionally illustrate very clearly why the [ʃ] creates a lower and [s] a higher aperiodic pitch impression: The lower spectral energy boundary is located at about 2300 Hz for [ʃ] but more than twice as high (between 5550 and 6000 Hz) for [s]. Then, for [ʃ], the first energy peak occurs right above the lower spectral energy boundary at about 3500 Hz, whereas the first (and only) spectral energy peaks of [s] are between 10,000 and 13,000 Hz. Finally, both [s] and [ʃ] show a clear decline of acoustic noise energy again above about 17,000 Hz. However, this decline is considerably stronger and steeper for [ʃ] than for [s].

2.2 Stimulus generation

2.2.1 Sentences and their F0 manipulation

The stimulus utterances consist of three pairs of syntactically marked question sentences that were also produced by the female speaker SB. The sentences are listed in Table 2. They contained monosyllabic target words in sentence-final position. These words—all of them nouns—differed in two ways. First, they formed minimal pairs, which ended in either [ʃ] or [s]. Second, the sibilants were preceded by one of three different phonologically short vowels, i.e. [ɪ], [a], or [ʊ]. We used short vowels, as they leave less room for F0 movements than long vowels. The idea was that shorter F0-based pitch sections in combination with probably stronger truncations of F0 movements by the target sibilants (cf. [24, 72]) would create ideal conditions for perceiving and processing the aperiodic pitch impressions of the target sibilants in terms of hypothesis 2 (a) and (b).

All target words are common words and part of the active vocabulary of native speakers of Northern Standard German; although, they occur with different frequencies in everyday conversations. For example, *Fisch*, *Pass*, and *Bus* are among the 10,000 most frequent German words, whereas *Viss*, *Pasch*, and *Busch* are not [73]. However, note that these frequency differences could, if at all, only cause a bias (cf. [74]) in the form of differences *between* the individual sentence pairs (i.e. vowel quality conditions). The decisive F0 context effects on target-word identification *within* each sentence pair are neither suppressed nor in other ways affected by word-frequency differences. Therefore, we can ignore the factor word frequency for the purposes of the present study.

SB produced the 3 × 2 sentences with a phonologically constant intonation pattern. It consisted of a prenuclear high-tone pitch accent on the sentence-initial verb, followed by a slightly dipped F0 transition that led over to a nuclear high-tone pitch accent on the sentence-final

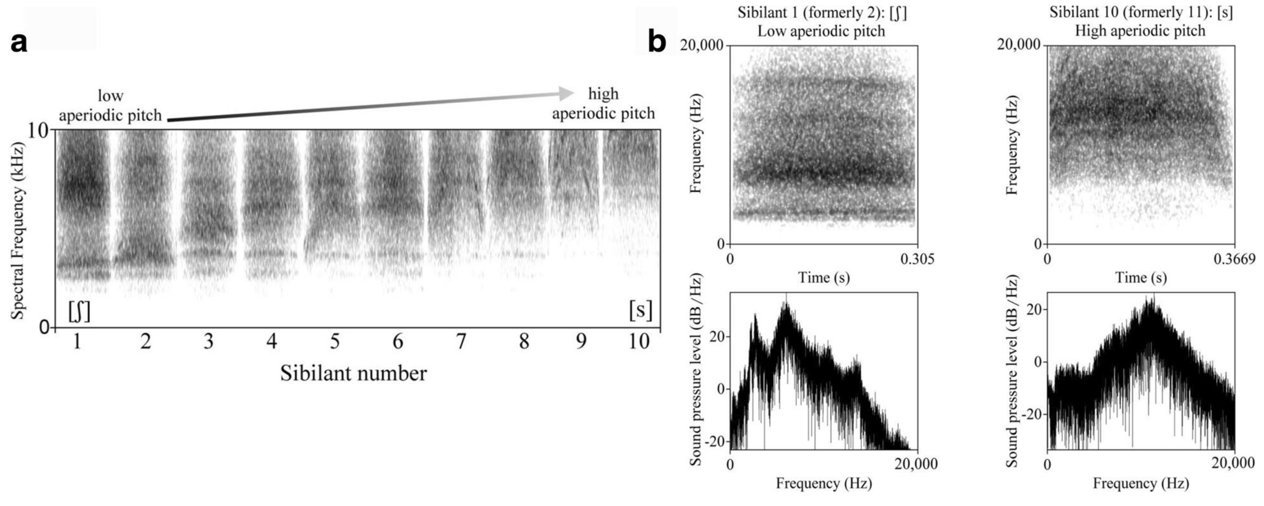


Fig. 4 a Spectral characteristics (0–10 kHz) of the ten sibilants in the sibilant continuum from [ʃ] to [s] that provides the basis of the stimuli in the main experiment. The sibilants correspond to tokens 3–12 of the original sibilant continuum; **b** more detailed frequency analysis of the two sibilants 1 and 10 (formerly 3 and 12) at the endpoints of the continuum. The *upper panel* shows spectrograms and the *lower panel* spectra (0–20 kHz) of the two sibilants

Table 2 The three pairs of stimulus sentences used in the perception experiment

Vowel context (phonolg. short)	Question sentence	English translation	Target word and sibilant
[ɪ]	"Kaufst Du noch Viss?"	Will you buy Viss? (a well-known German detergent brand)	[fɪs]
	"Kaufst Du noch Fisch?"	Will you buy fish?	[fɪʃ]
[a]	"Hast Du einen Pass?"	Do you have a passport?	[p^{h}as]
	"Hast Du einen Pasch?"	Did you throw a doublet? (with dices)	[p^{h}aʃ]
[ʊ]	"Siehst Du den Bus?"	Do you see the bus?	[ʊs]
	"Siehst Du den Busch?"	Do you see the bush?	[ʊʃ]

The pairs of sentences differ in the final target word or, more specifically, in the sentence-final sibilant. The experimentally relevant difference between the pairs concerns the short vowel preceding the sibilant. Underlining indicates nuclear pitch accents (solid lines) on the sentence-final target words and prenuclear pitch accents (dashed lines) on the sentence-initial verbs

target word. The nuclear-accent pattern ended in an intermediate mid-low fall that is typical of a list intonation (cf. [62]), see the dotted gray lines in Fig. 5a, b. This mid-level intonation was important to avoid resynthesis artefacts in the subsequent PSOLA-based F0 manipulation by which the naturally produced sentences were turned into the final experimental stimuli. The F0 manipulation was conducted in with PRAAT (version 5.4; cf. [67], see also http://www.fon.hum.uva.nl/praat/) in two successive steps.

In the first step, the naturally produced sentence intonations were replaced by two diametrically opposed tunes, stylized at eight contour points. Both tunes are equally compatible with syntactically marked questions in German. They just change the attitudinal meaning of the question (cf. [75]).

As is illustrated by the "Bus" sentence in Fig. 5a, b, the first part up to the prenuclear accent was left almost unchanged in both created tunes. But then, the tune in (a) continued with a longer, deeper dip towards a low nuclear pitch accent at 188 Hz. The nuclear accent was followed by a steep, high-rising phrase-final intonation. It reached a F0 level of 450 Hz before the sentence-final target sibilant. By contrast, F0 in tune (b) continued to rise after the prenuclear accent, reaching a maximum of 302 Hz right before the target word. From there, F0 changed into a steep fall throughout the target word, first to a level of 170 Hz at the low tonal target of the nuclear pitch accent, and then further through the phrase-final movement to a low terminal level of only 136 Hz before the sentence-final target sibilant.

In summary, the tune in Fig. 5a created a *high F0 context* before the target sibilant, whereas the tune in Fig. 5b embedded the target sibilant into a *low F0 context*. All six question sentences in Table 2 were resynthesized with tunes (a) and (b). In the experiment, the tunes constituted the two-level independent variable "F0 Context".

2.2.2 Splicing-in the sibilant continuum

The sentence-final [ʃ] and [s] sibilants were cut off from the resynthesized question sentences. The cutoff point was the onset of frication, which, in all cases, roughly coincided with the end of voicing. That is, we never removed more than 10–20 ms of voicing from the ends of the sentences. Then, the cutoff sibilants were replaced by all 10 sibilants of the sibilant continuum in Fig. 4. The 10 replacements formed the second independent variable "final sibilant".

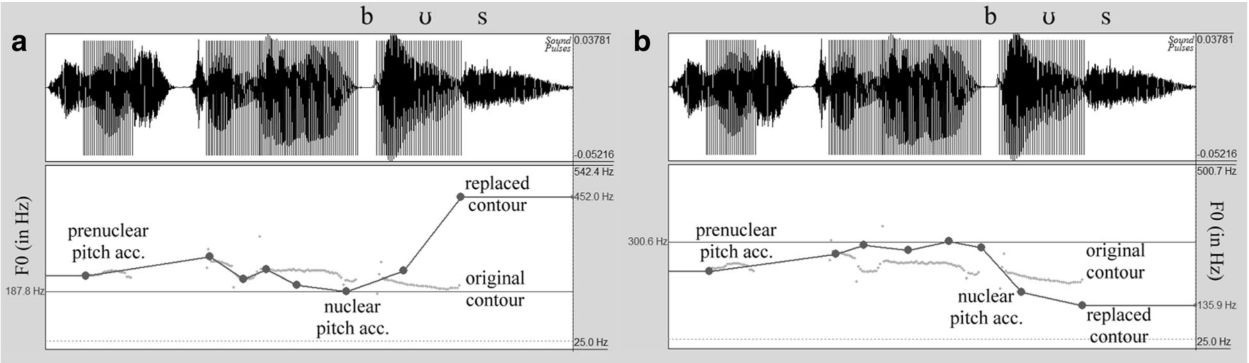


Fig. 5 Edited PRAAT screenshots of "Siehst Du den Bus?" (Do you see the bus?) showing the original (continuous) F0 contour (*gray dotted line*) and the two stylized tunes (*dark solid lines*)—rising to a high level in **a** and falling to a low level in **b**—that replaced the original F0 contour in each stimulus sentence

2.2.3 Stimulus number and independent control variables

The two consecutive steps of the stimulus generation described in sections 2.2.1 and 2.2.2 resulted in a total number of 120 stimuli. They represented the two key variables F0 context and final sibilant. However, in view of the stimulus sentences and the splicing procedure, our listeners' responses could additionally reflect two further context factors whose effects on [s] and [ʃ] perception are well documented: vowel quality and original sibilant.

As regards vowel quality, Mann and Repp [53] showed for a (synthetic) sibilant continuum between [ʃ] and [s] that listeners perceive considerably less instances of [ʃ] in [u] than in [a] contexts. The same effect occurred for a change from [a] to [i] in the study of Clayards et al. [76]. It is reasonable to assume that the present stimuli replicate these effects. That is, it is likely that [ʃ] perception increases at the expense of [s] perception from [ʊ] through [a] to [ɪ]. Furthermore, there is cross-linguistic evidence that the [ʃ]-vs-[s] contrast extends quite far into the preceding vowel, thus creating strong coarticulatory cues that are also perceptually relevant, particularly with respect to F2 [77–79]. It should therefore not be ruled out (despite the experimental evidence of [7]) that the original cutoff sibilant creates, through its traces in the preceding vowel, a bias that pushes or pulls sibilant identification in the created stimuli towards the cut off /s/ or /ʃ/ phoneme.

So, in order to properly represent and interpret listeners' responses and separate the effects of the two key independent variables final sibilant and F0 context from those of the two context factors above, the context factors were included in the statistical analysis of the perception data as control variables. "Vowel quality" was a three-level and "original sibilant" a two-level independent variable.

2.3 Conducting the experiment

The experiment was conducted on the basis of a PRAAT-MFC script (written by SB). It played each of the 120 stimuli three times in an overall randomized order, avoiding stimulus doublets. The randomization was different for each participant. Participants had the possibility to take a break after the 120th and the 240th stimulus.

The participants were 30 native speakers of Northern Standard German, 17 females and 13 males. They were between 19 and 23 years old and early-stage Bachelor students of Empirical Linguistics at Kiel University without any known hearing or speaking disorders. The early-stage Bachelor students were naive insofar as none them had taken part in a perception experiment before nor had anyone of them enrolled for lectures in prosody or speech acoustics.

The participants conducted the experiment on individual desktop PCs in a sound-treated lecture room at Kiel University. They listened to the stimuli via HiFi headphones at individual paces. The participants' task was simply to listen carefully to each of the stimuli and specify afterwards which word they had heard at the ends of the question sentences. They were informed that the sentence-final words would either be *Viss* or *Fisch*, *Pass* or *Pasch*, or *Bus* or *Busch*. Judgments were made on a 2AFC basis. To that end, each stimulus was presented in combination with two pictures on the PC screen. The two pictures represented the corresponding pair of target words. Figure 6a–c shows the three pairs of pictures that were used. The display in Fig. 6a was presented for all stimuli of vowel quality condition [ɪ]; Fig. 6b was combined with all stimuli of vowel quality condition [a]; and Fig. 6c was used to guide and elicit the listeners' judgments for the stimuli of vowel quality condition [ʊ]. Each display appeared after the end of the acoustic stimulus in order to avoid any cross-modal priming effects (cf. [80]). Participants were asked to click as soon as possible on the picture of the word that they identified in the question sentence or to guess if they were undecided. The order of the pictures shown on the screen was counterbalanced across the 30 participants. That is, for half of the participants, the left pictures showed *Viss*, *Pass*, and *Bus*; and for the other half, they showed *Fisch*, *Pasch*, and *Busch*.

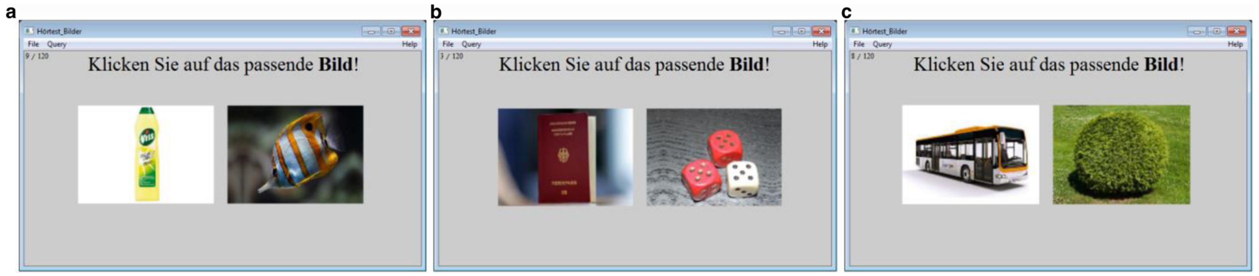
Reaction-time measurements were taken in combination with the word-identification judgments. Like in the continuum-validation experiments before, the use of different desktop PCs did not introduce any variability into the reaction-time measurements of the main experiment.

Conducting the experiment took about 25–30 min, including instructions, breaks, and a short concluding interview. The interview was based on a questionnaire that each participant was asked to fill out.

3 Results

3.1 Questionnaires

As regards the questionnaires, all but three participants assumed the stimuli to be naturally produced speech, without any manipulation. The other three participants did also not stumble upon any manipulation artefacts in the stimuli or the stimuli's quality in general. They judged the stimuli to be naturally sounding, but additionally stated that the stimuli were, in their ears, too uniform to be unmodified. The average naturalness score of the stimuli on a five-point scale from "artificial non-human-like speech" (1) to "natural everyday speech" (5) was 4.1. Furthermore, all participants rated the task either as "easy" or "very easy" on a five-point scale from "very difficult" (1) to "very easy" (5). The average score was 4.3. Both the perceived naturalness of the stimulus



Klicken Sie auf das passende **Bild**!

Fig. 6 Original screenshots of the three types of response displays presented to the listeners in the main experiment. Listeners identified the stimulus-final target words by clicking on the pictures for *Viss* or *Fish* (**a**), *Pass* or *Pasch* (**b**), or *Bus* or *Busch* (**c**)

sentences and the straightforward applicability of sour task underpin the validity and reliability of the actual response data, the results of which are summarized below.

3.2 Word-identification judgments

The categorical response data were statistically analyzed by means of a binomial logistic regression analysis (cf. [81]). It was conducted using the lme4 package in R [82]. The link function was logit. The logistic regression analysis was based on the four fixed effects F0 context (reference: low F0), final sibilant (continuous variable), vowel quality (reference: [a]), and original sibilant (reference: [ʃ]). Participants were included as a random-effects variable. The dependent (predictor) variable was the identification of the target words as ending in either [ʃ] or [s] per participant, stimulus, and stimulus repetition. Note that it was considered very unlikely that stimulus repetition would have a separate significant main effect on target-word identification, as the stimuli were presented to participants in individually randomized orders. Yet, following the suggestion of one of our reviewers, we conducted a further logistic regression analysis, this time with stimulus repetition included as a separate fixed effect. As was expected, this further analysis yielded no significant main effect of stimulus repetition (coeff. estimate (β) = 0.03, Std. error = 0.05, z value = 0.74, p = 0.458). The model's AIC was also slightly higher (3480.4) than for the model without stimulus repetition as a separate fixed factor. Furthermore, an analysis that included stimulus repetition in random slope per participant did not converge. Therefore, our results summary in Table 3 is based on the logistic regression analysis without stimulus repetition as a separate predictor.

The logistic regression analysis yielded highly significant main effects of the two main predictors F0 context and final sibilant. Moreover, there was a significant main effect of the control predictor vowel quality. The reference vowel [a] differed from both [ɪ] and [ʊ], with the former difference being slightly stronger than the latter (see βs in Table 3). The other control predictor, original sibilant, did not reach significance. In addition to the significant main effects, there were two significant two-way interactions; a weaker one between final sibilant and vowel quality ([a]/[ɪ] and [a]/[ʊ]), and a stronger one between final sibilant and F0 context (see again βs in Table 3).

Further details of the overall results pattern are displayed in Fig. 7a–c. The three figures show the percentages of [ʃ] identification, i.e. of (a) *Fisch*, (b) *Pasch*, and (c) *Busch* identification, respectively. The non-significant factor original sibilant was omitted, and the responses for its two factor levels were conflated in the figures. The resulting response frequencies were pooled across all participants and stimulus repetitions. Thus, each light and dark gray bar in Fig. 7a–c represents 180 judgments.

The black dotted lines show that the significant main effect of final sibilant manifests itself in all three panels of Fig. 7 as a change from clear [ʃ] to [s] identifications (e.g. from *Busch* to *Bus*) through a short transition section of more ambiguous sibilant identifications. This tripartition of the sibilant continuum into two perceptually steady parts and a transition part is supported by additional inferential statistics. Multiple paired sample t tests were carried out to compare the frequency of [ʃ] responses between all neighboring sibilants (i.e. adjacent steps in the sibilant continuum) in each vowel quality condition (i.e. for each panel of Fig. 7); [ʃ] responses were pooled across the three stimulus repetitions and the two original sibilant conditions, and thus varied between 0 and 6 per participant (n = 30) and sibilant number. The t tests' p levels were adjusted according to the Benjamini-Hochberg procedure.

All significant differences (i.e. p levels <0.05) in [ʃ] frequency between adjacent steps in the sibilant continuum are restricted to sibilants 3–7. No significant frequency differences between two neighboring sibilants were found among sibilants 1–2 and 8–10.

The main effect of F0 context shows up in Fig. 7a–c as the differences between the light and dark bars in each panel. As can be seen, the dark bars are almost consistently higher than the light bars. That is, the high F0 context condition supported [ʃ] identification and/or the low F0 context condition supported [s] identification in

Table 3 Table of covariates, results summary of the binomial logistic regression analysis

Fixed effects	Coeff. estimates (*β*)	Std. error	z value	Pr(>|z|)
Intercept	4.25	0.24	17.26	<0.0001
Final sibilant	−0.99	0.03	−38.11	<0.0001
F0 context	0.69	0.09	7.91	<0.001
Vowel quality [ɪ]	1.61	0.34	4.70	<0.01
Vowel quality [ʊ]	−0.98	0.05	−19.88	<0.0001
Original sibilant	0.13	0.52	0.26	n.s.
Final sibilant × vowel quality [ɪ]	−0.12	0.06	−2.05	<0.05
Final sibilant × vowel quality [ʊ]	−0.32	0.13	−2.47	<0.05
Final sibilant × F0 context	0.56	0.15	3.73	<0.01
AIC 3478.9				

the sentence-final target words. More specifically, a change of the F0 context from high to low or vice versa before the final target sibilant was able to shift word-identification judgments by up to 30%. To put this value into context, note the following: Each of the vowel quality conditions [ɪ] and [a] in Fig. 7a, b contains one stimulus for which the word identification judgments in high and low F0 contexts fall on opposite sides of the 50% [ʃ]-[s] boundary. Thus, whether one and the same target word was predominantly identified as either *Fisch*/*Pasch* (ending in [ʃ]) or *Viss*/*Pass* (ending in [s]) was for some of the present stimuli *only* determined by whether F0 ended high or low before the final sibilant.

The logistic regression curves calculated for the statistical analyses in Table 3 were further used to determine the exact 50% [ʃ]-[s] phoneme boundary in the two F0 context conditions and hence the extent to which the F0 context was able to shift this boundary in each vowel quality condition. In other words, we quantified by means of logistic regression the differences between the light and dark gray bars in terms of the 50% phoneme boundary for each panel of Fig. 7. The stimulus located at the 50% boundary would have a probability of 0.5 to be identified as ending in [ʃ] and a probability of 1–0.5 to be identified as ending in [s]. Thus,

$$ln\frac{0.5}{1\text{–}0.5} = \beta_0 + \beta_1 x$$

with β_0 and β_1 being the logistic regression coefficients and with $x = -\frac{\beta_0}{\beta_1}$.

The results of the phoneme boundary estimation are summarized in Table 4. As can be seen, the difference between the high and low F0 context shifted the 50% boundary in the [ʃ]-[s] sibilant continuum on average by about one complete continuum step in all three vowel quality conditions. These boundary shifts are all statistically significant according to paired sample *t* tests that compared, for each vowel quality, the estimated phoneme boundary locations (in terms of sibilant numbers) of the 30 participants between the two F0 contexts.

The main effect of vowel quality can be observed across the three panels of Fig. 7. Although the perceptual change from [ʃ] to [s] identification occurs under each vowel condition, the overall frequency of [ʃ] identification decreases successively from [ɪ] through [a] to [ʊ]. That is, all other variables being constant, participants heard more instances of *Fisch* than of *Pasch* and more instances of *Pasch* than of *Busch*. This vowel-specific perceptual organization of the sibilant continuum also involves a shift of the 50% phoneme boundary. As in the case of the F0 context effect, logistic regression curves were calculated to estimate for each vowel quality condition the exact [ʃ]-to-[s] cross-over point in the sibilant continuum. As is indicated in Fig. 7, this 50% point is located close to sibilant number 6 in the [ɪ] condition (5.89), and it moves close to sibilant number 5 in the [a] condition (5.07). In the [ʊ] condition, it is further shifted in favour of [s] and located almost in the middle of sibilants 4 and 5 (4.61). *t* tests for paired samples proved the significance of this shift in phoneme boundary location for both comparisons [ɪ] vs. [a] ($t[29] = 10.44$, $p < 0.001$) and [a] vs. [ʊ] ($t[29] = 6.85$, $p < 0.001$).

Finally, Fig. 7a–c show that both the effect of vowel quality and the effect of F0 context were clearly stronger in the judgment transition section at the center of the sibilant continuum. This fact manifests itself in the significant interactions between final sibilant and vowel quality and final sibilant and F0 context of the logistic regression analysis in Table 3.

3.3 Reaction times

In order to supplement the findings on target-word or sibilant identification, the participants' identification judgments were also analyzed with respect to their corresponding reaction times. Prior to conducting any statistical tests on these reaction times, they were normalized such that their measurement consistently started at the onset of the final sibilant, independently of target-sentence and target-word durations. Normalized reaction

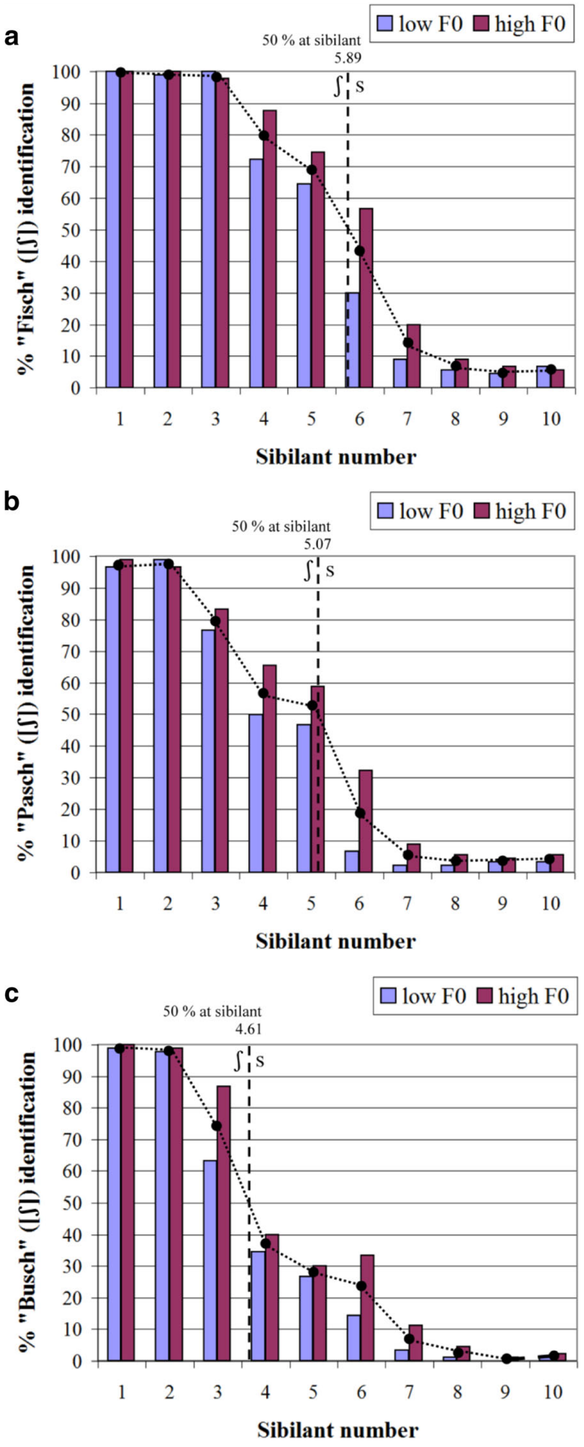


Fig. 7 Summary of the word-identification judgments. The three panels (**a**–**c**) show the relative frequencies of sentence-final target words being perceived as ending in [S] (i.e. *Fisch*, *Pasch*, *Busch*) in the rising (*dark bars*) and falling (*light bars*) F0 contexts; *n* = 180 for each bar. *Dotted lines* represent the average percentages across rising and falling F0 contexts and vertical dashed lines show the 50% cross-over point between [ʃ] and [s] identification, estimated by means of logistic regression curves

times ranged from 378 to 2190 ms. The average reaction time across all participants was 721 ms. In addition, skewness and kurtosis of the overall reaction-time distribution were checked, using the moments package of R. The check took into account that previous perception experiments often reported major deviations of reaction time curves from the normal distribution [83]. The values that characterize the reaction-time distribution in the present study clearly differ from zero for both skewness (−0.5) and kurtosis (2.45) and show that the distribution is slightly platykurtic and moderately skewed. However, although skewness and kurtosis are close to the limits for statistical tests that assume normally distributed data, they are both still within the commonly accepted ranges (cf. [84]).

On this basis, the reaction-time data were submitted to a linear mixed-effects model using R's lme4 package. Like in the binomial logistic regression analysis of the main judgment data, fixed effects were F0 context, final sibilant, vowel quality, and original sibilant. Stimulus repetition was used as a further fixed effect. Participants were included as a random-effects variable. Degrees-of-freedom (*df*, for effect-specific *p* value estimation) were computed from the model by means of the conservative Kenward-Roger approximation. The resulting model is summarized in Table 5.

Analogous to the results of the judgment data, the reaction-time analysis yielded significant main effects of final sibilant and F0 context, as well as a significant interaction between them. All other effects and interactions were non-significant. Therefore, the reaction-time results can be related to the sibilant continuum and its key predictors alone, without the need to take into account any of the control predictors.

The effect of final sibilant is due to an increase in reaction times at the center of the sibilant continuum, just as in Fig. 3 above. Furthermore, in the center of the sibilant continuum, reaction times were longer when sibilants predominantly identified as [ʃ] were perceived in a low-F0 context, and when sibilants predominantly identified as [s] were perceived in a high-F0 context. These additional increases in reaction time at the center of the sibilant continuum caused the F0-context effect, and their restriction to the center is responsible for the interaction of final sibilant and F0 context.

4 Discussion and conclusion

4.1 Hypotheses

Target-word identification patterns showed that listeners organized the 10-step sibilant continuum into three sections. Sibilants 1–2 were identified as instances of [ʃ] in more than 95% of all target words. Sibilants 8–10 were identified as instances of [s] in about 90% or more of all target words. There were no significant differences between the target word identification frequencies among

Table 4 Mean 50% phoneme boundary locations in the [ʃ]-[s] sibilant continuum, estimated by means of logistic regression curves derived from response patterns in the high and low F0 contexts of each vowel quality condition

Vowel quality	Mean 50% [ʃ]-[s] boundary location in high F0 context	Mean 50% [ʃ]-[s] boundary location in low F0 context	Mean [ʃ]-[s] boundary shift high vs. low F0	Test statistics of paired sample *t* test
[ɪ] *Fisch/Viss*	6.41	5.35	1.06	$t[29] = 4.83, p < 0.001$
[a] *Pasch/Pass*	5.63	4.50	1.13	$t[29] = 6.22, p < 0.001$
[ʊ] *Busch/Bus*	5.24	3.99	1.25	$t[29] = 7.95, p < 0.001$

sibilants 1–2 as well as among sibilants 8–10. That is, sibilants 1–2 and 8–10 triggered clear and constant [ʃ] or [s] identifications and thus represented perceptually steady continuum sections. Sibilants 3–7 constituted a transition section that linked the two steady sections of the sibilant continuum. Almost every upward step from sibilant 3 to 7 in the transition section caused a significant decrease in [ʃ] identification. Therefore, the final sibilant effect predicted in hypothesis 1 is clearly borne out by the response data: The [ʃ]-[s]-continuum led to a change in phoneme identification from /ʃ/ to /s/, which, in turn, caused a corresponding change in target word identification.

Target-word identification patterns also included a strong and significant effect of F0 context. This effect was exactly as predicted by hypothesis 2 (a): Sibilants following the high F0 context were more frequently interpreted as [ʃ], whereas sibilants following the low F0 context were more frequently perceived as [s]. An obvious explanation for this effect is that listeners did perceive the aperiodic pitch impressions caused by the sibilants and related them to the F0 context. That is, analogous to the previous findings on the perception of coarticulation and the "subtraction mechanism" sketched in Fig. 2, listeners expected sibilants to have higher aperiodic pitch levels in high-F0 contexts and hence interpreted more of the intrinsically higher pitched [s]-like sibilants in this context as higher pitched realizations of /ʃ/. Likewise, listeners expected sibilants to have lower aperiodic pitch levels in low-F0 contexts and hence interpreted more of the intrinsically lower pitched [ʃ]-like sibilants in this context as lower pitched realizations of /s/. Or, returning to the principle explained in section 1.3, listeners interpreted a part of the high aperiodic pitch impression of [s]-like sibilants as being due to the high-F0 context and "subtracted" this part from the fricative. In consequence, [s]-like sibilants in the high-F0 context became perceptually more [ʃ]-like. In the opposite direction, listeners interpreted a part of the low aperiodic pitch impression of [ʃ]-like sibilants as belonging to the low-F0 context and "subtracted" this part from the fricative. In consequence, [ʃ]-like sibilants in the low-F0 context became perceptually more [s]-like.

Table 5 Table of covariates, results summary of the linear mixed-effects model

Fixed effects	Coeff. estimates (*β*)	Std. error	*t* value	Pr(>|t|)
Intercept	745.46	11.23	66.40	<0.0001
Final sibilant	−22.18	9.76	−2.27	<0.01
F0 context	43.68	11.96	3.65	<0.001
Vowel quality [ɪ]	8.75	12.06	0.73	n.s.
Vowel quality [ʊ]	0.50	3.11	0.16	n.s.
Original sibilant	0.13	0.52	0.26	n.s.
Stimulus repetition	0.38	2.77	0.14	n.s.
Final sibilant × vowel quality [ɪ]	−0.86	1.93	−0.44	n.s.
Final sibilant × vowel quality [ʊ]	2.01	2.57	0.78	n.s.
Final sibilant × F0 context	7.34	1.57	4.68	<0.001

n.s. not significant

Following a reviewer's request, we point out in this context that the assumption of a "subtraction mechanism" which underlies speech perception and allows listeners to compensate for variation in sound segments is still compatible with variation-driven sound change. As is stressed by Ohala and Feder [37] and summarized by, for example, Kleber et al. [40], sound changes can occur despite a "subtraction mechanism" due to a misalignment between the production and perception of variation [85]. For example, L2 speakers of a language who are not entirely familiar with the coarticulatory patterns of that language cannot (fully) subtract the corresponding variation and hence still perceive the coarticulatory variation in speech sounds. If this variation is then reproduced, it creates a "minisound change" ([39]; 2826) that can spread and become conventionalized and phonologized under certain conditions. A further example of a misalignment between production and perception is subtracted phonetic variation that is erroneously reassigned to and perceptually integrated in the wrong speech unit, for example, the adjacent vowel instead of the adjacent consonant. The repeated mispronunciation of this vowel again represents a "minisound change". Furthermore, looking at the many small idiosyncratic differences in sound production and variation, and assuming that each speaker's "subtraction mechanism" is fed by and hence tuned to his/her own production patterns, it is likely that no speaker is consistently and fully able to subtract the variation in speech sounds coming from a different speaker [40]. Under these circumstances, "minisound changes" and subsequent sound changes are only a matter of frequency, salience, reproducibility, language/group needs, and, perhaps, phonotactics.

Hypothesis 2 (b) is also supported by the results. As a further consequence of the F0 context effect described above, the high F0 context additionally caused a significant phoneme boundary shift within the [ʃ]-[s]-continuum in favor of /ʃ/, i.e. the phoneme boundary moved closer to the [s]-end of the continuum. In a low F0 context, the phoneme boundary within the [ʃ]-[s]-continuum was significantly shifted in favor of /s/, i.e. it moved closer to the [ʃ]-end of the continuum. Overall, the F0 context effect was able to shift the boundary by about one step in the sibilant continuum. This may not seem a lot, but it is actually a considerable effect. Note that in languages like German, French, and English, the mean CoG difference between [ʃ] and [s] realizations is about 1500 to 2000 Hz [26, 28, 44–51]. One step in the sibilant continuum corresponds to a CoG difference of between 600 and 800 Hz. Thus, in terms of the CoG measure, the present effect of F0 context on sibilant perception and the sibilant boundary /ʃ/ and /s/ is about half as large as the natural difference between [ʃ] and [s] realizations in German and other languages. In fact, according Gordon et al. [46], there even seem to be quite a few languages whose CoG differences between /ʃand /s/ are on average in the same order of magnitude (600–800 Hz) as the F0-induced context effect and boundary shift found in the present study. This includes Gaelic, Hupa, Chickasaw, Western Apache, and a variety of Turkish. However, it might be that the mean differences in the study of Gordon et al. are only smaller than those of Niebuhr and colleagues because they measured CoG in a narrower frequency range. This possibility needs to be checked.

Finally, hypothesis 3 is consistent with the results as well. The F0-context effect was stronger for the sibilants 3–7 from the center of the sibilant continuum than for sibilants 1–2 and 8–10 from both ends of the continuum. This finding shows up statistically in a significant interaction of final sibilant and F0 context. Importantly, the fact that the F0 context effect primarily occurs in the center of the sibilant continuum does not undermine the relevance (external validity) of the present findings. The crucial point from the perspective of the present research question is *that* the F0 context effect occurred, not *where* in the continuum it occurred. Moreover, sibilant productions of /ʃ/ and /s/ that are as clearly articulated as those of sibilants 1–2 and 7–10 of the present continuum hardly ever occur in connected (spontaneous or read) speech. With reference to CoG measurements, typical /ʃ/ and /s/ realizations in German as well as in other languages are characterized by CoG values of about 6–7 and 8–9 kHz, respectively (see [25, 26, 28, 29], and the references provided in section 1.3). As can be seen in Table 1, this is exactly the CoG range of stimuli 3–7 for which the observed effects of F0 context were most strongly pronounced.

In conclusion, the results of the present study are in agreement with hypotheses 2 (a) and (b) and 3. As in previous studies of, for example, Niebuhr [24], Kohler [31], Mixdorff et al. [32, 33], Heeren [27], and Welby and Niebuhr [35], the data clearly show the perceptual relevance of segmental intonation. Going beyond this basic relevance, evidence is provided here that listeners perceive the variation in the sound quality of fricatives indeed as a variation in aperiodic pitch, even if they were not explicitly asked to do so (i.e., unlike in the study of [27]). Moreover, listeners obviously relate this aperiodic pitch impression to the adjacent F0 context. That is, the variation in aperiodic pitch is not merely functionally interpreted, as is assumed by Kohler [31] with reference to basic ethological concepts like frequency code, effort code, and production code (cf. [86–88]). If the listeners' approach to variation in aperiodic pitch was merely functionally oriented, then hypothesis 2 (a) and (b) would have not been confirmed and there would have been no significant F0 context effect at all. However, the opposite is true. The relation that is established by listeners between a voiceless fricative and its F0 context is actually one of perceived pitch. The term "segmental intonation" that was introduced by Niebuhr [25] is perceptually appropriate. Furthermore, taking into account that the results are interpretable in terms of a subtraction of aperiodic pitch from the sibilants, and in view of the solid evidence (e.g. from the micro-perturbation of F0 by consonants, see section 1.3) that the subtracted phonetic substance is reassigned and integrated in the perception of the context factor it stems from, the present results indirectly suggest that the subtracted aperiodic pitch is integrated in the perception of the F0 or intonation contour. Not least for this reason, segmental intonation can well be one part of the explanation why utterance tunes typically appear "*subjectively continuous*" ([34]; 275) in the ears of listeners—despite the fact that a considerable percentage of all utterances in speech communication (between 25 and 35% in Western Germanic languages like English and German, cf. [34]) is actually voiceless. The conclusion of the present study also matches with more recent findings of Mixdorff et al. [32, 33] and Welby and Niebuhr [35], which all suggest that listeners can fill-in interrupted F0 sections, if the interruption does not manifest itself as complete silence (like for unaspirated stop consonants), but is at least partly filled by fricative noise.

4.2 Effects of control variables

The effect of vowel quality is in line with that of previous studies, for example, by Mann and Repp [53] and Clayards et al. [76]. They found that sibilants were less likely to be identified as [ʃ] in [u] than in [a] contexts. A parallel effect emerged when [a] contexts were compared to [i] contexts. That is, [a] contexts triggered fewer [ʃ] identifications than [i] contexts. These effects are ascribed to the principle that "*perception parallels production*" ([53]; 215): Listeners expect coarticulatory reflections of the vowels in the sibilants and "subtract" these reflections from the perceived sibilant noise before they identify the sibilant phoneme. It is reasonable to assume that such an expectation-driven cognitive process also caused the vowel quality effect of the present study.

In fact, the perception-parallels-production principle and its subtractive effect on phoneme perception is the same that is also assumed here to underlie the F0 context effect—the only exception being that the "segmental intonation" effect in fricative production may not simply be coarticulatory in nature, see the explanations in section 1.1 on larynx height.

In the case of the vowel quality effect, the back and high tongue shape of the /u/ in combination with lip rounding bundles almost all acoustic energy below 1000 Hz and gives this sound its inherently "dark" quality. A part of the tongue and lip constellations for /ʊ/ will also continue into a subsequent [ʃ]; and if listeners expect and subtract this coarticulatory "darkening" from the sibilant, their perception is, in comparison to /a/ contexts, shifted towards /s/. The same explanation can be applied to /a/ contexts in comparison to /ɪ/ contexts.

The perception-parallels-production principle could also have caused an effect of the other control variable: original sibilant. But, such an effect did not materialize. The reason is probably that by cutting off the original sibilant phoneme, we also removed its anticipatory coarticulation pattern in the preceding vowel. It is indeed true and supports the provided explanation that the removal of the original sibilant also removed some part of the vowel formant transitions (including that of F2), which were found to be particularly important for identifying the following sibilant's place of articulation (cf. [53, 78, 79]). Alternatively (or additionally), it is possible that formant transition residuals of the original sibilant are irrelevant perceptually as long as there is no phonetically closely related neighboring phoneme, i.e. a "spectrally confusable fricative" in the terminology of Wagner et al. [7]. Wagner et al. found that misleading formant transitions created by fricative cross-splicing in pseudowords did not affect the identification rate of /s/ and /ʃ/ for German listeners, because the two fricatives are too different from each other to be spectrally confusable.

4.3 Further insights from reaction times

In general, reaction times measured in the present perception experiment were all at a normal low level, for example, with reference to similar 2AFC decision tasks and stimuli in Poeppel et al. [89], Gerrits, [65], Schneider et al. [90], or Cangemi and D'Imperio [91]. This fact lends further support to the conclusion drawn from the questionnaire analysis in section 3.1 that the stimuli's sound quality and naturalness were overall good, and that the task was easy to do for listeners as well as short enough to prevent fatigue (see also the lack of a significant effect of stimulus repetition in the linear mixed-effects model, Table 5).

The effect of the final sibilant on reaction times manifested itself as an increase in the center of the sibilant continuum. This increase is most likely caused by the fact that sibilants in the center, i.e. closer to the phoneme boundary, were more ambiguous as to their perceptual identification as either /s/ or /ʃ/ than sibilants from the periphery of the continuum. Based on this assumption, the reaction-time results provide further evidence in support of hypothesis 3.

More ambiguous instances within a phonemic contrast require longer cognitive processing times than clearer instances, for example, because there are fewer stored exemplars against which these sounds can be matched by listeners, or because listeners have to make a more comprehensive analysis of contextual factors before a specific phonemic interpretation becomes possible. Irrespective of its exact reason, the final-sibilant effect replicated an earlier finding by Chen [92], who argued that reaction-time measurements can basically substitute discrimination judgments in experiments looking for categorical perception. Chen stated that discrimination judgments and reaction-time measurements should be equally suitable to complement identification results in a categorical-perception setting as their data curves are both expected to develop a clear peak value in the area where the identification curve crosses the 50% threshold. Shedding light on this reaction-time-related question could be an interesting task of follow-up studies. Further research perspectives are sketched in the following section.

The effect of F0 context on reaction times, i.e. the additional increase in the center of the sibilant continuum, is probably due to a contextual mismatch. That is, it is assumed that sibilants from the center of the continuum that were identified as /ʃ/ created a too high aperiodic pitch impression to match well with the low F0 context. Likewise, sibilants from the center of the continuum that were identified as /s/ created a too low aperiodic pitch impression to match well with the high F0 context. In other words, the F0 context effect on reaction times is explained by inconsistencies in F0-based and aperiodic pitch perception. In this sense, the effect provides further support for the conclusion that listeners are not just perceptually sensitive to the F0-related variation in the sound quality of fricatives (i.e., to "segmental intonation"). Rather, they treat this variation literally as "segmental intonation" and relate it to the F0 context when perceiving an utterance tune.

5 Outlook

The participants of our experiment judged the task to be easy and the stimuli to sound natural and free from manipulation artefacts. The decisive effect of F0 context on sibilant identification is highly significant in terms of both lexical judgments and reaction times. Reaction times are overall relatively short, i.e. participants needed on average only about 700 ms to decide whether the stimulus-final target word ended in /ʃ/ or /s/. All these facts imply that our experimental findings can be considered valid and reliable.

Yet, one task of follow-up studies is to test whether the present findings can be replicated for different phonological intonation patterns and hence different F0 contexts as well as for different languages and other types of fricatives than sibilants. A broader empirical basis would substantially underpin the present conclusion that the term "segmental intonation" is phenomenologically appropriate in its literal sense, i.e. that it denotes a direct contribution of the aperiodic pitch impressions of noise segments in speech (voiceless fricatives as well as postaspiration sections after voiceless plosives) to the perceived pitch contour of utterances.

Creating a broader and more solid empirical basis could also include the use of synthesized fricative continua (or synthesized speech in general) in which the stepwise changes in mean CoG and potentially confounding factors in the spectral energy distribution can still be better controlled than in the present study. The present study has demonstrated, though, that it is generally possible to generate uniformly structured CoG stimulus continua based naturally produced sibilant tokens. However, firstly, such natural productions involved an in-depth knowledge of the relevant articulatory changes that are required for creating the uniform CoG steps; secondly, selecting the right tokens for the CoG continuum involved phonetically trained ears. The in-depth articulatory knowledge and its skilled implementation in fricative realization as well as the subsequent phonetic listening skills in fricative selection might not always be available in all research projects and for all languages. Moreover, based on the author's experience with fricative production, acoustics, and perception gained during the work on segmental intonation, it is likely that generating uniformly structured CoG stimulus continua could be much harder for fricatives like [ɸ], [f], [θ], and [ç] (whose spectrum is fairly flat and whose CoG range is inherently small in natural speech production, cf. [6]) than it is for sibilants and dorsal or glottal fricatives like [x] and [h]. So, using synthesized fricative continua probably becomes unavoidable at a certain stage, and it would be interesting to compare how these synthetic continua perform (perceptionwise) in relation to naturally produced ones.

A further obvious line of follow-up research must determine in more detail how segmental intonation and F0-based intonation are actually integrated in perception. Based on experiments of Mixdorff et al. [32, 33] and Welby and Niebuhr [35], it is still only marginally understood, to what extent, in which way, and under which circumstances the aperiodic pitch impressions of fricatives contribute to the subjective continuity of utterance intonation contours. For example, do aperiodic pitch impressions directly fill-in missing F0 sections or do aperiodic pitch impressions only function as acoustic triggers for a signal-external interpolation and filling-in of F0 gaps? Addressing these questions would not just be interesting from a phonetic point of view; finding answers would also help improving speech recognition and synthesis algorithms and contribute to updating current intonational phonology frameworks in such a way that they go beyond F0 and integrate the two traditionally separated layers of the speech signal: segments and prosodies.

Finally, one reviewer asked whether it is possible in the future to provide more direct evidence for the integration of segmental intonation in the perception of speech melody. We think this is tricky to test without explicitly drawing the participants attention to the target fricatives' aperiodic pitch impressions. However, it is without a doubt a worthwhile task, and one option could be to let participants draw the perceived intonation contours of utterances whose fricatives are cross-spliced between different F0 contexts. An initial feasibility test for the use of such a drawing task in a follow-up experiment has already been successfully conducted.

Acknowledgements

I would first and foremost like to thank Jonathan Harrington whose ideas on the investigation of segmental intonation paved the way for the present experiment. Moreover, I am greatly indebted to my internship student, Stefanie Berger, for her help with producing and preparing the experimental stimuli. Further thanks are due to Simon Betz, Katharina Wendler, and Plinio Barbosa for their kind and patient help with the statistical analyses of the data. I am also very grateful to Marszena Zygis and two further anonymous reviewers for their constructive, insightful comments, and concrete suggestions on the earlier drafts of this paper. Last but not the least, I would like to express my gratitude to Timo Röttger, Simon B. Ritter, and all colleagues from the Phonetics Institute in Cologne for the many fruitful and inspiring discussions on the issue of segmental intonation.

Competing interests

The author declares that he has no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 24 August 2016 Accepted: 24 July 2017
Published online: 02 August 2017

References

1. Grice, M., Baumann, S., Benzmüller, R. (2005). German intonation in autosegmental-metrical phonology. In SA. Jun (Eds.), *Prosodic typology: The phonology of intonation and phrasing* (pp. 55–83). Oxford: Oxford University Press.
2. Kohler, K. J. (1991). Prosody in speech synthesis: the interplay between basic research and TTS application. *Journal of Phonetics, 19*, 121–138.
3. Kohler, K. J. (1997). Modelling prosody in spontaneous speech. In Y. Sagisaka, N. Campbell, H. Higuchi (Eds.), *computing prosody. Computational models for processing spontaneous speech* (pp. 187–210). New York: Springer.
4. Mixdorff, H. (2012). The application of the Fujisaki model in quantitative prosody research. In O. Niebuhr (Eds.), *Understanding Prosody: Context, Function, Communication* (pp. 55–74). Berlin/New York: de Gruyter.
5. Peters, J. (2015). *Intonation - Kurze Einführungen in die germanistische Linguistik 16*. Heidelberg: Winter.
6. Johnson, K. (2012). *Acoustic and auditory phonetics* (3rd ed.). Oxford: Wiley-Blackwell.
7. Wagner, A., Ernestus, M., & Cutler, A. (2006). Formant transitions in fricative identification: the role of native fricative inventory. *Journal of Acoustical Society of America, 120*, 2267–2277.
8. Terhardt, E. (1974). Pitch, consonance, and harmony. *Journal of the Acoustical Society of America, 55*, 1061–1069.
9. Traunmüller, H. (1987). Some aspects of the sound of speech sounds. In M. E. H. Schouten (Eds.), *The psychophysics of speech perception* (pp. 293–305). Dordrecht: Martinus Nijhoff.
10. Thomas, I. B. (1969). Perceived pitch in whispered vowels. *Journal of the Acoustical Society of America, 46*, 468–470.
11. Higashikawa, M. (1994). Perceptual, acoustical and aerodynamic study of whispering. *Nippon Jib Gak Kaiho, 97*, 1268–1280.
12. Higashikawa, M., Minifie, F. D. (1999). Acoustical-perceptual correlates of 'whisper pitch' in synthetically generated vowels. *Journal of Speech Language & Hearing Research, 42*, 583–591.
13. Abramson, A. S. (1972). Tonal experiments with whispered Thai. In A. Valdman, *Papers in linguistics and phonetics to the memory of Pierre Delattre* (pp. 31–44). The Hague: Mouton.
14. Krull, D. (2001). Perception of Estonian Word Prosody in Whispered Speech. In Proc. 8th Nordic Prosody Conference (pp. 153–164). Norway: Trondheim.
15. Meyer-Eppler, W. (1957). Realization of prosodic features in whispered speech. *Journal of the Acoustical Society of America, 29*, 104–106.
16. Whalen, D. H., & Xu, Y. (1992). Information for Mandarin tones in the amplitude contour and in brief segments. *Phonetica, 49*, 25–47.
17. Konno, H., Kanemitsu, H., Toyama, J., & Shimbo, M. (2006). Spectral properties of Japanese whispered vowels referred to pitch. *Journal of the Acoustical Society of America, 120*, 3378.
18. Liu, S., & Samuel, A. G. (2004). Perception of Mandarin lexical tones when f0 information is neutralized. *Language and Speech, 47*, 109–138.
19. Nicholson, H., & Teig, A. H. (2003). How to tell beans from farmers: cues to the perception of pitch accent in whispered Norwegian. *Nordlyd, 31*, 315–325.
20. Jensen, M. K. (1958). Recognition of word tones in whispered speech. *Word, 14*, 187–196.
21. Miller, J. D. (1961). Word tone recognition in Vietnamese whispered speech. *Word, 17*, 11–15.
22. Wang, D. L., Hu, G. (2006). Unvoiced speech segregation. In *Proc. IEEE ICASSP Tolouse* (pp. 953–956). France.
23. Balise, R., & Diehl, R. L. (1994). Some distributional facts about fricatives and a perceptual explanation. *Phonetica, 51*, 99–110.
24. Niebuhr, O. (2008). Coding of intonational meanings beyond F0: evidence from utterance-final /t/ aspiration in German. *Journal of the Acoustic Society of America, 142*, 1252–1263.
25. Niebuhr, O. (2009). Intonation segments and segmental intonations. In *Proc. 10th Interspeech Conference* (pp. 2435–2438). Brighton.
26. Niebuhr, O. (2012). At the edge of intonation—the interplay of utterance-final F0 movements and voiceless fricative sounds. *Phonetica, 69*, 7–27.
27. Heeren, W. F. L. (2015). Coding pitch differences in voiceless fricatives: whispered relative to normal speech. *Journal of the Acoustical Society of America, 138*, 3427–3438.
28. Niebuhr, O., Lill, C., Neuschulz, J. (2011). At the segment-prosody divide: the interplay of intonation, sibilant pitch and sibilant assimilation. In *Proc. 17th ICPhS* (pp. 1478–1481). Hong Kong, China.
29. Ritter, S., Röttger, T. B. (2014). Speakers modulate noise-induced pitch according to intonational context. In *Proc. 7th International Conference of Speech Prosody* (pp. 1-5). Dublin.
30. Żygis, M., Pape, D., Jesus, L. M. T., Jaskuła, M. (2014). Intonation of statements and polar questions in whispered, semi-whispered and normal speech modes. In *Proc. 7th International Conference of Speech Prosody* (pp. 1-5). Dublin.
31. Kohler, K. J. (2011). Communicative functions integrate segments in prosodies and prosodies in segments. *Phonetica, 68*, 26–56.
32. Mixdorff, H., Hönemann, A., Niebuhr, O., Draxler, C. H. (2014). Perceived prominence reflected by imitations of words with and without F0 continuity. In *Proc. 7th International Conference of Speech Prosody* (pp. 1-5). Dublin.
33. Mixdorff, H., Niebuhr, O. (2013). The influence of F0 contour continuity on prominence perception. In *Proc. 14th Interspeech Conference* (pp. 230–234). Lyon.
34. Jones, D. (1909). *Intonation curves*. Leipzig: Teubner.
35. Welby, P., Niebuhr, O. (2016). The influence of F0 discontinuity on intonational cues to word segmentation: a preliminary investigation. In *Proc. 8th International Conference of Speech Prosody* (pp. 40–44). Boston.
36. Gow, D. W. (2003). Feature parsing: feature cue mapping in spoken word recognition. *Perception Psychophys, 65*, 575–590.
37. Ohala, J. J., & Feder, D. (1994). Listeners' identification of speech sounds is influenced by adjacent 'restored' phonemes. *Phonetica, 51*, 111–118.
38. Fowler, C., Brown, J., Sabadini, L., & Weihing, J. (2003). Rapid access to speech gestures in perception: Evidence from choice and simple response time tasks. *Journal of Memory and Language, 49*, 396–413.
39. Harrington, J., Kleber, F., & Reubold, U. (2008). Compensation for coarticulation, /u/–fronting, and sound change in standard southern British: an acoustic and perceptual study. *Journal of the Acoustical Society of America, 123*, 2825–2835.
40. Kleber, F., Harrington, J., & Reubold, U. (2012). The relationship between the perception and production of coarticulation during a sound change in progress. *Language and Speech, 55*, 383–405.
41. Goldstein, E. B. (2014). *Sensation and perception* (9th ed.). Wadsworth: Cengage.
42. Landgraf, R. (2015). Simulating complex speech-production environments. In O. Niebuhr, R. Skarnitzl (Eds.), *Tackling the complexity of speech* (pp. 97–110). Prague: Epocha.
43. Scherer, K. R. (2003). Vocal communication of emotion: a review of research paradigms. *Speech Communication, 40*, 227–256.
44. Behrens, S., & Blumstein, S. E. (1988). On the role of the amplitude of the fricative noise in the perception of place of articulation in voiceless fricative consonants. *Journal of the Acoustical Society of America, 84*, 861–867.
45. Goodacre, J., & Nakajima, Y. (2005). The perception of fricative peaks and noise bands. *Journal of Physiological Anthropology, 24*, 151–154.
46. Gordon, M., Barthmaimer, P., & Sands, K. (2002). A cross-linguistic study of voiceless fricatives. *Journal of the International Phonetic Association, 32*, 141–174.
47. Heinz, J. M., & Stevens, K. N. (1961). On the properties of voiceless fricative consonants. *Journal of the Acoustical Society of America, 33*, 589–596.
48. Maniwa, K., Jongman, A., & Wade, T. (2009). Acoustic characteristics of clearly spoken English fricatives. *Journal of the Acoustical Society of America, 125*, 3962–3973.
49. Soli, S. (1981). Second formants in fricatives: acoustic consequences of fricative-vowel coarticulation. *Journal of the Acoustical Society of America, 70*, 976–984.
50. Tabain, M. (2001). Variability in fricative production and spectra: Implications for the hyper-and hypo-and quantal theories of speech production. *Language and Speech, 44*, 57–94.
51. Toda, M. (2007). Speaker normalization of fricative noise: considerations on language-specific contrast. In *Proceedings of the 16th ICPhS* (pp. 825–828). Saarbrücken.
52. Russ, C. V. J. (2010). *The sounds of German*. Cambridge: Cambridge University Press.
53. Mann, V. A., & Repp, B. H. (1980). Influence of vocalic context on perception of the [ʃ]-[s] distinction. *Perception and Psychophysics, 28*, 213–228.

54. Munson, B., Jefferson, S. V., & McDonald, E. (2006). The influence of perceived sexual orientation on fricative identification. *Journal of the Acoustical Society of America, 119*, 2427.
55. Winn, M. B., Rhone, A. E., Chatterjee, M., & Idsardi, W. J. (2013). The use of auditory and visual context in speech perception by listeners with normal hearing and listeners with cochlear implants. *Frontiers in Psychology, 4*, 824.
56. Boersma, P., & Hamann, S. (2008). The evolution of auditory dispersion in bidirectional constraint grammars. *Phonology, 25*, 217–270.
57. Lehiste, I. (1970). *Suprasegmentals*. Cambridge: Cambridge University Press.
58. Klatt, D. H., & Cooper, W. E. (1975). Perception of segment duration in sentence contexts. *Communication and Cybernetics, 11*, 69–89.
59. Huggins, A. (1968). How accurately must a speaker time his articulations? *IEEE Transactions on Audio and Electroacoustics, 16*, 112–117.
60. Rossing, T. D., & Houtsma, A. J. M. (1986). Effects of signal envelope on the pitch of short sinusoidal tones. *Journal of the Acoustical Society of America, 79*, 1926–1933.
61. Gussenhoven, C. (1999). Discreteness and gradience in intonational contrasts. *Language and Speech, 42*, 283–305.
62. Kohler, K. J. (2004a). Categorical speech perception revisited. In *Proc. From Sound to Sense: 50+ years of discoveries in speech communication* (pp. 157–162). Cambridge.
63. Gerrits, E., & Schouten, M. E. H. (2004). Categorical perception depends on the discrimination task. *Perception & Psychophysics, 66*, 363–376.
64. Pisoni, D. B. (1975). Auditory short-term memory and vowel perception. *Memory & Cognition, 3*, 7–18.
65. Gerrits, E. (2001). *The categorisation of speech sounds by adults and children. PhD thesis*. Utrecht: Utrecht University.
66. Beck, J. (2014). Experiment-MFC: Erstellung und Auswertung eines Perzeptions experiments in Praat. *KALIPHO, 2*, 81–113.
67. Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International, 5*, 341–345.
68. Harris, K. S. (1958). Cues for the discrimination of American English fricatives in spoken syllables. *Language and Speech, 1*, 1–7.
69. Kleber, F., John, T., & Harrington, J. (2010). The implications for speech perception of incomplete neutralization of final devoicing in German. *Journal of Phonetics, 38*, 185–196.
70. Noteboom, S. (1981). Speech rate and segmental perception or the role of words in phoneme identification. *Advances in Psychology, 7*, 143–150.
71. Walley, A. C., & Flege, J. E. (1999). Effect of lexical status on childrens' and adults' perception of native and non-native vowels. *Journal of Phonetics, 27*, 307–332.
72. Rathcke, T. (2013). On the neutralizing status of truncation in intonation: a perception study of boundary tones in German and Russian. *Journal of Phonetics, 41*, 172–185.
73. Jones, R., & Tschirner, E. (2005). *Frequency dictionary of German*. London: Routledge.
74. Fox, R. A. (1984). Effect of lexical status on phonetics categorization. *Journal of Experimental Psychology: Human Perception and Performance, 10*, 526–540.
75. Kohler, K. J. (2004b). Pragmatic and attitudinal meanings of pitch patterns in German syntactically marked questions. In G. Fant, H. Fujisaki, J. Cao, Y. Xu (Eds.), from traditional phonology to modern speech processing. *Festschrift for professor Wu Zongji's 95th birthday* (pp. 205–214). Beijing: Foreign Language Teaching and Research Press.
76. Clayards, M., Gaskell, G., Niebuhr, O. (2011). Comparing French and English listeners' on-line perception of assimilated speech. In *Proc. 10th International Symposium of Psycholinguistics*. Spain: Donastia San-Sebastian. http://www.bcbl.eu/events/psycholinguistics/speakers/desde0/ver/330/. Accessed 29 July 2017
77. Datscheweit, W. (1990). Frication noise and formant-onset frequency as independent cues for the perception of /f/, /s/ and /ʃ/ in vowel-fricative-vowel stimuli. In *Proc. International Conference on Spoken Language Processing* (pp. 561–564). Kobe.
78. Ooijevaar, E. S. M. (2011). *Cue weighting in the perception of Dutch sibilants, MA thesis*. RMA Linguistics, University of Amsterdam.
79. Whalen, D. H. (1991). Perception of the English /s/–/ʃ/ distinction relies on fricative noises and transitions, not on brief spectral slices. *Journal of the Acoustical Society of America, 90*, 1776–1785.
80. Tabossi, P. (1996). Cross-modal semantic priming. *Language and Cognitive Processes, 11*, 569–576.
81. Jaeger, T. F. (2008). Categorical data analysis: away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language, 59*, 434–446.
82. Baayen, H. R. (2008). *Analyzing linguistic data. A practical introduction to statistics using R*. Cambridge: Cambridge University Press.
83. Baayen, H., & Milin, P. (2010). Analyzing reaction times. *International Journal of Psychological Research, 3*, 12–28.
84. Gravetter, F., & Wallnau, L. (2014). *Essentials of statistics for the behavioral sciences* (8th ed.). Wadsworth: Belmont.
85. Beddor, P. (2009). A coarticulatory path to sound change. *Language, 85*, 785–821.
86. Gussenhoven, C. (2002). Intonation and interpretation: phonetics and phonology. In *Proc. 1st International Conference of Speech Prosody* (pp. 47–57). France: Aix-en-Provence.
87. Ohala, J. J. (1984). An ethological perspective on common cross-language utilization of F0 of voice. *Phonetica, 41*, 1–16.
88. Ohala, J. J. (1994). The frequency codes underlies the sound symbolic use of voice pitch. In L. Hinton, J. Nichols, J. J. Ohala, *Sound symbolism* (pp. 325–347). Cambridge: Cambridge University Press.
89. Poeppel, D., Guillemin, A., Thompson, J., Fritz, J., Bavelier, D., & Braun, A. R. (2004). Auditory lexical decision, categorical perception, and FM direction discrimination differentially engage left and right auditory cortex. *Neuropsychologia, 42*, 183–200.
90. Schneider, K., Dogil, G., Möbius, B. (2011). Reaction Time and Decision Difficulty in the Perception of Intonation. In *Proc. 10th INTERSPEECH Conference* (pp. 2221–2224). Florence.
91. Cangemi, F., & D'Imperio, M. (2013). Tempo and the perception of sentence modality. *Laboratory Phonology, 4*, 191–219.
92. Chen, A. (2003). Reaction time as an indicator to discrete intonational contrasts in English. In *Proc. EUROSPEECH 2003* (pp. 97–100). Geneva.
93. Coleman, J., Grabe, E., Braun, B. (2002). Larynx movements and intonation in whispered speech. Summary of research supported by British Academy Grant No. SG-36269. http://www.phon.ox.ac.uk/files/pdfs/project_larynx_summary.pdf. Accessed 20 May 2017.
94. Kohler, K. J. (1990). Macro and micro F0 in the synthesis of intonation. In J. Kingston, M. E. Beckamn (Eds.), *Papers in Laboratory Phonology I* (pp. 115–138). Cambridge: Cambridge University Press.