Do children better understand adults or themselves?  
An acoustic and perceptual study of the complex sibilant system of Polish

*Marzena Żygis  
Leibniz-Centre General Linguistics & Humboldt University, Berlin, Germany  
zygis@leibniz-zas.de  
Corresponding author

Daniel Pape  
McMaster University, Hamilton, Canada  
paped@mcmaster.ca

Marek Jaskuła  
Westpomeranian University of Technology, Szczecin, Poland  
Marek.Jaskula@zut.edu.pl

Laura L. Koenig  
Adelphi University Garden City, New York, USA, Haskins Laboratories, New Haven CT, USA, & Leibniz-Centre General Linguistics, Berlin, Germany  
koenig@haskins.yale.edu

Abstract

This paper reports a developmental production-perception study of the three-way Polish sibilant contrast /s, ʂ, ɕ/ in typical children (N=76). Children ages 2;11–7;11 produced words with sibilants in word-medial and initial position. They then identified the same words they produced, and the words as produced by an unknown adult female. Results show higher identification accuracy for the adult productions across all ages. Production and perception data suggest that the alveolo-palatal /ɕ/ is acquired first, and that it is differentiated mainly by formant patterns. In the perceptual test, most errors were found for child-produced /ʂ/, and this persisted into the oldest age group. Early acquisition of /ɕ/ has been observed in other languages and may reflect motoric considerations as well as a focus on formant-related information in child speech perception. Cue weighting appears to change over age in sibilant-specific ways. This work contributes to the study of cross-language differences in acquisition, provides an acoustic characterization of child-produced Polish sibilants, and helps elucidate the acoustic characteristics that children rely on in making perceptual judgments of fricatives.
1. Introduction

This paper presents a developmental perception–production study focusing on the Polish sibilant contrast among dento-alveolar /s/, retroflex /ʂ/, and palatal /ɕ/. Three-way sibilant contrasts are quite rare, and descriptions of them are even more sparse. Sibilants, as a class, are also acquired rather late, so that it is informative to follow production and production changes over a multi-year period. The aims of the work were three-fold: a) to assess children’s accuracy in perceiving their own speech compared to that of an adult model; b) to explore the acoustic variables that characterize the three fricatives, and which could lend insight into the perceptual findings; and c) to infer the order of acquisition among the three Polish fricatives and compare it to other languages with complex sibilant systems. Research relevant to these questions is laid out in the next three sections.

1.1 Children’s perception of adult vs. child speakers

Extensive work on the perception of speech by typically-developing infants and children indicates that the acquisitional process initially draws on biologically-given mechanisms and is subsequently shaped by the linguistic environment (for a review, see Werker & Hensch, 2015). Aspects of speech perception continue to be refined into the school-age years (e.g., Hazan & Barrett, 2000; Nitttrouer, 2002), i.e. well after the shift away from language-general processing that begins late in the first year of life (Werker & Hensch, 2015). A noteworthy feature of the developmental literature is that children have almost always been tested on their perception of adult-based forms. Listening materials have included a) speech produced naturally by adult speakers (e.g., Fernald, Perfors, & Marchman, 2006; Polka, Colantino, & Sundara 2001; Swingley & Aslin, 2000); b) synthetic tokens based on adult models (e.g., Hazan & Barrett, 2000; Mayo & Turk, 2004); and c) hybrid stimuli with naturally-produced segments spliced onto synthetic continua (e.g., Nitttrouer, 2002). As a result, rather little is known about child perception of child-produced speech, whether the speaker is the individual child (self-perception) or other children. However, this topic has gained interest in recent years (see Bernier and White, 2019a, 2019b; Cooper, Fecher, and Johnson, 2018; all reviewed in more detail below). One impetus for this work has been to clarify the nature of children's underlying representations as they develop their lexicons and phonological systems, and how those representations relate to the input that children receive, both from themselves and from others around them. Research has also called attention to possible differences between self- and other-perception as a factor underlying some developmental speech-sound disorders (Shuster, 1988; Strömbergsson, Wengelin, & House, 2014). Thus, the issue has both theoretical and practical implications.

Multiple theoretical perspectives (cf. Bernier & White, 2019a, b; Cooper et al., 2018) suggest that children’s perceptual accuracy (e.g., correct labeling) might differ for adult- vs. child-produced speech, but the predictions are conflicting. Under the assumption that young children usually receive extensive exposure to their caregivers’ voices, it could follow that early lexical and phonological representations are mainly based on adult speech, in accordance with episodic (e.g., Goldinger, 1996) or exemplar (e.g., Pierrehumbert, 2003) models. These models predict higher accuracy for an adult voice because it provides a better match to the stored representation than a child’s voice. One could likewise suppose, however, that children are quite familiar with their own voices as well, and therefore show an advantage for self-produced speech. This is somewhat tempered by the fact that self-perception is affected by bone conduction (Maurer & Landis, 1990). A stronger argument for a self-produced speech advantage comes from common coding perspectives (e.g., Prinz, 1992; cf. discussion in Tye-Murray et al., 2013). In these models, action
and perception have a shared internal representation, and actions should be perceived more accurately when they match or approximate one’s own performance. A self-perception advantage has been documented for a variety of adult actions, including music performance, dancing, clapping, and throwing darts (Knoblich & Sebanz, 2006). Tye-Murray, Spehar, Myerson, Hale, and Sommers (2013) also observed that adults were usually more accurate when lip-reading themselves as compared to unknown individuals.¹ A final consideration is that, by the time children reach the school-age years, they will typically have gained experience with a large number of child voices, which could have the effect of expanding internal representations so that child and adult voices are perceived with approximately equivalent facility.

Two inter-related developmental findings are also relevant in the context of self-perception of speech in children: The ‘fis phenomenon’ (Berko & Brown, 1960) and covert contrasts (e.g., Scobbie, Gibbon, Hardcastle, & Fletcher, 2000). In the first, children object to an adult imitation of a child’s perceived substitution (here, \( [ʃ] \rightarrow [s] \)): “Not a [fɪs], a [fɪs].” In the second, children produce a phonetic difference that is, for adults, subphonemic. Many examples of covert contrasts have been recognized over the years, including distinctions of voicing (Macken & Barton, 1980), plosive place of articulation (viz., alveolar~velar, e.g. McAllister Byun et al., 2016), and, notably for present purposes, sibilant fricatives (Li et al., 2009). One possible interpretation of both of these phenomena is that some children produce a distinction that they, in fact, perceive. Such contrasts may involve atypical acoustic differences (Li et al., 2009), secondary cues that are redundant in adult speech (Forrest & Rockman, 1988), or differences along the expected dimension (e.g., voice onset time) that remain within an adult phonemic category (Macken & Barton, 1980). Covert contrast implies that a child has some form of phonemic difference that is not yet manifested phonetically in an adult-like fashion. This could explain why producing a covert contrast is a positive prognostic indicator for children undergoing therapy for speech sound disorders (Tyler, Figurski, & Langsdale, 1993): Their internal representations are in some sense more highly-specified than those of children who produce no contrast whatsoever.

Only a few empirical studies have evaluated child perception of child speech in typical development. Dodd (1975) carried out an early experiment on this question. Fourteen children ages 2;3–4;7 (years;months) produced words in a naming task. For each child, words with consistent errors (e.g., consonant omission or substitution) were selected. In subsequent sessions, children (the original 14 plus another cohort of 14, ages 3;0–4;9) performed a picture-identification task in response to the words produced by a) an unknown adult (using no errors) and b) the child him/herself, or an unknown child. In all cases, children’s responses were more accurate for the adult speaker. Dodd concluded that children constructed linguistic representations based on adult forms. More recently, Hazan and Markham (2004) recorded words produced by women, men, and 13-year-olds (45 speakers in total) and played them to adults, 11–12-year-olds and 7–8-year-olds (45 listeners in total). All listener groups showed higher accuracy levels (i.e., higher intelligibility scores) for adult speakers than for adolescents. Similar results were obtained by Cooper, Fecher, Schuerman et al. (2015) assessed the perception of participants’ own speech compared to an ‘average’ speaker, using a vocoder to remove source information. The authors reported an advantage for the average speaker over self-perception for easy but not hard words. Given the use of vocoded speech and construction of an average, it is difficult to compare these results to those of the child-based studies discussed in the text.

¹ We are aware of only one study of self- vs. other-perception of oral speech in adults. Schuerman et al. (2015) assessed the perception of participants’ own speech compared to an ‘average’ speaker, using a vocoder to remove source information. The authors reported an advantage for the average speaker over self-perception for easy but not hard words. Given the use of vocoded speech and construction of an average, it is difficult to compare these results to those of the child-based studies discussed in the text.
and Johnson (2018). These authors obtained word productions from 54 mother-child pairs, with child ages 2;6–3;0. In a subsequent recording session, children were recorded in an eye-tracking protocol as they listened to words produced by themselves and their mothers along with words produced by an unfamiliar child and mother. The results showed an effect of age, but no effect of familiarity and no age × familiarity interaction: Children shifted their gaze to the target picture more quickly, and looked at the picture longer, when the speaker was an adult (known or unknown), suggesting faster and more accurate recognition of adult forms.

In contrast, two recent studies by Bernier and White (2019a, 2019b), using eye-gaze measures to assess two-year-olds’ perception of an unknown adult and child speaker (approximately 7 years of age), found no main effects of speaker age. In Bernier and White (2019b), speaker age also did not interact with the other variables of novel vs. familiar labels and level of toddler experience with child speech (quantified according to time spent in daycare vs. at home). Bernier and White (2019a), however, obtained one three-way interaction involving the age factor: For the child speaker, looking times to correct and incorrect productions were comparable regardless of whether errors were typical or atypical in child speech. For the adult, however, there was a large difference in looking time when words contained error patterns common in child speech, but no difference when the errors were atypical. The authors concluded that young children may process child and adult speech somewhat differently via a “complex interplay of speaker, type of mispronunciation, and specific contrast” (pg. 4137).

On the whole, therefore, the extant data do not indicate that children perform better when listening to child speech. Of the studies reviewed here, only two (Cooper et al., 2018; Dodd, 1975) explored self-perception. In Dodd’s study, all child productions had phonemes in error. Cooper et al. obtained productions of 32 words varying widely in phonemic content and complexity (evidently without overt errors; cf. footnote 1). These paradigms do not permit a detailed exploration of children’s perception-production relationships. One goal of the current work was to compare children’s labeling with acoustic analyses of their speech, to gain insight into what they could be attending to in the speech signal using an appropriately controlled dataset.

1.2 Acoustic descriptions of the Polish sibilants

The most relevant acoustic parameters for Polish sibilants can be divided into the influence of spectral characteristics of the sibilant on the one hand, and the coarticulatory effects of that sibilant on neighboring vocalic segments on the other hand. With respect to the spectral shape of the sibilant noise, two distinct measures are normally used. One measure, the frequency of the (major) first highest spectral peak, corresponds directly to the size of the front cavity resonance of the sibilants, and thus increases for more fronted articulations and decreases with more posterior places of articulation. The second measure, the centre of gravity, returns a single-value estimate of the spectral distribution of amplitude values over all frequencies (or sometimes a subrange of the full frequency range). Although the first (major) spectral peak and the centre of gravity often yield similar values, the centre of gravity has the the disadvantage that stronger noise source activation and additional (higher) spectral peaks can shift the centre of gravity value towards higher frequencies, thus obscuring the influence of the first (major) spectral peak.

---

2 Cooper et al. (2018, page 17) state that "Participants were encouraged to produce the word in citation form and were prompted to repeat the item as necessary." Based on this description, it appears that their stimuli did not contain overt errors.
In Polish, adult data show that the dental /s/ is clearly separated from the retroflex /ʂ/ and the alveolo-palatal /ɕ/, with a very high first spectral peak and centre of gravity (around 8000 Hz and higher) compared to a much lower first spectral peak and centre of gravity for both the retroflex and alveolo-palatal (Lisker, 2001; Nowak, 2006; Żygis et al., 2012). In fact, the retroflex and alveolo-palatal often show very similar first (major) spectral peak and centre of gravity values (Nowak, 2006; Żygis et al. 2012) around 3000 Hz, thus making it impossible to distinguish the two sibilants based on frication noise spectral shapes alone. Furthermore, it has been shown that retroflexes and sometimes alveolo-palatals show a second strong (major) peak around 7 kHz (see Żygis & Pape 2016 for a complete discussion, but also Lisker 2001) whose articulatory cause is currently not satisfactorily explained. This additional strong second peak will significantly influence the centre of gravity measurements but not the first spectral peak measurements. The retroflex and alveolo-palatal can, however, be distinguished by means of formant values, most notably the second formant (F2), at the neighboring vowel boundary. Alveolo-palatals show higher second formants due to their palatalized articulation.

With respect to child productions, Łobacz and Dobrzańska (1999) investigated the acoustics of Polish sibilants in 35 items containing [s, z, tʃ, ɖʐ, ʂ, z, ʂ, ʐ, tʃ, ɖʐ, ɕ, ʑ, tʃ, ɖʐ] produced by 19 children aged 4–7 years old. Examining the four highest fricative spectral peaks above 2000 Hz, they found a clear difference between the affricates and fricatives, with the latter being more similar to the adult spectra. They also showed that spectra for [tʃ, ɖʐ], and [tɕ, ɖʐ] were very similar, similar to the adult findings reported above.

In general, the literature on how children differentiate sibilants in production has largely concentrated on a small number of acoustic measures, often just the COG (M1, centroid). Here, we expand on the work of Li (2008) and Łobacz and Dobrzańska (1999) to obtain measures of the first four moments, spectral peak frequencies, and the first three formants of the preceding and following vowels to assess how fricative production develops with age in Polish.3 Thus, a final goal was to contribute to the acoustic descriptions of sibilants in child speech and to use these in seeking explanations for children’s perceptual behavior.

1.3 The development of three-way sibilant contrasts

The few cross-linguistic comparisons of sibilant development suggest this sound class tends to be acquired later than many other sounds, but details vary across languages (Hua & Dodd, 2000; Li, 2008; So & Dodd 1995). Much of this work has been based on adult perceptual judgments of children’s productions; acoustic descriptions of sibilants are limited to English /s ʃ/ (e.g., Cristià, 2010; Holliday, Reidy, Beckman, & Edwards, 2015; Maas & Mailend, 2017; Nittrouer, Studdert-Kennedy, & McGowan, 1989; Zharkova, Hardcastle, & Gibbon, 2018); Japanese /s f/ (Li, 2012; Li, Edwards, & Beckman, 2009); and the three-way /ś ɕ/ contrast in mainland Mandarin Chinese, or Putonghua [standard mainland Mandarin Chinese] (Li, 2008; Li & Munson, 2016).

---

3 Some previous work has also used spectral slope measures to characterize fricative acoustics (e.g., Żygis, Pape, & Jesus, 2012). We did not do that here because a) spectral peaks, used as one reference point in the slope measure, vary widely across the three places of articulation here, and b) our reference data were based on adult speakers, not children. Moreover, the spectral peak measures would be expected to vary with age so that the computation of the slope would not be consistent across ages.
Comparing across diverse languages, Li (2008) carried out acoustic analyses of 2–3-year-olds' sibilant productions in English (two-way contrast), Japanese (two-way contrast) and Mandarin (three-way contrast, similar to Polish). The data showed that the second formant (F2) and the first and second spectral moments (M1, M2; cf. Forrest, Weismer, Milenkovic, & Dougall, 1988) showed language-specific developmental patterns that corresponded to the nature of the adult contrast: English learners mainly differentiated /s ʃ/ by M1, also known as the centroid, or centre of gravity [COG]; Japanese learners showed /s ʃʲ/ differences in both the centroid and standard deviation (M2); and Chinese learners had a more complex pattern whereby all three acoustic measures (M1, M2, F2) were used in the emerging distinction. These results could indicate that the acquisitional process differs for complex, three-way sibilant contrasts as compared to simpler two-way contrasts. Alternatively, these differences may simply be language-specific patterns unrelated to the complexity of the sibilant contrast. More cross-language data are needed to distinguish between these possibilities. Another aim of this study is to contribute instrumental data on Polish, which provides another case of such rare three-way sibilant contrasts.

Previous studies on Polish sibilant acquisition by children are mainly based on perceptual impressions of adult listeners (the exception being Łobacz and Dobrzanska, 1999, mentioned in section 1.2). For instance, Sołtys-Chmielowicz (1998) analyzed the pronunciation of 1063 preschool children, ages 3–8 years, who were asked to name pictures. The words recorded included different types of segments that were annotated by the author according to her percept. With respect to alveolo-palatals /c, z, ċ, ć, dź/ , Sołtys-Chmielowicz reported that 85.9% children aged 3;1–3;6 pronounced the sounds correctly. By comparison, 71.9% children from the same age group were found to correctly pronounce /s, z, ć, dź/. Finally, the retroflexes (frontal and alveolar sounds in the author’s terminology) were pronounced correctly by 25% of children in the same age group. In another study, Bryndal (2015) investigated the acquisition of Polish sounds by 200 pre-school children (aged 3;0–7;2) also by employing a picture naming task. The recordings of sibilants were transcribed by the author and an additional transcriber. The study reports results for each sibilant separately: rates of correct production for the youngest group of children aged 3;0–3;11 were 91% for /c/, 84% for /z/, 78% for /ć/, and 75% for /dź/. As for dento-alveolars, /s/ was classified as 65% correct, /z/ as 59%, /ć/ as 66%, and /dź/ as 56%. The most difficult sibilants for children were retroflexes: the youngest children pronounced /g/ at 37% correct, /z/ at 31%, /ć/ at 34%, and /dź/ at 34%.

Thus, both large-scale developmental studies of Polish show a rather clear pattern: the first sounds acquired by children are alveolo-palatals, followed by dento-alveolars. The latest sounds acquired by children are retroflexes. A similar acquisitional sequence has been observed for Putonghua: Hua and Dodd (2000), using a percentage correct criterion, reported the acquisition order as /c/ > /s/ > /g/, at ages 1;6–2;0, 2;1–2;6, and 3;7–4;0, respectively. A final goal of this work was to evaluate this previously-established acquisitional order in light of fuller acoustic findings, and children’s perception of their own speech.

4 To put these data in the context of the Polish consonant system /p, p’, b, b’, f, f’, v, v’, m, m’, t, d, s, z, c, z, s, ść, ć, dz, ć, dz, ć, dź, ć, dz, n, p, w, l, r, j, k, g, x/, we observe that both the trill /r/ and the retroflex sibilants belong to the latest sounds acquired by children. The rate of correct /r/ production was only 34% for children aged 3;0–3;11, i.e. comparable to the retroflex sibilants. Other fricatives are acquired considerably earlier: The youngest group pronounced /p’/ at 97% correct, /f’/ at 94%, /x’/ at 91%, /v’/ at 91%, and /v’/ at 81% (Bryndal 2015:101). Another later-acquired consonant for the same age group is [v’] (81%).
1.4 Summary and questions

Children's production and perception of sibilants have not been widely explored across languages, especially when it comes to complex contrasts. As a result, it is unclear to what extent acquisitional processes are language-specific or show commonalities across languages with similar contrasts. Instrumental studies have not usually employed multiple acoustic measures nor sampled across ages. Finally, very few studies have contrasted children’s perception of their own speech as compared to an adult model, and those studies were not designed to explore how child perceptual judgments relate to acoustic aspects of the stimuli. A better understanding of self- vs. other-perception is particularly interesting when it comes to sounds, like sibilants, that are late to be produced in an adult-like fashion for many children (Hua and Dodd, 2000; Smit et al., 1990). To address these gaps, the present study evaluates production and perception of the three-way phonemic contrast among /s, ʂ, ɕ/ by Polish children. Acoustic data are evaluated to ascertain changes in age regarding perception of the three sibilants, and perception of one’s own vs. an adult voice. The perceptual results are interpreted in light of the acoustic data. Finally, we bring the acoustic and perceptual data to bear on the order of sibilant acquisition in Polish, which in past studies has relied on adult perceptual judgments.

In summary, our questions were as follows:

a) Do children show higher accuracy when listening to their own (self-produced) speech compared to the voice of an adult?

b) What are the acoustic characteristics of the three Polish fricatives in children’s speech, and how do these relate to the perceptual data?

c) How does the order of acquisition in Polish compare to that of Putonghua, the one other three-way sibilant contrast that has been described in the developmental literature?

The article is organized as follows. In section 2 we will present the methodology of the experiment and in section 3 its results. We will start with the acoustic data (3.1) and continue with the perceptual results (accuracy and RTs) in 3.2. In section 3.3 we will link the perceptual results with acoustic properties of the stimuli. Discussion of the results is presented in section 4. Section 5 concludes.

2. Methods

2.1. Informants

The recordings were made with 81 preschool and school-aged children (39 female), all monolingual native speakers of Standard Polish living in Szczecin (Northern Poland). Their ages ranged from 2;11–7;11 years old. The parents of the participants signed a consent after being informed about the experiment in detail.5

It should be noted that not all participants were able to finish the perceptual parts, due to different reasons, as e.g. they wanted to go back to their group or we realized that the child did not understand the task. In those cases we terminated the experiment. Also, the classroom teachers informed us that three of the children were developmentally delayed. Finally, we also excluded

---

5 The experiment followed the principles of the Declaration of Helsinki. At the outset of the study, local universities in Poland (Szczecin, the affiliation of the third author) and University Wroclaw did not have an ethical review board.
one child aged 106 months due to a large age difference to the second oldest child (95 months old). In combination, these factors led to the removal of five children. Thus, the final dataset, for both perception and production, came from 76 children (39 female). For the presentation of the results (see section 3) we divided the category Age into five groups as presented in Table 1. However, in our statistical modeling the category Age (in months) functions as a continuous variable.

Table 1: Children’s age groupings used in the graphs

<table>
<thead>
<tr>
<th>Group</th>
<th>Age (year; month)</th>
<th>Number of children</th>
<th>Number of observations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group 1</td>
<td>2;11–3;9</td>
<td>21</td>
<td>903</td>
</tr>
<tr>
<td>Group 2</td>
<td>3;10–4;7</td>
<td>17</td>
<td>785</td>
</tr>
<tr>
<td>Group 3</td>
<td>4;8–5;5</td>
<td>19</td>
<td>884</td>
</tr>
<tr>
<td>Group 4</td>
<td>5;6–6;3</td>
<td>12</td>
<td>534</td>
</tr>
<tr>
<td>Group 5</td>
<td>6;4–7;11</td>
<td>7</td>
<td>318</td>
</tr>
</tbody>
</table>

The total number of items submitted to a further statistical analysis amounted to 3424.

2.2. Experimental design

2.2.1 Data collection protocol

The experimental session began with a short training session. In this training an experimenter gave the instructions to a child by showing the pictures on the screen and asking the child to name them. The experimenter also mentioned that the name of the doll is Kasia (see Fig. 1). Requiring the children to name the pictures first provided some assurance that the younger children would not respond randomly during the perception portions. Next, a short training session was offered for the perception study. The experimenter explained that the child should press the button that corresponded to the appropriate picture. The frame of the picture was of the same color as the button and the horizontal layout of pictures and buttons was parallel; cf. Fig. 2, below). After making sure that the child understood the tasks, the experiment began.

Fig. 1: Pictures illustrating the minimal triplet: /kasa/ “cash register”, /kaʂa/ “groats”, and /kaɕa/ “Cathe, prop.name”
After the training session, the experiment proceeded in three parts. In the first, the children were asked to name pictures displayed one at a time on a screen. In the second, they saw three pictures, listened to the stimuli produced by an adult, and chose the picture of the target word by pressing a button. Finally, in the third part, they listened to their own productions recorded in part I and chose the target word by button-press. During the perception tasks participants’ answers (right/wrong) as well as reaction times (RTs) were recorded.

For the accuracy and reaction time data, our main focus is on differences as a function of speaker (child/self vs. adult) and sibilant. It is widely known that reaction times shorten with development for motoric, linguistic, and cognitive tasks (e.g., de Ruiter et al., 2018; Kiselev et al., 2009; Ren et al., 2021; Thomas & Nelson, 2001), and it can be difficult to separate out motoric contributions to RT changes with age (but see Favilla, 2006 and Kiselev et al., 2009 for explorations into this topic). Along similar lines, increased accuracy with age may reflect both attentional factors and increased acoustic differentiation of the sibilants along with developmental changes in perception. For each age group, we assume that children’s fastest RTs and highest accuracy provide insight into their attentional and motoric capabilities, and focus on differences as a function of speaker and sibilant.

2.2.2 Stimuli

In the picture naming task (part I) participants produced bisyllabic words with initial and medial sibilants /s, ʂ, ɕ/ in a low vowel /a/ context. The word-medial sibilants formed a minimal triplet: /kasa/ “cash register”, /kaʂa/ “groats”, and /kaɕa/ “Cathe, prop.name”. The word-initial sibilants were produced in the words /sanki/ “sleigh”, /ɕatka/ “net”, and /ʂafa/ “wardrobe”. (Note that Polish does not have a minimal triplet with the sibilants in word-initial position that we could employ for a study with young children; we thus include these words mainly to provide a comparison with the more informative minimal triplet.) Finally, there were three distractors: /swɔnkɔ/ “sun”, /ʐaba/ “frog”, and /rɨba/ “fish”. The words were always stressed on the first syllable and they were frequent words of Polish (see Appendix for the exact frequencies in children-directed speech (Table 1) and adult speech (Table 2). The children did not encounter difficulties in naming the pictures. The pictures were shown three times in a randomized order for each participant. Thus, we obtained 27 word recordings for each child (9 words x 3 repetitions). Eight words were used for the perceptual testing. All sibilants were included in the perceptual analysis.

6 We limited our stimulus set to the low vowel /a/ because it is the only available contrast appearing in a minimal triplet appropriate for a production-perception study with children in terms of morpho-phonological simplicity and word frequency. The vowel /a/ also allows a broad comparison of the sibilants in question with other studies, where the low vowel context is dominant. Furthermore, we excluded voiced sibilants for reasons of (i) the unpredictable influence of the laryngeal voicing source on the spectral properties of the sibilants resulting in a much more variable spectral shape, (2) the possibility of a different acquisitional pattern for voiced sibilants compared to voiceless ones and (3) the more valuable cross-linguistic acquisitional comparisons of voiceless sibilants since voiced sibilants are much less common than voiceless ones in the languages of the world (see, e.g., Ohala & Solé. (2010).

7 The matrix presenting the arrangement of the stimuli we used in the experiment is provided in the Appendix. Note that each picture occurred in all three screen locations over the course of the experiment. Each child received a separate randomization.
The same stimulus set was also recorded from one adult female native speaker of Polish. One token of each word was selected as a stimulus for the perception part of the experiment. The choice of the final tokens was made by one of the experimenters and checked by another one, both native speakers of Polish. The tokens showed comparable loudness, pitch and duration. No striking perceptual cues were noticed.

Note that we did not exclude child productions that were perceived by the experimenter to be phonetically ‘in error’ during the naming part of the study. This was in recognition of the growing appreciation for the prevalence of covert contrast in young child speech (e.g., Munson et al., 2010). That is, we allowed for the possibility that some children might produce an atypical contrast that they could, in fact, perceive themselves. Note also that, in the listening study (see next section), children only heard their own productions and that of the adult speaker, i.e. they were never asked to judge the speech of other children.

2.2.3 Design of perception task

In the perception components, the same pictures were shown (see Fig. 1), accompanied by one acoustic stimulus. Here we dispensed with the control item /riba/ “fish” and included only two distractors, i.e. /swɔnkɔ/ “sun” and /ʐaba/ “frog” to limit experiment duration. Children were asked to select on the screen which word they identified based on the presented stimulus. In total, we obtained 48 perceptual responses by each child (8 words x 3 productions/presentations = 24 responses based on adult and 24 on child recordings). The experiment took 15 minutes to complete with a 4–5 min. break between the production and perception parts. During the break, one experimenter engaged the child, who had an opportunity to choose a souvenir to compensate for taking part in the study. A second experimenter extracted one production of each recorded word for presentation in the perception portion. In most cases the first production was selected, unless subsequent productions by that child were clearer or better in quality.

2.3 Instrumentation for data collection

For the purpose of the experiment we built an Ubuntu Mate-based tool which we called Linguistino. This set-up contained a Raspberry Pi single board computer (SCB), a set of buttons, a microphone, and an LCD monitor, as presented in Fig. 2. By making use of Linguistino we were able to use our own design, record children’ voices and adjust the answer box (colours and placements of the buttons) in the perceptual part of our experiment. The reaction times (RTs) were recorded from the onset of the acoustic stimulus which was played after the pictures were shown on the monitor display. Children were not instructed to wait until the end of the word before responding. If the child’s answer in the perception test was incorrect, i.e. did not match the appropriate picture, we removed that response from RT data analysis.

---

8 The adult stimuli were recorded in a sound-proof laboratory in Berlin. There were 7 tokens of each word (9 words, 63 tokens in total). This adult speaker was known to two children who took part in the study. The perceptual results of these children were not evidently different from those of their peers.

9 We also ran an informal perceptual study run with two native speakers of Polish who listened to all childrens’ stimuli and decided which sound they heard. Since they agreed only in 70% we decided not to include the adult annotation in our analysis. The results of this study are also included in the Appendix (Tables A4-A7).
For the needs of the project a program was written in JavaScript in client-server architecture. The layer that a child saw during the experiment (client) was run in the Chromium browser, while other elements such as recording of data (response buttons, response time, sound recording, identification data) were handled on the server side by Node.js. Furthermore, related technologies such as Bootstrap/jQuery were used to write a user interface displayed in the browser.

The experiment took place in a quiet room in the school and kindergarten. The children were seated 15 cm in front of the microphone (Rebel MH-35A) when performing the picture naming task. In the perception part the microphone was turned to the side. The setup is shown in Fig. 2.

![Fig. 2: The experimental setup, showing the microphone, response box, and pictures illustrating the triplet: /kaʂa/ “groats”, /kaɕa/ “Cathe, prop.name”, and /kasa/ “cash register”](image)

2.4. Acoustic methods

For the acoustic analysis, we computed multitaper spectra with a 512 point Hamming window at the temporal midpoint of each sibilant (see Žygis et al., 2012, for the advantages of multitaper over other analysis methods). The audio file sampling rate was 44.1 kHz (16 bits), yielding a single midpoint window of length 11.6 ms. The midpoint was defined as halfway between the onset and offset of the sibilant frication noise. The onset and offset of fricative noises were determined by a forced-alignment algorithm (MausAlign) and then manually corrected for errors based on waveforms and spectrograms in Praat. For the multitaper analysis, power spectral density was estimated via the Thomson multitaper method available in the MathWorks Signal Processing Toolbox Version 6.2 (2007). This used a linear combination with unity weights of individual spectral estimates. The frequency of the highest spectral peak was computed for the full frequency range from 20 Hz to 22050 Hz from the multitaper spectral object. Furthermore, for the full frequency range we computed the four spectral moments (M1–M4) based on the formulae given in Praat (version 5.2, Boersma & Weenink 2018): centre of gravity (COG), standard deviation, skewness, and kurtosis of the spectrum. As laid out above (section 1.2), including the spectral peak along with COG should give a more accurate picture of the spectral shape of Polish sibilants.

Next, formant frequencies F1, F2, F3 were measured at the end of the preceding and at the beginning of the following vowel (based on the onset/offset of stable visible formant structures) using PRAAT’s formant extraction algorithm (“To Formant... burg”) with adapted parameter settings for child vocal tracts (max. formant: 8000 Hz; max. number of formants: 5; window length: 25 ms; pre-emphasis from 50 Hz) as a first approximation of the correct formant value. The automatically extracted spectral maxima of the described Praat algorithm were then carefully manually corrected as needed based on LPC spectra, FFT broadband spectra (at the selected
timepoints) and the FFT broadband spectrogram. Thus, for each timepoint, each automatic measurement value (obtained by Praat’s (adapted) automatic formant extraction algorithm) was manually checked and, if necessary, corrected, based on (1) the temporal formant movement information visible in the spectrogram, (2) the spectral information visible in the broadband FFT spectral slice (5ms window length, identical to the spectrogram settings) and (3) the spectral information visible in a generated LPC spectral slice (Praat LPC object with LPC order = 16, sampling rate of the underlying audio file = 16000 Hz). This hybrid approach of measuring formants for notoriously difficult child speech data based on the combination of the three methods ensures that eventual automatic formant tracking errors are noticed and thus can be manually corrected.

For each CVCV minimal triplet, we also calculated the acoustic durations of the sibilant and the previous and following vowels. For the words with sibilants in initial position, we obtained the duration of the following vowel and the sibilant.

2.5. Statistics

All statistical analyses were conducted in the R Studio software (version 1.1.453, RStudio Team 2018). We describe the methods used in the various results sections in turn. Additional comments about statistical interpretation are provided in the respective methods sections.

2.5.1 Acoustic measures

For acoustic measures, we first ran a random forest model which is used for selection of the most important variables, especially in cases where many potentially collinear predictor variables are available (see Tagliamonte & Baayen, 2012). For this analysis we used the packages party (Strobl, Boulesteix, Kneib, Augustin, & Zeileis, 2008) and caret (Kuhn, 2019). Our random forest consisted of 2000 individual trees, which aimed to determine which parameters excerpted from the acoustic analysis (see (1) below) predicted differences among the three sibilant categories. We also added Age and Sex of the child to the random forest. For each tree, the prediction accuracy was recorded based on the out-of-bag portion. The procedure was repeated after permuting each predictor variable. Finally, the difference between the two accuracies was averaged over all trees and normalized by the standard error (Kuhn, 2019). The only difference with word-initial position was that we did not include F1, F2, F3 and duration of the preceding vowel for the obvious reason that the vowel was not present. In this case the number of parameters was reduced from 14 to 10.

(1) Parameters:

a) F1, F2, and F3 at the end of the preceding vowel (V1),
b) F1, F2, and F3 at the beginning of the following vowel (V2),
c) Duration of sibilant, V1, and V2
d) Frequency of the highest spectral peak
e) COG, standard deviation, skewness, kurtosis

The output of the analysis is a dimensionless measure which shows the relative importance of the variables for the sound classification.

Furthermore, by using the package “VGLM” (Yee & Moler, 2020), a multinomial analysis testing the influence on the variable SOUND with three sublevels [s, ʂ, ɕ] was conducted by including the parameters in (1) as well as participants’ SEX and AGE. We also included interactions of age with parameters (1a)–(e). Prior to the analysis, all predictors were centered to reduce multicollinearity. Whereas the random forest analysis evaluates all three fricatives as a set
(separately for a word-medial and initial position), the multinomial analysis considers each two-way sibilant contrast individually, to assess how the acoustic distinction changes over time.

2.5.2 Perceptual measures

For the perceptual analysis we calculated generalized linear mixed effects models by using the R packages lmer4 (Bates et al., 2018) and afex (Singman et al., 2021). Our models investigated the influence of the fixed effects of (listener’s) AGE, STIMULUS TYPE [child, adult], SOUND [s, ʂ, ɕ], and TRIAL order and their interactions on dependent variables of ANSWER [correct, incorrect] and REACTION TIME. In the case of the nominal variable ANSWER the family in the glmer function was “binomial”. Prior to this analysis we centered AGE and TRIAL. We also used random structures justified by the model including random intercepts and slopes for Participant (listener). Most of our models included only the by-participant slope for Stimulus because adding more random structure resulted in the non-convergence of the models (despite selecting appropriate optimisers and removing correlations between the intercept and slopes). We started with full models, removed non-significant interactions and selected the best fit model by means of likelihood ratio tests using the ANOVA function in R. We ran separate models for words with sibilants in the word-medial position (the triplet) and in the word-initial position which did not create minimal pairs. In linear mixed effects models with Reaction Time as the dependent variable, the RT was logarithmized as it was (positively) skewed. After the log transform, eight individual word productions had values greater than ±3 SDs away from the mean. Since this was such a small quantity of data, spread over word types, we retained these productions in the results. This decision was also in accordance with the three sigma rule (see, e.g., Grafarend (2006).

2.5.3 Relating perception and production

Finally, we conducted another random forest model testing the importance of acoustic parameters of the sibilants in the perception of these sounds following the procedure outlined in section 2.5.1 Doing so provided insight into which acoustic features of the sibilants best predicted the perceptual results for different age groups. Here we also divided the participants according to age groups. We ran those models for /s/ vs. other sibilants, /ʂ/ vs. other sibilants, and /ɕ/ vs. other sibilants.

3. Results

We start with presenting acoustic results, aiming to 1) characterize sibilant production in Polish-speaking children, and gain insight into how the acoustic parameters differentiate the three sibilants. This is followed by perceptual data (accuracy and reaction times), and lastly, analyses that relate the acoustics with the children’s perception. In all cases we present results pertaining to the word-medial contrast, i.e. /kasa/, /kaʂa/, and /kaɕa/ and to the word-initial position, i.e. /sanki/, /ćatka/, and /ʂafa/ separately, as the first set creates a minimal triplet and the second set differs in sounds other than the sibilants.

3.1. Acoustic results of child productions

Fig. 3 presents multitaper spectra of /s/, /ʂ/ and /ɕ/ from all child productions (three for each word) excerpted at the midpoint of frication in the word-medial position, split by the five age groups. The spectral plots for the word-initial position are presented in the appendix of this paper (Appendix section 2.2). The spectra for individual productions are shown by green lines and the
overlaid mean spectrum is represented by the black line. We observe large inter-speaker variation in the children’s spectra for all sibilants. It can be inferred that, as expected, children’s productions become more distinct from the early ages (2;10 to 3;9 years) towards the older age groups (e.g. 6;3 to 7;11): The /s/ and /ʂ/ spectra for the youngest age group look almost identical with no clear separation of the first spectral peak at around 5kHz and the second major peak at around 10 kHz, whereas in the older age group the two sibilants /s/ and /ʂ/ are clearly separated by spectral shape, with /s/ emphasizing the front cavity peak around 10kHz and /ʂ/ emphasizing the front cavity peak around 5kHz. These spectral profiles are similar to those seen for adult speakers (see spectral plots for the adult speaker of this study in section 3.2). In contrast, the sibilant /ɕ/ does not show consistent or major changes in spectral shape from younger to older children.
Fig. 3: Individual multitaper spectra (green lines) with overlaid mean spectrum (black lines) obtained for all /s/ productions (top panel), for all /ɕ/ productions (middle panel), and for all /ʂ/ productions (bottom panel) for the minimal pairs in word-medial position. The columns show the different age groups of the children, from the youngest age group (2;10 to 3;9 years) in the left column to the oldest age group (6;3 to 7;11 years) in the right column.

In the following we will quantify differences among fricatives by means of the acoustic parameters listed in (1) above. We start with a random forest analysis which will help us to
understand which parameters play a major role in differentiating the three sibilant categories in production.

3.1.2 Random forest analysis of sibilant productions: Acoustic classification of /s/, /ʂ/, /ɕ/

This analysis (see section 2.5.1) compares the relative reliability of different acoustic features in child productions to separate the data based on the target. We start with contrasts in the word-medial position and then proceed to the analysis of the word-initial position.

The random forest analysis for word-medial position reveals that the second formant of the following vowel and centre of gravity are the most important cues in sibilant categorization. They are followed by spectral cues of frication such as skewness, peak frequency, and standard deviation as well as formants including F2 of the preceding vowel, and F1 and F3 of the following vowel. From the other predictors age also plays a role but its contribution is rather small. Fig. 4 presents the results for the minimal triplet. Note that values from the acoustic analysis are scaled after conversion to the range [0, 100]. The maximum value, i.e. the maximum importance, is scaled to 100 and the minimum to 0. The remaining values are scaled proportionally and shown by means of the horizontal bars. Thus, the latter bars do not provide any information about the significance of the predictors but only show the relative importance of different acoustic cues to sibilant categorization for the children's stimuli.

![Variable importance of different acoustic features for separating the word-medial sibilants in children’s productions](image)

The results of this random forest analysis are in line with previous literature (Nowak, 2006; Żygis & Padgett, 2010) where F2 measured at the beginning of the following vowel and centre of gravity (COG) of the sibilant spectrum were shown to be most decisive in the classification of Polish sibilants in adult listeners.

Fig. 5 shows how these two measures separate fricatives across all five child age groups. It appears that in the youngest children the sibilants are almost randomly spread. It also seems, however, that in the second group the alveolo-palatal /ɕ/ is least spread and as such may have started to separate from other sounds early. In the third group, where children have already finished the fourth year of life, the alveolo-palatal is indeed the least spread sound and its separation from other sounds continues. Finally, the retroflex and the dento-alveolar sibilants start to separate. A
few sibilant productions still overlap in the oldest group, but it should be taken into account that only two acoustic parameters, i.e. COG and F2 frequency, are reflected in this representation (see (1) for other parameters).

Fig. 5: COG and following F2 frequency for sibilants in the word-medial position across the five age groups. To improve readability of the figure, 37 extreme data points with F2s that were higher than 3200 Hz and lower than 1000 Hz were removed.

The following figures will further illustrate how the two parameters, F2 of V2 and COG, develop over time. If we consider F2 independently of COG it turns out that this formant separates the alveolo-palatal right at the beginning of acquisition process; see Fig. 6. F2 does not reliably separate /s/ and /ʂ/. 
Fig. 6: F2 of the vowel following /s, ɕ, ʂ/ in the word-medial position as a function of age

By contrast, COG does not contribute to the differentiation of the sounds at the very beginning because this spectral parameter overlaps for all sibilants. It also does not contribute to differentiating the alveolo-palatal to the extent that F2 does. However, it starts to contribute to the separation of /s/ after the third year; see Fig. 7.

Fig. 7: COG of word-medial /s, ɕ, ʂ/ spectra as a function of age

As far as word-initial position is concerned, a similar scenario is found: The random forest analysis shows that F2 of the vowel following the sibilant and the centre of gravity are the most important cues in sibilant categorization; see Fig. 8. The duration of the following vowel is also
important although less influential.\textsuperscript{10} Other parameters such as age, sex, skewness, sibilant duration, and frequency peak are of secondary importance. See also Figs. A3 and A4 in the Appendix showing F2 and COG of sibilants in the word-initial position.

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{importance.png}
\caption{Variable importance of different acoustic features for separating the word-initial sibilants in children’s productions}
\end{figure}

In summary, F2 of the following vowel and COG are the most decisive parameters in both word-medial and initial position. The order of other variables whose contribution was secondary was different for the two positions. Some of these differences presumably arise from word structure, viz. the fact that a minimal triplet was not available for initial position.

3.1.3 Multinomial analysis: Pairwise sibilant contrasts as a function of age

Whereas the previous analysis has shown which parameters are important for sibilant classification, the multinomial analysis (section 2.5.1) compares the sibilants pairwise in interaction with child Age (again leaving out the adult data). The analysis will show which parameters play a significant role in the development of children's production of the Polish sibilant contrast, i.e. how individual sound categories continue to be better differentiated over time, evaluating each sibilant contrast separately. This additional method supplements the random forest

\textsuperscript{10} We note that the medial sounds differed among the three words used to assess the word-initial contrast, namely /sanki/, /catka/, and /ʂafa/. Shortening of vowels before voiceless consonants in syllable coda position has been widely noted across languages (e.g., Chen, 1970), i.e. this syllable-level factor might provide an explanation for the role of vowel duration in these words. However, Keating (1985) reported finding no such durational difference for Polish-speaking adults. Most work on this question has assessed vowel duration before single voiced vs. voiceless obstruents, whereas our stimuli included clusters. On the basis of our data, therefore, it is not possible to determine whether the importance of vowel duration here reflects characteristics of the syllable onsets, viz., the sibilants themselves, or the syllable codas.
analysis, which is based on a different statistical procedure including random choice of the parameters each time the modeling is performed.

The analysis (see Table 2) reveals that formants at the boundary of the following vowel are of particular importance in separating /ɕ/. Whereas for the /s/ vs. /ɕ/ contrast F2 and F3 of the following vowel obtained the highest z-values, F2, F1, and F3 play the most important roles in the creation of the /ɕ/ vs. /ʂ/ contrast over time. Other parameters that were significant but whose contributions were smaller were the spectral properties of standard deviation and kurtosis, and F1 of V2 for /s/ vs. /ɕ/, and standard deviation and COG for /ɕ/ vs. /ʂ/. Finally, it is striking that for the /s/ vs. /ʂ/ contrast the COG played the most important role, followed by F3 of the following vowel and spectral properties standard deviation and kurtosis. F2 of the following vowel was not important at all, but instead F2 of the preceding vowel played a role followed by F1 of the following vowel.
Table 2: Results of the multinomial analysis for the sibilants in word-medial position based on children’s stimuli

<table>
<thead>
<tr>
<th>Word-medial</th>
<th>Coefficient</th>
<th>Odds ratio</th>
<th>z-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>a) /s/ vs. /ɕ/</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F2 of V2*Age</td>
<td>-1.783</td>
<td>0.168</td>
<td>-7.62</td>
<td>p&lt;.001</td>
</tr>
<tr>
<td>F3 of V2*Age</td>
<td>0.871</td>
<td>2.390</td>
<td>4.38</td>
<td>p&lt;.001</td>
</tr>
<tr>
<td>STD*Age</td>
<td>0.908</td>
<td>2.479</td>
<td>4.18</td>
<td>p&lt;.001</td>
</tr>
<tr>
<td>Kurtosis*Age</td>
<td>0.736</td>
<td>2.088</td>
<td>3.32</td>
<td>p&lt;.001</td>
</tr>
<tr>
<td>F1 of V2*Age</td>
<td>0.345</td>
<td>1.412</td>
<td>2.26</td>
<td>p&lt;.05</td>
</tr>
<tr>
<td>b) /ɕ/ vs. /ʂ/</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F2 of V2*Age</td>
<td>1.485</td>
<td>4.417</td>
<td>6.49</td>
<td>p&lt;.001</td>
</tr>
<tr>
<td>F1 of V2*Age</td>
<td>-0.566</td>
<td>0.910</td>
<td>-3.77</td>
<td>p&lt;.001</td>
</tr>
<tr>
<td>F3 of V2*Age</td>
<td>-0.492</td>
<td>0.610</td>
<td>-2.39</td>
<td>p&lt;.05</td>
</tr>
<tr>
<td>STD*Age</td>
<td>-0.523</td>
<td>0.747</td>
<td>-2.36</td>
<td>p&lt;.05</td>
</tr>
<tr>
<td>COG*Age</td>
<td>0.373</td>
<td>1.452</td>
<td>2.01</td>
<td>p&lt;.05</td>
</tr>
<tr>
<td>c) /s/ vs. /ʂ/</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>COG*Age</td>
<td>0.523</td>
<td>1.687</td>
<td>3.98</td>
<td>p&lt;.001</td>
</tr>
<tr>
<td>F3 of V2*Age</td>
<td>0.493</td>
<td>1.638</td>
<td>3.17</td>
<td>p&lt;.01</td>
</tr>
<tr>
<td>STD*Age</td>
<td>0.399</td>
<td>1.491</td>
<td>2.56</td>
<td>p&lt;.05</td>
</tr>
<tr>
<td>Kurtosis*Age</td>
<td>0.430</td>
<td>1.537</td>
<td>2.32</td>
<td>p&lt;.05</td>
</tr>
<tr>
<td>F2 of V1*Age</td>
<td>-0.345</td>
<td>0.707</td>
<td>-2.28</td>
<td>p&lt;.05</td>
</tr>
<tr>
<td>F1 of V2*Age</td>
<td>-0.257</td>
<td>0.746</td>
<td>-1.99</td>
<td>p&lt;.05</td>
</tr>
</tbody>
</table>

For the word-initial position, a different scenario is found. For all three contrasts it is the COG which plays the most important role over time. In fact, it is the only parameter for the /ʂ/ vs. /ɕ/ contrast which significantly interacted with age; see Table 3.
Table 3: Results of the multinomial analysis for the sibilants in word-initial position based on children’s stimuli

<table>
<thead>
<tr>
<th>Word-initial</th>
<th>Coefficient</th>
<th>Odds ratio</th>
<th>z-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>a) /s/ vs. /ɕ/</td>
<td>COG*Age</td>
<td>0.545</td>
<td>1.724</td>
<td>3.82</td>
</tr>
<tr>
<td></td>
<td>Sex*Age</td>
<td>0.563</td>
<td>1.756</td>
<td>2.18</td>
</tr>
<tr>
<td>b) /ʂ/ vs. /ɕ/</td>
<td>COG*Age</td>
<td>-0.282</td>
<td>-0.125</td>
<td>-2.39</td>
</tr>
<tr>
<td>c) /s/ vs. /ʂ/</td>
<td>COG*Age</td>
<td>0.761</td>
<td>2.141</td>
<td>5.11</td>
</tr>
<tr>
<td></td>
<td>F2 of V2*Age</td>
<td>-0.708</td>
<td>0.492</td>
<td>-3.06</td>
</tr>
<tr>
<td></td>
<td>Duration of V2*Age</td>
<td>-0.295</td>
<td>0.744</td>
<td>-2.11</td>
</tr>
</tbody>
</table>

In summary, the analysis shows that several parameters play a significant role in the process of sibilant acquisition in Polish. In particular, for the word-medial alveolo-palatal /ɕ/, formants of the following vowel are essential. It was not only F2 but also F1 and F3 at the following vowel onset which differentiated the /ɕ/ from /ʂ/ and /s/ in the acquisition process. For the /ʂ/ vs. /s/ contrast, F2 of the following vowel turns out to be not significant, but instead spectral properties such as COG, STD and kurtosis (as well as formants) reflect the acquisition of the contrast. The word-initial contrast where the items differed in other sounds did not show this result. In all cases COG was a decisive parameter.

3.2 Perceptual results

In the following sections we describe the results of the perception experiment for the children judging (i) their own production of the three sibilants and (ii) the adult’s sibilant productions. The statistical analyses are described in section 2.5.2.

Before presenting the perception results we will give the acoustic properties of the adult productions for the perception experiment in order to be able to interpret the perceptual results for different speakers (self vs. adult). As with the children’s data, we present here the results for the medial contrast; data for the initial sibilants are given in the appendix. These plots reflect the single production that was used in the perception experiment.
Fig. 9: Multitaper spectra measured at the acoustic midpoint for the adult sibilant stimuli in the medial position. The left panel shows the /s/ spectrum, the middle panel the /ɕ/ and the right panel the /ʂ/ production for the minimal triplet.

Table 4: Parameters of the acoustic analysis of the adult productions used in the perception experiment (medial position)

<table>
<thead>
<tr>
<th>Parameter</th>
<th>kasa</th>
<th>kacea</th>
<th>kaşa</th>
</tr>
</thead>
<tbody>
<tr>
<td>duration [ms]</td>
<td>157</td>
<td>169</td>
<td>155</td>
</tr>
<tr>
<td>spectral peak [Hz]</td>
<td>7407</td>
<td>4307</td>
<td>2756</td>
</tr>
<tr>
<td>centre of gravity [Hz]</td>
<td>7752</td>
<td>5120</td>
<td>2611</td>
</tr>
<tr>
<td>SD [Hz]</td>
<td>1302</td>
<td>1790</td>
<td>528</td>
</tr>
<tr>
<td>skewness</td>
<td>1.8571</td>
<td>2.6047</td>
<td>5.6464</td>
</tr>
<tr>
<td>kurtosis</td>
<td>9.1332</td>
<td>10.99</td>
<td>83.2736</td>
</tr>
<tr>
<td>formant values (F1, F2, F3 ) for the preceding vowel [Hz]</td>
<td>917, 1676, 2620</td>
<td>920, 1860, 2650</td>
<td>871, 1547, 2898</td>
</tr>
<tr>
<td>formant values (F1, F2, F3) for the following vowel [Hz]</td>
<td>840, 1552, 2729</td>
<td>854, 1876, 2700</td>
<td>805, 1580, 2783</td>
</tr>
</tbody>
</table>

All acoustic parameters, i.e. the acoustic spectra and the measurements of the most important acoustic parameters (see Fig. 9 and Table 4) are in line with spectral shapes and acoustic parameters given in the literature for the three Polish sibilants (see e.g., Lisker 2001; Nowak 2006; Żygis et al. 2017).

3.2.1 Labeling accuracy

Our perceptual results show that children discriminate the word-medial contrast /s, ʂ, ɕ/ better when produced by an adult, and their labeling accuracy is significantly worse when they listened to their own stimuli (t = -4.05, p<.001). Not surprisingly, the discrimination improves with increasing age (t = 5.81, p<.001); see Fig. 10. Recall that we have divided the category Age into five groups for illustrative purposes (see Table 1 above). However, in our statistical modeling the category Age functions as a continuous variable (provided in months).

It is important to note that perceptual accuracy improved with age for both speakers (adult and child/self). When children listened to the adult speaker, increasing accuracy with age can be ascribed to a combination of cognitive (e.g., attentional) factors as well as increased perceptual acuity. For some speaker/sibilant conditions (see Fig. 12), accuracy was quite high even in the
youngest children, i.e. attentional factors seem to play only a small role. In the case of child-produced speech, on the other hand, increasing accuracy with age presumably also reflects children’s greater production accuracy. In this analysis, TRIAL was not significant, i.e. accuracy did not change over the course of the experiment.

![Graph showing proportions of correct/incorrect answers as a function of stimulus and age for the triplet](image)

**Fig. 10:** Proportion of correct/incorrect answers as a function of the stimulus (adult, child) and age for the triplet.

Furthermore, our results reveal that, on average, /ɕ/ is the easiest sound for children to discriminate, but the difference is only significant with respect to /s/ (t = -2.42, p<.05). However, if we look at the stimulus type, a slightly different pattern emerges (see Fig. 11 and Table 5, left columns). For the adult stimuli, /ɕ/ obtains a slightly higher number of correct answers as compared to /ɕ/ but the difference is not significant. Hence, /ɕ/ and /ʔ/ are the easiest sounds to discriminate. The most difficult sound is the dento-alveolar /s/. When children listen to the stimuli produced by themselves, the alveolo-palatal /ɕ/ obtains the highest number of correct answers; /ʔ/ and /s/ receive lower and almost equal numbers of correct answers (see Table 5). This difference is also not significant. The interaction StimulusType*Sound for all stimuli types is significant (Stimulus [Child]*/ʔ/: t = -2.84, p<.01). Fig. 12 shows accuracy for both stimuli types and the three sibilants across ages. Three children achieved 35.3%, 40.9% and 47% of correct answers when listening to

---

11 An anonymous reviewer asked whether older children's choices are more correct due to their improved discrimination capacity or because they are more mature, “which allows them to stay focused and not get distracted easily”. Results presented in Figure 12 clearly show that even very small children (2;11–3;9) are able to discriminate the sibilant-initial words very well when they are produced by an adult person (about 95% of correct answers). That is, about a 5% error rate could potentially be ascribed to age-dependent errors in perceiving correctly-produced tokens.

12 We selected only the adult stimuli when calculating this difference and conversely, we selected only childrens’ data when calculating differences regarding the perception of childrens’ stimuli.
their stimuli and one child achieved 40% correct answers for the adult stimuli. Other children were above 50%. Note that 33% is the chance level due to three answer possibilities.

Fig. 11: Probabilities of correct answers as a function of the stimulus (adult, child) and the word-medial contrast /s, ʂ, ɕ/, averaging all child age groups. The dots show mean values and the whiskers 95% confidence intervals.

Fig. 12: Proportion of correct/incorrect answers as a function of the stimulus (adult, child) and age for the word-medial contrast.
Table 5: Proportion in % of correct answers obtained for stimuli produced by adults and children for sibilians in word-medial (left) and word-initial (right) positions. In parentheses the number of answers is provided.

<table>
<thead>
<tr>
<th></th>
<th>ka/s/a</th>
<th>ka/ʂ/a</th>
<th>Ka/ɕ/a</th>
<th>/s/anki</th>
<th>/ʂ/afa</th>
<th>/ɕ/atka</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Stimulus_Adult</strong></td>
<td>81.2 (173)</td>
<td>92.9 (197)</td>
<td>89.2 (190)</td>
<td>97.6 (205)</td>
<td>96.2 (205)</td>
<td>95.7 (198)</td>
</tr>
<tr>
<td><strong>Stimulus_Child</strong></td>
<td>59.6 (130)</td>
<td>60.9 (131)</td>
<td>72.7 (160)</td>
<td>97.2 (205)</td>
<td>94.9 (204)</td>
<td>90.6 (193)</td>
</tr>
<tr>
<td><strong>Summary</strong></td>
<td>70.3 (303)</td>
<td>76.8 (328)</td>
<td>80.8 (350)</td>
<td>97.4 (410)</td>
<td>95.6 (409)</td>
<td>93.1 (391)</td>
</tr>
</tbody>
</table>

For words with sibilians in the word-initial position, the results show that the adult stimuli again trigger a higher number of correct answers than children’s (t = 2.89, p<.01). The perception also improves with increasing age (t = 4.19, p<.001). The results are presented in Fig. 13. As expected for these non-minimal contrasts, there were more correct answers as compared to the triplet: 95.4% of words were correctly recognized with sibilians in the word-initial position, whereas the discrimination varied between c. 70% and 80% for words with word-medial sibilians (see Table 5).

![Fig. 13: Proportion of correct/incorrect answers as a function of the stimulus (adult, child) and age for the word-initial sibilians](chart.png)

When we look at individual sounds, the results show that most incorrect answers were obtained when /ɕ/atka and /ʂ/afa needed to be identified (/ʂ/ vs. /s/ n.s.) and the highest number of correct answers was obtained for /s/anki; see Fig. 14. Our statistical modeling shows a significant difference between /ɕ/atka and /s/anki (t = -3.00, p<.01). The interaction StimulusType*Sound is not significant, i.e., the patterns are comparable for child- and adult-produced stimuli.
3.2.2 Reaction times

The reaction time data for word-medial position reveal that children identified /ɕ/ most easily (the shortest RT), followed by /s/ (t = 3.43, p<.001) and /ʂ/ (t = 3.10, p<.01). However, no significant difference was found in the reaction times for /s/ and /ʂ/. RTs decrease with increasing age (t = -10.93, p<.001) and increasing TRIAL number (t= -5.25, p<.001). Fig. 15 presents the results.
Fig. 15: Reaction time as a function of age in the word-medial contrast. The boxes correspond to the 25th to 75th percentile range, black horizontal lines represent medians, and whiskers correspond to ±1.5 inter-quartile range; outliers, i.e., data above or below this range are represented as dots.

When children listened to words with sibilants in word-initial position (with other differing sounds as well) the only significant factors that influenced the RT were AGE and TRIAL. Younger children were slower in choosing the answer than older ones ($t = -10.05$, $p<.001$) and children’s reaction times were quicker in later trials ($t = -4.42$, $p<.001$). There was no effect of stimulus type or sound (see Fig. 16).
Finally, the reaction time was shorter when children listened to their own stimuli as opposed to the adult stimuli (mean RT = 2.00 s for adult stimuli and 1.78 s for their own stimuli). However, this result applies only to contrasts appearing in the word-medial position (t = -3.48, p<.001). Fig. 17 presents RTs for the adult and child stimuli for sibilants in the word-medial position (left) and word-initial position (right).

Fig. 16: Reaction time as a function of age in the word-initial contrast. The boxes correspond to the 25th to 75th percentile range, black horizontal lines represent medians, and whiskers correspond to ±1.5 inter-quartile range; outliers, i.e., data above or below this range are represented as dots.

Fig. 17: Reaction time as a function of the stimulus (adult, child) and age in the word-medial position (left) and word-initial position (right). The boxes correspond to the 25th to 75th percentile range, black horizontal lines represent medians, and whiskers correspond to ±1.5 inter-quartile range; outliers, i.e., data above or below this range are represented as dots.

3.3. Random forest analysis relating children’s self-perception to sibilant acoustics

In this section we explore which acoustic parameters of the child-produced fricatives (only those selected for the perceptual experiment) are most closely related to the children's perceptual responses. For this analysis we focus on the minimal triplet, since those words are differentiated by the sibilants only, and not other sounds in the word, and therefore provide the clearest picture of sibilant perception. To this end we conducted a random forest analysis by using the children's identification of a certain sibilant (e.g. they clicked on [s] after hearing a stimulus). For this purpose we conducted three separate random forest analyses, one for selecting [s] and not the other two sibilants, one for [ɕ] and not selecting the other two sibilants, and finally for selecting [ʂ] and not the other two sibilants. These selections entered the random forest model as dependent variables. We used the rich acoustic parameter space of all children’s speech productions as predictors, thus using all mentioned acoustic parameters (of child productions, see section 3.1) as predictors of the model. The aim of this analysis was to find out which acoustic parameters of the child-produced sibilants would drive a response for a certain sibilant (and thus not the other two) from the children. Please note that this random forest analysis is only adequate for the children's productions with their high acoustic variability. It is not possible to conduct a comparison for the adult stimuli since all children judged the same, identical stimuli, so those productions lack the necessary acoustic variability to serve as predictors. Fig. 18 presents the strength of each parameter in children’s perception of all three sibilants.
The results show that the spectral property COG is a decisive parameter for /s/ followed by skewness. As far as /c/ vs. other sounds is concerned, F2 of the following and preceding vowel plays the most important role, again followed by skewness. For /s/ vs. other sounds it is the spectral parameter COG followed by F2 of the preceding vowel which play the most pronounced role in the children's perception.

We also split the results according to the age groups to see if changes in the children's perception could be observed. The data (see Fig. 19) reveal that for the youngest children (2;11–3;9 years) it is a spectral moment (skewness) and formants which play the most important role in the perception of [s] vs. other sounds. It is also striking that the F2 of the following vowel is almost as important as skewness (the bars are of approximately equal length). At later ages the spectral properties such as frequency peak, skewness, and COG become more important than formants. Note that all parameters were included in the analysis but only the first five were shown in these figures for the sake of readability.
In the case of /ɕ/ vs. other sounds, formants, especially F2 of the preceding and following vowels, provide the most important cues for the perception across all age groups apart from the oldest one where frequency peak followed again by F2 of the preceding and following vowel are the most important cues; see Fig. 20.

Fig. 20: Perception of Ka/ɕ/a vs. other words by children split by age groups

Finally, in the case of /ʂ/ vs. other sounds the results evince that younger children rely more on formants of the preceding vowel and following vowel whereas older children concentrate on COG along with vowel formants. For the oldest group it is COG and duration of the preceding and following vowel which are most relevant; see Fig. 21.
In summary, the results show that different parameters are responsible for perceptual classification of sibilants in children of different ages. However, if we look at younger children it appears that they pay more attention to formants independent of the sibilant and the cue weighting changes during the acquisition process.

4. Discussion

The goals of this work were a) to explore whether and how children’s labeling of sibilants depended on whether they were listening to an adult or themselves; b) to relate acoustic measures of the fricatives to the perceptual results; and c) to assess the order of acquisition of the three Polish sibilants, and compare it to other languages with complex sibilant systems. We first comment on our perceptual methods and results (section 4.1) and then address these three issues in turn.

4.1 Listener age and perceptual results

As noted in section 2.2.1, developmental changes in motor control and attention may affect measures of reaction times and accuracy. Such factors should apply equally to all stimuli, however. For present purposes, more important than overall increases in accuracy and shorter reaction times with age are how those measures differed as a function of speaker and sibilant. Accuracy rates of 95% for adult-produced sibilant-initial words in the youngest age group provide assurance that these children were not behaving randomly. Not surprisingly, accuracy for the word-medial (minimal triplet) case was lower, but still much better than chance (76%). A method that does not require a manual response, such as eye tracking, may show better discrimination than what we observe; nevertheless, our data do show that the task was well within the capabilities of the children, so that it is valid to assess differences across conditions (self-produced vs. adult speech and across the three sibilants).

4.2 Speaker age effects on perceptual results

Various theoretical considerations have suggested possible differences between perception of child (or self-produced) vs. adult speech, but those predictions are conflicting and few empirical studies present data that address this question. The current results consistently show that children’s labeling accuracy for both word-medial and word-initial /s, ʂ, ɕ/ was significantly higher for words produced by an adult than when they listened to themselves. These results do not show an advantage for self-produced speech. They are, however, consistent with the findings of Hazan and Markham (2004) for adolescents, as well as Cooper et al. (2018) for toddlers, showing a general advantage when listening to adult forms. The acoustic data provide some explanation for this: At younger ages, the three Polish sibilants are less well-separated by their contextual formants (F2 in flanking vowels) and spectra (COG values).

It should be noted that in the perceptual part of the experiment the children heard the adult stimuli first, followed by their own productions. Conceivably, poorer labeling accuracy for child speech could be attributed to a fatigue effect. However, the short duration of the study (about 15 minutes) likely mitigated this concern.

---

13 We note that the literature on covert contrast, as most traditional studies of child speech, has been based on adult perception of child forms. Indeed, covert contrast itself has been framed as an acoustic contrast that adults may not perceive (cf. Li et al., 2009). To our knowledge, evidence that typical children can perceive their own covert contrasts has been anecdotal.
minutes in total, with a break between the production and perception parts) should be well within the attention span of children of this age. Faster reaction times for self-produced speech are also at odds with an explanation based on fatigue; on the contrary, they suggest a learning or a practice effect, as has been observed in other RT studies in children (e.g., Thomas & Nelson, 2001). Finally, if fatigue effects on RTs were confounded with age, we would expect the adult-child RT difference to be the greatest in the youngest children. Visual inspection of RTs for adult and child stimuli (Fig. 17) does now show such a pattern.

An interesting aspect of the results is that children’s RTs were slightly faster for self-perceived speech in medial position (the minimal triplet), yet their accuracy was nevertheless lower than for the corresponding adult forms. Although the average difference was small (22 ms: 2.00 s for the adult speaker vs. 1.78 s for child speakers), this is a paradoxical finding. We suggest the following as a possible explanation. Several studies have provided data to support the hypothesis that children show more extreme coarticulatory behavior than adults (e.g., Nittrouer et al., 1989; Noiray et al., 2019)—with the important caveat that consonant identity and the degree of coarticulatory resistance may affect such patterns (Noiray et al., 2019). We hypothesize that vowel formant patterns may diverge as a function of the upcoming consonant earlier in child than adult speech. Such anticipatory information would only be available for the medial fricative, and not the word-initial case. Indeed, the random forest analysis for the word-medial acoustics (Figure 4) does show some contributions of V1 measures to the acoustic differentiation of the triplet. Further, V1 formant patterns are available in the acoustic signal earlier than the sibilant noises. This information in V1 could lead children to a faster decision for their own voices for medial consonants. However, the variability inherent to child speech also means that such formant information may be noisy, so that such earlier responses are not necessarily more accurate. Additional analyses would be required to verify this possibility.

4.3 Relating perception and production

Unlike past developmental studies of Polish, we obtained acoustic data, which provide insight into our perceptual findings and the acquisitional order in production. The analyses of the acoustic data show that several parameters play important roles in children’s differentiation of the Polish sibilants. In particular, it turns out that for the word-medial alveolo-palatal /ɕ/ the formants of the following vowel are essential. It was not only F2 but also F1 and F3 at the following vowel onset which separated /ɕ/ from /ʂ/ and /s/ in the acquisition process (Table 3). For the /ʂ/ vs. /s/ contrast, F2 of the following vowel turns out to be not significant, but instead spectral properties such as COG, STD and kurtosis (as well as formants) reflect sibilant acquisition in production. The random forest analysis for the children as a group also shows that F2 was important in driving perception of /ɕ/, whereas spectral features were more important for /ʂ/ and /s/.

The random forest analysis predicting the perceptual data from stimulus acoustics in different age groups showed that the youngest children consistently relied on F2 in differentiating the three sibilants; in older children the noise characteristics of the fricatives became more important. This is consistent with previous findings indicating that younger children rely more on dynamic than static information in the process of phonemic categorization (Nittrouer, 2002; Nittrouer & Studdert-Kennedy, 1987), possibly a result of children being more attuned to holistic word patterns than phonetic details of individual segments (Nittrouer and Studdert-Kennedy, 1987; Vihman, 2015; Walley, 2008; Werker & Tees, 1999). In the current context, two sets of perceptual findings are particularly germane: Nittrouer and Studdert-Kennedy (1987) observed that children’s labeling slopes across a fricative continuum were shallower than those of adults. Further, Nittrouer (2002)
observed that partial correlations between labeling functions and acoustic characteristics of formant and fricative spectra showed a pattern of increasing correlation coefficients with age for fricative spectra, and decreasing coefficients for vowel formants. This could explain why the separation of /s/ and /ʂ/ in perception takes place later: Not only are the productions acoustically less distinct (see Fig. 5), but the sounds are mainly separated by fricative spectral parameters such as COG, to which young children are less sensitive.

Our results do suggest one modification to the conclusions of Nitttrouer and Studdert-Kennedy: That study, focusing on the /s/-ʃ/ distinction in English, described a general developmental movement away from an early reliance on formants to a later reliance on fricative noise characteristics. The current data (see Figs. 19–21) suggest that not all sibilants show the same developmental trajectory in this respect. Our data for /s/ show a pattern of early reliance on formants, shifting later to more emphasis on spectral characteristics of the sibilant. The /ɕ/ was mainly differentiated by formants across ages. The data for /ʂ/ are the most complex (cf. also Bukmaier & Harrington, 2016), with children of different ages alternating between reliance on formants and spectral features. Even the oldest children in our study were still perceiving self-produced /ʂ/ at only about 65% accuracy (see Fig. 5), suggesting that acquisition of this fricative in production may last well into the school-age years. Additional cross-language data are needed to further understand how formants vs. fricative spectral features are used in children’s perceptual processes, and how those processes change over time.

4.4 Order of acquisition in Polish and Putonghua

Here we focus on the minimal triplet, where the sibilant is critical to differentiating words in production as well as perception. In this condition, we found that labeling accuracy for /ɕ/ was fairly high for both child and adult productions. (Qualitatively, for adult productions, the accuracy was highest for /ʂ/, but the difference between /ʂ/ and /ɕ/ was not significant). RTs were also fastest for /ɕ/. The acoustic data also showed (Figs. 5, 6, 7) that the alveolo-palatal was differentiated in children's production before the other fricatives. Based on these data, we propose that /ɕ/ is the earliest sound acquired by Polish-learning children, followed by /s/ and /ʂ/. Even in the oldest age group, perceptual accuracy for child-produced /ʂ/ remained rather poor (Fig. 5). This finding, in combination with high perception accuracy for adult-produced /ʂ/, suggests that children’s low perception accuracy reflected poor acoustic differentiation, i.e. production difficulties. Our conclusions are generally in line with the order /ɕ/ > /s/ > /ʂ/ proposed in previous studies of Polish (e.g., Kaczmarek, 1953; Łobacz, 1996; Soltys-Chmielowicz, 1998; Łukaszewicz 2000; Bryndal 2015), whose conclusions were based on small data sets or obtained via impressionistic adult perception. Our study, based on acoustics and children’s own perceptions, also suggests that the ranking between /s/ and /ʂ/ may vary across children. Further, the children's reaction times for their own stimuli (Fig. 15) were consistent in showing the fastest RTs for /ɕ/, but the order for /s/ and /ʂ/ varied, suggesting some individual variation.

We can compare these findings those for /s ʂ ɕ/ in Putonghua, the only other three-way sibilant system for which, to our knowledge, developmental data are available (Hua & Dodd, 2000; Li & Munson, 2016). In both Polish and Putonghua, /ɕ/ is the first acquired sibilant, but in Mandarin the second sibilant to emerge is the retroflex /ʂ/ and the latest is /s/ (Li and Munson, 2016). Following Li and Munson (2016) and Li (2012), we hypothesize that the early acoustic differentiation of the alveolo-palatal in both languages is due to motor control mechanisms: /ɕ/ is produced with the tongue dorsum, and control over this articulator should precede that for the fine movements of the apex or tongue blade required to produce /s/ or its retroflex counterpart /ʂ/. High
perceptual accuracy for both adult- and child-produced /ɕ/ may also relate to the nature of early child speech perception (cf. section 4.3), in which children seem to be more sensitive to formant patterns than fricative spectral noises. Finally, although quantitative studies are still lacking, it has been reported that /ɕ/ is commonly used in motherese in Polish, i.e. adults when speaking to children replace /s/ and /ʂ/ by [ɕ] (Czaplicki et al., 2016). This is probably linked to sound iconicity where palatal and palatalised sounds are associated with “smallness”. One possible explanation of this phenomenon lies in the higher frequency regions characterizing palatal sounds that are associated with smallness (Ohala 1994). As described in the next paragraph, increased frequency of a form could facilitate acquisition.

Regarding the other sibilants, we suggest that the frequency of occurrence of the sibilants in those languages may play a role. As shown by Li and Munson (2016), the frequency order of all sibilants in Putonghua is /ʂ/ > /ɕ/ > /s/), i.e. /ʂ/ occurs more frequently than /s/ does. On the contrary, in Polish, the frequency order of all sibilants is /s/ > /ɕ/ > /ʂ/, i.e. /s/ occurs more often than /ʂ/; see Kłosowski (2017). The relative rarity of /ʂ/ may make it more difficult than /s/ for at least some children. Frequency effects could impact a) the strength of phonological representations (e.g., Pierrehumbert, 2003) and/or b) effects of greater practice in perception and production (e.g., Edwards, Beckman, and Munson (2015). Edwards et al. (2015) drew their conclusions from the cross-linguistic paidologos [παιδολογος] project (http://www.ling.ohio-state.edu/~edwards). Our results for children 3–8 years may provide some support for the conclusions of that research group, who assessed acquisition over ages 2–5 years.

As pointed out by a reviewer, another possible explanation of acquisitional patterns may lie in the perceptual saliency. Previous studies on the perception of Polish sibilants have shown that /ɕ/-/ʂ/ pairs are the most difficult to discriminate. This conclusion is based, among other things, on the perception of Polish sibilants by English (Lisker 2001) and French speakers (Shoemaker 2014) speakers. These findings would imply that /s/ is the most salient sound. However, the results of our and previous studies (e.g. Sołtys-Chmielowicz 1998, Bryndal 2015) have shown that /s/ is not the earliest sound acquired by children. Hence, the perceptual saliency does not seem to be the primary reason for the order of the sibilant acquisition in Polish. Hua and Dodd (2000) have suggested that explaining cross-language acquisitional patterns requires attention to how sounds play out within specific language systems. More cross-linguistic data are needed to expand our understanding of language-general and language-specific patterns in three-way sibilant contrasts.

4.5 Limitations and future work

Whereas previous studies of sibilant acquisition of Polish have been carried out (Kaczmarek, 1953; Łobacz, 1996; Sołtys-Chmielowicz, 1998; Łukaszewicz 2000; Bryndal 2015), the current study contributed detailed acoustic measures, acoustic-perceptual correspondences, and comparisons between perception of adult- and child-produced forms. To keep the task manageable for our youngest listeners, the sibilant set was limited to the voiceless case and single word sets.

14 However, the higher frequency cannot be the only reason for this association, as the [s]-sounds are characterized by even a higher frequency than palatal ones. A deeper analysis of this acoustic-perceptual-semantic relation is needed to better understand this phenomenon. Note also that the “smallness” or “nearness” is not only associated with palatal consonants but also with front vowels (Haynie et al. 2014).
for the two positions (word-initial, word-initial). Future work is needed to understand the acquisitional patterns for voiced sibilants, in syllable coda position, and across differing coarticulatory contexts. Assessing sibilants in multiple vowel contexts would also allow a better understanding of coarticulatory patterns. The inclusion of other vowel contexts such as /i/ and /u/ would affect the spectra due to palatalization and rounding, respectively; thus, a future study could examine these effects with regard to sibilant acquisition. It has been shown for adults that vowel formants are important for the perception of sibilants, mainly for the alveolo-palatal sound but also for retroflexes. Nowak (2006) found increased perceptual identification mistakes when a following low vowel context did not correspond to the correct place of articulation (with less of an effect for higher vowels /e u/). Both retroflexes and alveolo-palatals were very susceptible to the removal of vowel information, whereas dentals were robustly identified based on the fricative noise alone. On the other hand, Lisker (2001), along with showing that retroflex and alveolo-palatal fricative noises were highly confusable, demonstrated that isolated vowels (cut out from sibilant contexts) achieved best sibilant identification for alveolo-palatals, but the other two sibilants were still robustly identified based on vowel information alone.

Finally, we note that the age of the children taking part in this experiment was skewed: most children (57) were between 2;10–5;5 months and only 19 between 5;6–7;11. This gives us more insight into the behavior of younger children and less so into the behavior of older children. It is also the case that our oldest age group had rather few participants, so more studies on long-term aspects of fricative acquisition should also be conducted. However, recall that our variable Age was submitted as a continuous variable for statistical modeling which mitigated this problem to some extent.

5. Conclusions

The data presented here show that Polish-speaking children 3–8 years of age had higher labeling accuracy for an adult voice as compared to their own voice. Acoustic analyses of the sibilants indicate that F2 and the centre of gravity are the primary acoustic features that differentiate the sibilants in production. Analyses relating the acoustics of the children's productions with their labeling suggest that the most important cues differ across fricatives and change as a function of age. Together, the acoustic and perceptual analyses suggest that typical children acquire Polish /ɕ/ first, and at least some children show greater difficulty with /ʂ/ compared to /s/. The early acquisition of /ɕ/ is consistent with acquisitional studies of Mainland Chinese, and may reflect motoric aspects of the sound, and/or the fact that acoustically this sound differs from the other sibilants largely in formant characteristics of adjacent sounds. These results add detail to our understanding of sibilant acquisition within and across languages.

Acknowledgements

We thank the Associate Editor and three anonymous reviewers for comments and suggestions that improved this work. Marzena Żygiis was supported by the BMBF (grant number 01UG1411) and the Leibniz Society. Daniel Pape’s work was partly funded by the Natural Sciences and Engineering Research Council of Canada (NSERC) Early Career Researcher grant (RGPIN-2018-06518, 2018). We also thank Ania Gogolin for drawing the pictures for the experiment, and all our informants for taking part in this study.
References


Appendix

1. Word frequencies

Table A1: Frequency of words based on CHILDES Polish Child-Directed Speech Corpus (CDS) including 799,439 word tokens from seven corpora (128 children) ([https://childes.talkbank.org/access/Slavic/Polish/Polish-CDS.html](https://childes.talkbank.org/access/Slavic/Polish/Polish-CDS.html)). The second column shows morphological forms of the same paradigm in singular form.

<table>
<thead>
<tr>
<th>Word</th>
<th>Frequency of morphological forms in singular</th>
<th>Total frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>kasa /kasa/</td>
<td>kasa 14, kasy 4, kasę 5, kasę 1</td>
<td>27</td>
</tr>
<tr>
<td>Kasia /kāca/</td>
<td>Kasia 148, Kasię 3, Kasi 33, Kasi 11, Kasiu 77</td>
<td>272</td>
</tr>
<tr>
<td>kasza /kaʃa/</td>
<td>kasza 24, kaszy 5, kaszę 8</td>
<td>37</td>
</tr>
<tr>
<td>sanki /sanki/</td>
<td>sanki 56, sanek 1, sankami 7, sankach 40</td>
<td>104</td>
</tr>
<tr>
<td>siatka /ɕatka/</td>
<td>siatka 11, siatki 6, siatce 7, siatkę 7, siatką 1</td>
<td>32</td>
</tr>
<tr>
<td>szafa /ʂafa/</td>
<td>szafa 17, szafy 46, szafie 45, szafę 31, szafą 6</td>
<td>145</td>
</tr>
<tr>
<td>ryba /riba/</td>
<td>ryba 85, ryby 91, rybę 19, rybą 1, rybo 10, rybie 3</td>
<td>209</td>
</tr>
<tr>
<td>słonko /swɔnkɔ/</td>
<td>słonko 58, słonku 4, słonkiem 1</td>
<td>63</td>
</tr>
<tr>
<td>żaba /ʐaba/</td>
<td>żaba 49, żaby 35, żabie 2, żabę 11, żabą 1, żabo 1</td>
<td>99</td>
</tr>
</tbody>
</table>

Table A2: Frequency of words based on National Corpus of Polish with 300 million words ([http://nkjp.pl/poliqarp/](http://nkjp.pl/poliqarp/)). The second column shows morphological forms of the same paradigm in singular form.

<table>
<thead>
<tr>
<th>Word</th>
<th>Frequency of morphological forms in singular</th>
<th>Total frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>kasa /kasa/</td>
<td>kasa 1000, kasy 90, kasie 63, kasę 51, kasę 68, kasę 2</td>
<td>1274</td>
</tr>
<tr>
<td>Kasia /kāca/</td>
<td>Kasia 1000, Kasię 244, Kasię 196, Kasię 80, Kasiu 166</td>
<td>1686</td>
</tr>
<tr>
<td>kasza /kaʃa/</td>
<td>kasza 361, kaszy 41, kaszę 35, kaszę 54, kaszo 1</td>
<td>492</td>
</tr>
<tr>
<td>sanki /sanki/</td>
<td>sanki 122, sanek 33, sankom 2, sankami 61, sankach 40</td>
<td>258</td>
</tr>
<tr>
<td>siatka /ɕatka/</td>
<td>siatka 50, siatki 66, siatce 131, siatkę 38, siatką 119, siatko 1</td>
<td>405</td>
</tr>
<tr>
<td>szafa /ʂafa/</td>
<td>szafa 49, szafy 71, szafie 44, szafę 132, szafą 60, szafę 2</td>
<td>358</td>
</tr>
<tr>
<td>Word</td>
<td>Examples</td>
<td>Frequency</td>
</tr>
<tr>
<td>------</td>
<td>----------</td>
<td>-----------</td>
</tr>
<tr>
<td>ryba /řiba/ “fish”</td>
<td>ryba 32, ryby 120, rybie 47, rybę 236, rybą 95, rybo 10</td>
<td>540</td>
</tr>
<tr>
<td>słonko /swɔŋkɔ/ “sun”</td>
<td>słonko 40, słonka 30, słonku 26, słonkiem 17</td>
<td>113</td>
</tr>
<tr>
<td>żaba /ʐaba/ ‘frog’</td>
<td>żaba 63, żaby 80, żabie 50, żabę 72, żabą 42</td>
<td>307</td>
</tr>
</tbody>
</table>
2. Additional information about the experimental design

2.1. Matrix used in the perception experiment

\[
\begin{array}{cccc}
[1, 1, 2, 3, "L"], \\
[1, 2, 1, 3, "M"], \\
[1, 2, 3, 1, "R"], \\
[2, 1, 2, 3, "M"], \\
[2, 1, 3, 2, "R"], \\
[3, 1, 2, 3, "M"], \\
[3, 1, 2, 3, "R"], \\
[4, 2, 1, 3, "L"], \\
[4, 2, 1, 3, "M"], \\
[4, 2, 1, 3, "R"], \\
[5, 2, 1, 3, "M"], \\
[5, 2, 1, 3, "R"], \\
[6, 2, 2, 3, "L"], \\
[6, 2, 2, 3, "M"], \\
[6, 2, 2, 3, "R"], \\
[7, 2, 2, 3, "M"], \\
[7, 2, 2, 3, "R"], \\
[8, 2, 2, 3, "M"], \\
[8, 2, 2, 3, "R"];
\end{array}
\]

First column: numbers stand for individual pictures: 1 = Kasia ("Cathe"), 2 = kasza ("cash register"), 3 = kasa ("wardrobe"), 4 = sanki ("sleigh"), 5 = szafa ("groats"), 6 = siatka ("net"), 7 = żaba ("frog"), 8 = słońce ("sun")

Second column: first (left) position of the picture on the computer display

Third column: second (middle) position of the picture on the computer display

Fourth column: third (right) position of the picture on the computer display

Fifth column: the correct position (expected answer) of the picture related to a stimulus; L=left, M=mid, R=right;

The rows were randomized for each listener by means of the algorithm fisher Yates in Python

```javascript
var multArr = [0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23];
var order = fisherYates(multArr)
```

2.2. Acoustic spectra for all child productions in initial position
Fig. A1: Individual multitaper spectra (green lines) with overlaid mean spectrum (black lines) obtained for all /s/ productions (top), for all /ɕ/ productions (middle), and for all /ʂ/ productions (right) in word-initial position.

2.3. Acoustic spectra and parameters for the adult production in initial position
Fig. A2: Multitaper spectra measured at the acoustic midpoint for the adult sibilant stimuli in the initial position. The left panel shows the /s/ spectrum, the middle panel the /ɕ/ and the right panel the /ʂ/ spectrum.

Table A3: Parameters of the acoustic analysis of the adult sibilant productions used in the perception experiment (initial position)

<table>
<thead>
<tr>
<th>Parameter</th>
<th>/sanki/</th>
<th>/ɕatka/</th>
<th>/ʂafa/</th>
</tr>
</thead>
<tbody>
<tr>
<td>duration [ms]</td>
<td>168</td>
<td>146</td>
<td>146</td>
</tr>
<tr>
<td>spectral peak [Hz]</td>
<td>9819</td>
<td>4307</td>
<td>2584</td>
</tr>
<tr>
<td>centre of gravity [Hz]</td>
<td>8658</td>
<td>4732</td>
<td>2627</td>
</tr>
<tr>
<td>SD</td>
<td>1828</td>
<td>1663</td>
<td>940</td>
</tr>
<tr>
<td>skewness</td>
<td>0.6199</td>
<td>1.8773</td>
<td>3.5991</td>
</tr>
<tr>
<td>kurtosis</td>
<td>1.7595</td>
<td>10.6168</td>
<td>21.7851</td>
</tr>
<tr>
<td>formant values (F1, F2, F3) for the preceding vowel [Hz]</td>
<td>772, 1563, 2849</td>
<td>920, 1893, 2717</td>
<td>821, 1563,2667</td>
</tr>
<tr>
<td>formant values (F1, F2, F3) for the following vowel [Hz]</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Fig. A3: F2 of the vowel following /s, ɕ, ʂ/ in the word-initial position as a function of age

Fig. A4: COG of /s, ɕ, ʂ/ in the word-initial position as a function of age

3. Results of a preliminary perceptual study with two listeners
Table A4: Results of the perceptual test. The inter-rater agreement calculated as Cohen’s kappa was 0.71.

<table>
<thead>
<tr>
<th></th>
<th>/s/</th>
<th>/ʂ/</th>
<th>/ɕ/</th>
</tr>
</thead>
<tbody>
<tr>
<td>Annotator 1:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>correct</td>
<td>163 (67.6%)</td>
<td>102 (42.5%)</td>
<td>204 (85%)</td>
</tr>
<tr>
<td>Annotator 2:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>correct</td>
<td>185 (76.8%)</td>
<td>123 (51.3%)</td>
<td>186 (77.5%)</td>
</tr>
</tbody>
</table>

The following tables show the annotations by both adult listeners according to age categories.

Table A5: Data for /s/

<table>
<thead>
<tr>
<th></th>
<th>2;11-3;9</th>
<th>3;10-4;7</th>
<th>4;8-5;5</th>
<th>5;6-6;3</th>
<th>6;4-7;11</th>
</tr>
</thead>
<tbody>
<tr>
<td>Annotator 1:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>correct</td>
<td>25 (40.3%)</td>
<td>32 (59.3%)</td>
<td>52 (86.7%)</td>
<td>31 (88.6%)</td>
<td>22 (81.5%)</td>
</tr>
<tr>
<td>Annotator 2:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>correct</td>
<td>40 (64.5%)</td>
<td>36 (66.7%)</td>
<td>52 (86.7%)</td>
<td>31 (88.6%)</td>
<td>25 (92.6%)</td>
</tr>
</tbody>
</table>

Table A6: Data for /ʂ/

<table>
<thead>
<tr>
<th></th>
<th>2;11-3;9</th>
<th>3;10-4;7</th>
<th>4;8-5;5</th>
<th>5;6-6;3</th>
<th>6;4-7;11</th>
</tr>
</thead>
<tbody>
<tr>
<td>Annotator 1:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>correct</td>
<td>9 (14.5%)</td>
<td>16 (29.6%)</td>
<td>35 (58.3%)</td>
<td>19 (55.9%)</td>
<td>21 (77.8%)</td>
</tr>
<tr>
<td>Annotator 2:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------------</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
</tr>
<tr>
<td>correct</td>
<td>19</td>
<td>19</td>
<td>40</td>
<td>23</td>
<td>21</td>
</tr>
<tr>
<td></td>
<td>(30.6%)</td>
<td>(35.2%)</td>
<td>(66.7%)</td>
<td>(67.6%)</td>
<td>(77.8%)</td>
</tr>
</tbody>
</table>

Table A7: Data for /s/

| 2;11-3;9     | 3;10-4;7 | 4;8-5;5 | 5;6-6;3 | 6;4-7;11 |

<table>
<thead>
<tr>
<th>Annotator 1:</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>correct</td>
<td>44</td>
<td>46</td>
<td>51</td>
<td>33</td>
<td>27</td>
</tr>
<tr>
<td></td>
<td>(71.0%)</td>
<td>(85.2%)</td>
<td>(85.0%)</td>
<td>(97.1%)</td>
<td>(100%)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Annotator 2:</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>correct</td>
<td>37</td>
<td>39</td>
<td>50</td>
<td>32</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>(59.7.6%)</td>
<td>(72.2%)</td>
<td>(83.3%)</td>
<td>(94.1.6%)</td>
<td>(92.6%)</td>
</tr>
</tbody>
</table>

For comparison, Holliday et al. (2015) analyzed percentages of correct production for English /s/ and /ʃ/ in children ages 2–5. At about 36 months of age, logistic regression model predicted accuracy of about 45% for /s/ and 55% for /ʃ/. 