Acoustic and perceptual evaluation of category goodness of /t/ and /k/ in typical and misarticulated children’s speech

Sofia Strömbergsson, Giampiero Salvi, and David House
Department of Speech, Music and Hearing, School of Computer Science and Communication, KTH Royal Institute of Technology, Lindstedtsvägen 24, SE-100 44 Stockholm, Sweden

(Received 20 February 2014; revised 9 April 2015; accepted 24 April 2015)

This investigation explores perceptual and acoustic characteristics of children’s successful and unsuccessful productions of /t/ and /k/, with a specific aim of exploring perceptual sensitivity to phonetic detail, and the extent to which this sensitivity is reflected in the acoustic domain. Recordings were collected from 4- to 8-year-old children with a speech sound disorder (SSD) who misarticulated one of the target plosives, and compared to productions recorded from peers with typical speech development (TD). Perceptual responses were registered with regards to a visual-analog scale, ranging from “clear [t]” to “clear [k].” Statistical models of prototypical productions were built, based on spectral moments and discrete cosine transform features, and used in the scoring of SSD productions. In the perceptual evaluation, “clear substitutions” were rated as less prototypical than correct productions. Moreover, target-appropriate productions of /t/ and /k/ produced by children with SSD were rated as less prototypical than those produced by TD peers. The acoustical modeling could to a large extent discriminate between the gross categories /t/ and /k/, and scored the SSD utterances on a continuous scale that was largely consistent with the category of production. However, none of the methods exhibited the same sensitivity to phonetic detail as the human listeners. © 2015 Acoustical Society of America. [http://dx.doi.org/10.1121/1.4921033]

I. INTRODUCTION

The description and assessment of impaired speech production is most often based on the perceptual evaluations of listeners and expressed with reference to phonetic transcription. Phonetic transcription on a broad level necessarily involves a reduction of acoustic-phonetic detail into phonological categories. Therefore, the use of transcription as an analytic instrument to assess and identify problems in children’s speech production might not always produce a complete or fair description of the child’s actual production. A child who is perceived as not producing a distinction between two sounds might in fact signal the phonological contrast by other means, by producing a sub-phonemic distinction that passes undetected by listeners (Gibbon, 1999; Hewlett, 1988; Macken and Barton, 1980; Munson et al., 2010). Evidence of such subtle acoustic-phonetic distinctions—“covert contrast”—has been reported for many contrasts, e.g., for place of articulation of stops (Forrest et al., 1990; Hewlett, 1988; Scobbie et al., 2000). Differentiation between phonemes in speech production—even if it is signaled sub-phonemically, by other means than by the phonologically critical features—is a positive prognostic factor for children with speech sound disorders (SSDs) (Tyler et al., 1993). Hence, looking at such signals in the clinical intervention of SSDs would be very unfortunate and can have negative implications for intervention and therapy. Moreover, without an objective and fine-grained instrument of evaluating a child’s efforts at producing a speech target through the course of intervention, subtle signs of progression might not be registered. In this study, we explore adults’ perceptual sensitivity to phonetic detail and covert contrasts in the evaluation of children’s efforts at producing the voiceless stops /t/ and /k/ by using a visual-analog rating scale instead of traditional transcription. We then describe the initial steps towards the long-term goal of achieving an objective acoustic analysis of children’s speech by evaluating how two acoustic analysis methods (spectral moments analysis and discrete cosine transformation analysis) compare to the results of human evaluators.

A. Perceptual evaluation: Possibilities and limitations

In speech perception, the phenomenon of categorical perception is well-documented, e.g., by Liberman et al. (1957); when listeners are given a binary choice between two speech sounds (phonemes), and faced with the task of judging whether ambiguous speech stimuli are realizations of either one or the other, listeners with the same language background are in high agreement on where the phonemic boundary between these two sounds lies. Furthermore, the listeners will disregard acoustic variation on either side of phonemic boundaries (i.e., within-category variation), but be sensitive to variation occurring at phonemic boundaries (i.e., across-category variation). However, when the perceptual task involves a gradient judgment rather than a binary choice, e.g., to evaluate category goodness with regards to a visual-analog scale (VAS) between two sound sounds, listeners have been shown to utilize the whole scale, thus exhibiting sensitivity to finer acoustic-phonetic detail—far from and close to categorical boundaries (Munson et al., 2010).
similar responses, i.e., responses corresponding to human listeners’ rating of category goodness, could be generated automatically it would be of great clinical and educational value. If implemented in computer-based speech training tools—for children with disordered speech, or for second language learners—an objective acoustic measure of category goodness could provide valuable feedback to the users in their attempts at approaching adequate speech targets.

Even if the perceptual evaluation of speech sounds does not involve phonetic transcription or other categorical decisions, other factors may also obscure physically measurable detail in the speech signal. One such factor is the potential perceptual bias in the listener, e.g., driven by lexical or phonotactic expectations, or by other aspects of the listener’s linguistic experience. For instance, the context in which the speech sound is presented has been shown to affect listeners’ perception of the sound; if phonologically ambiguous speech sounds are presented in word and nonword contexts, listeners tend to prefer phonological interpretations that make words. For example, in synthesized continua (varying in voice onset time) between the word “dask” and the nonword “task” (Ganong, 1980). The phonological context in which a speech sound occurs may also influence listeners’ perceptual evaluations, such that more frequent phonotactic interpretations are preferred over those that are less frequent. Hence, listeners evaluating a phonologically ambiguous CV sequence between /ri/ and /li/ will prefer the /ri/-interpretation when the CV sequence follows a /t/ (hence showing a preference for the more frequently occurring sequence /tri/ over the phonotactically less probable sequence /tli/), whereas they will prefer the /li/-interpretation when the CV sequence follows an /s/ (hence preferring the admissible sequence /sli/ over the less probable sequence /stri/). Such phonotactic probability effects have been reported by, for example, Massaro and Cohen (1983) and Pitt and McQueen (1998).

The listeners’ own levels of experience perceiving disordered speech may also affect their perceptual judgments. For example, whereas inexperienced listeners are biased towards interpreting /t/-/k/ productions as the more frequently occurring /t/, experienced listeners do not exhibit the same preference (Munson et al., 2012). Taken together, these reports highlight the need to compensate for the potential influence of listener biases when interpreting perceptual responses of human listeners.

**B. Acoustic analysis of disordered speech**

Early research on the acoustic differentiation between velar and dental/alveolar stops (i.e., the distinction between /k/ and /t/, and between /g/ and /d/) in adult speakers identified phonologically critical cues to place of articulation in the transition from the plosive into the following vowel (Fant, 1960; Liberman et al., 1954). Subsequent work has applied locus equations (Lindblom, 1963), relying on trajectory properties of the second formant (F2) in this region, in classifying place of articulation in plosives (Krull, 1990; Lindblom, 1963; Sussman et al., 1991). Although these approaches have been quite successful for distinguishing place of articulation in voiced plosives, distinguishing the same place contrast in unvoiced plosives requires other methods. For the distinction between voiceless /t/ and /k/ in languages such as Swedish, where aspiration and absence of a voicing limit the possibility of tracking formants in the transition into the vowel, differentiating acoustic cues have been sought in the release burst of the plosive (Blumstein and Stevens, 1979; Forrest et al., 1990; Forrest et al., 1988; Marin et al., 2010). For example, in their study of children’s successful and unsuccessful efforts at producing /k/, Forrest et al. (1990) applied a spectral moments analysis to capture dynamic aspects in the interval from the burst release to the onset of the vowel. In their analysis, three spectral parameters—mean frequency, skewness, and kurtosis—were used to successfully (with an accuracy of 82%) classify initial voiceless plosives by their place of articulation, in utterances produced by children with typical speech production. However, when the same model was applied to the speech produced by the children with disordered speech, who were all perceived as substituting [t] for /k/, they identified contrasting production only one of four children. Interestingly, this child was the only one of the four who produced appropriate production of /k/ after treatment, supporting the proposal in Tyler et al. (1993): the production of a covert contrast between two speech sounds is a positive prognostic factor.

Spectral moments analysis has been used by many researchers in the acoustic differentiation between /t/ and /k/ in children’s speech (Forrest et al., 1994; Nissen and Fox, 2009; Nittouer, 1995; Tyler et al., 1993), and although most rely only on the first, third, and fourth moments (corresponding to spectral mean, skewness, and kurtosis), Nissen and Fox (2009) report that also the second moment (corresponding to spectral variance) contributed significantly in the differentiation between different places of articulation.

Although hitherto not reported as frequently, an alternative approach to the acoustic analysis of the /t/-/k/ contrast is described by Marin et al. (2010), namely, Bark-scaled discrete cosine transformation (DCT) analysis. Similar to the spectral moments analysis used by Forrest and colleagues, the DCT analysis also relies on time-varying spectral information in the burst of the plosive. However, instead of representing spectral information linearly, Marin and colleagues follow the suggestion by Kewley-Port (1983), and use a nonlinear approach—the Bark scale—in order to better reflect the frequency resolution in the human auditory system. The use of the Bark scale is particularly motivated when aiming to predict and simulate perceptual performance by human listeners.

Considering the difficulties in human listeners to disregard phonemic perception and instead focus on subphonemic detail, an evaluative measure based solely on the information in the acoustic signal would respond to the demand for objective assessment methods in clinical and educational settings. If implemented successfully, this could function as an objective judge registering finer-grained detail, e.g., the acoustic-phonetic differentiation in a child’s speech production between /t/ and /k/, signaled by other
means than by those that are phonologically critical to proficient language users. As fronting of velar consonants, i.e., the perceived substitution of alveolar consonants [t, d, n] for the velar targets /k, g, n/ is one of the most frequently reported error patterns in children with SSDs, occurring across many languages (Nijland, 2009; Rvachew and Brosseau-Lapré, 2012), a large group of children would potentially benefit from more basic and clinical research into this particular speech error pattern. And considering that most research on the acoustic distinction between /k/ and /t/ is based on American English, there is also value in exploring this variation in other languages. In Swedish, for example, /t/ is produced with a dental rather than an alveolar place of articulation (Stoel-Gammon et al., 1994). Hence, it is not evident that acoustic and perceptual findings reported for American English are directly transferrable to Swedish.

C. Speech sound disorders and motor control

As referred to above, descriptions of impaired speech production are most often based on the perceptual evaluations of adult listeners, and expressed with reference to phonetic transcription. Furthermore, acoustic studies of impaired speech production have often—quite understandably—been restricted to describing the nature of the specific error patterns, while paying less attention to the children’s general articulatory skills. However, reports have been presented indicating that children with SSDs may have more general problems with speech motor control compared to their peers, even affecting aspects of speech that are unrelated to their documented error patterns. For example, Edwards (1992) found that some children with SSD with documented deficits in their production of consonants, exhibited impaired compensatory articulation of vowels in a bite-block experiment, compared to peers with typical speech and language development. Moreover, reduced diadochokinetic rates have also been observed for children with SSDs, even in children whose main problem is not considered to be an impairment of speech-motor control (Shriberg et al., 1986). Measurable acoustic differences have also been reported between seemingly target-appropriate productions of /t/ and /k/ in children with SSDs who are in the process of acquiring this contrast, compared to target-appropriate productions produced by peers with typical speech and language development (Forrest et al., 1994). Together, these reports illustrate that detailed acoustic and physiological measures of more general speech production skills may reveal subtle deficiencies in speech produced by children with SSDs, which may not be detected by standard assessment methods. Furthermore, the reports suggest the potential existence of a more general speech-motor deficit in (at least some) children with SSD. To date, however, studies that include fine-grained perceptual measures of speech produced by children with SSDs have been rare, and, hence, little is known of whether the reported subtle physiological and acoustic idiosyncrasies are perceptually discernible. The present study also aims to address this gap by examining phonetic detail in successful and unsuccessful efforts at producing /t/ and /k/ in children with SSD, in both the perceptual and the acoustic domain.

D. Research questions

In this investigation, the following questions are explored:

1. Is there perceptual evidence of covert contrast in children’s misarticulated productions classified as “clear substitutions” of one speech sound for another?
2. Are there any perceptual differences between sounds produced by children with a SSD compared to their typically developing (TD) peers, even within target-appropriate productions?
3. How well do acoustic models of prototypical “correct” productions predict human perceptual evaluations of category-goodness?

II. METHOD

A. Materials

Recordings were collected from 117 Swedish children ranging from 4 to 9 years of age. All of the children spoke Swedish as their mother tongue. Thirty-five of the children had been diagnosed as having a speech sound disorder (specifically, a phonological disorder), and produced speech with either velar fronting or dental backing (substituting [t] for /k/, or vice versa). These 35 children were recruited through speech-language pathologists in Stockholm; 5 of them to participate in a pilot intervention study, and 30 of them for participation in studies investigating children’s perception of their own speech production (Strömbergsson, 2013; Strömbergsson et al., 2014). For the same purpose, children with no known problems with speech and/or language were recruited from pre-schools and schools in the same region. In these studies, the children with typical speech production (TD) had been recruited in two different age groups, TD_young (n = 50, age range: 4:0 to 6:0, mean age = 5:1, SD = 7.0 months) and TD_old (n = 32, age range = 6:3 to 9:2, mean age = 7:9, SD = 8.6 months). The age of the children with SSD varied between 4:0 and 7:8, with a mean of 5:2 (SD = 10.8 months). For more descriptive details of these children, see the studies referred to above. The recorded items were isolated words beginning with either /tV/ or /kV/, with primary stress on the first syllable (see the Appendix). The words were elicited by playing a pre-recorded adult voice producing the stimuli, and asking the children to produce the same word in isolation. In total, 4002 isolated words beginning with either /t/ (n = 1518) or /k/ (n = 2484) were recorded. (Of these, 886 recordings with initial /k/ were collected from an unpublished intervention study, where initial /k/ was the target of intervention. These words were the same as listed in the Appendix.) All recordings were made at the children’s schools or pre-schools, in separate rooms with limited noise, with a 16-kHz sampling rate and 16-bit resolution, in a lossless format. Two different headset microphones were used during the course of collecting the recordings: a Sennheiser PC151 headset and a Sennheiser m@b 40 headset.

All recorded items were segmented and aligned with the HMM-based nAlign (Sjölander, 2003). Through this
procedure, the beginning and end of the word, as well as segmental boundaries, were identified. In order to ensure consistency of presentation, all initial and final silences (i.e., before and after the recorded words) were set to 1 s of silence. Then, all 4002 items were evaluated by the first author (a Swedish-speaking certified speech-language pathologist), with regards to whether the production of the initial consonant was successful or not. Thus, initial consonants were classified as belonging to one of the six categories listed in Table I (these are the same categories as used in Munson et al., 2010). Items not fitting into any of these categories (e.g., productions where the initial consonant was masked by noise, or otherwise inaudible) were labeled as “distortion.” This also pertained to items where the alignment procedure described above had been unsuccessful—as noted when the item included more or less speech than the expected word. All items labeled as “distortion” were excluded from further analysis, leaving a set of productions containing 3627 items. This set included productions from all 177 children, the majority of the children (n=64) contributing 23 items each. The number of items per child varied from 12 to 147, with 5 children with SSD (the ones included in the intervention study) contributing more than 100 items each (recorded across multiple sessions). As can be noted in Table I, velar fronting (i.e., the substitution of [t] for /k/) is considerably more common in the data than the opposite pattern, as reflected in the category Ts being eight times as large as the Ks category. The 5 children included from the intervention study contribute appreciably to this skewness; however, the asymmetry is also observed in the remaining 30 children with SSD. This reflects the distribution of these error patterns in the population of children with SSDs, where dental backing (i.e., the substitution of [k, g, η] for /t, d, n/) is considerably less frequent than velar fronting (Dodd, 2005; Hodson and Paden, 1983; Nijland, 2009).

Intra-annotator reliability for this classification, as calculated by means of Cohen’s kappa for a 5% subset of the data (180 items), was 0.87 (95% confidence interval, CI: 0.81–0.93). A second annotator [a fourth-year speech-language pathology (SLP) student] was engaged to annotate another 5% of the data, and the inter-annotator agreement between these annotations and those of the first author, as calculated by means of Cohen’s kappa, was 0.90 (95% CI: 0.84–0.95).

<table>
<thead>
<tr>
<th>Category</th>
<th>Description</th>
<th>N (TD)</th>
<th>N (SSD)</th>
<th>N (all)</th>
</tr>
</thead>
<tbody>
<tr>
<td>K</td>
<td>Correct production: [k] for /k/</td>
<td>1024</td>
<td>436</td>
<td>1460</td>
</tr>
<tr>
<td>T</td>
<td>Correct production: [t] for /t/</td>
<td>973</td>
<td>410</td>
<td>1383</td>
</tr>
<tr>
<td>Ks</td>
<td>Clear substitution: [k] for /t/</td>
<td>1</td>
<td>53</td>
<td>54</td>
</tr>
<tr>
<td>Ts</td>
<td>Clear substitution: [t] for /k/</td>
<td>8</td>
<td>424</td>
<td>432</td>
</tr>
<tr>
<td>Km</td>
<td>Between /t/ and /k/, but more like [k]</td>
<td>13</td>
<td>77</td>
<td>90</td>
</tr>
<tr>
<td>Tm</td>
<td>Between /t/ and /k/, but more like [t]</td>
<td>15</td>
<td>193</td>
<td>208</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td>2034</td>
<td>1593</td>
<td>3627</td>
</tr>
</tbody>
</table>

From the alignment procedure described above, the start and end points of all segments (including the occlusion and burst of the plosive) were identified in all 3627 recorded items. The initial CV-sequence was then automatically extracted from the productions, in order to allow presentation of speech isolated from its lexical context. A balanced set of items (CV sequences) was extracted automatically, by randomly selecting 20 items from each of the six categories listed in Table I, resulting in a listening script of 120 items. Ten of these stimuli were randomly selected and doubled in the script, in order to allow for calculation of intra-annotator reliability. Five items (not occurring in the listening script) representing both clear and intermediate productions, were selected as practice items. Hence, the listening script presented to the listeners in the perceptual evaluation consisted of 5 introductory items (excluded from analysis), and 130 test items.

B. Perceptual evaluation

Ten native Swedish listeners participated in the perceptual evaluation. The listeners were all SLP students, with beginner experience of transcribing and evaluating misarticulated speech. The stimuli were presented in random order, different for all participants. The task for the listeners was to evaluate the stimuli with regards to a seemingly step-less scale (a visual-analog scale, VAS) from “t” to “k” to the right. For each stimulus item, the listeners provided their evaluation by moving the marker to its desired location on the VAS. Although invisible to the participants, the extreme left end of the VAS was represented in the graphical interface as 0, and the extreme right end of the VAS was represented as 100, and each position in between was represented as a numerical value between these numbers. Hence, 0 would correspond to a “clear [t]” and 100 would correspond to a “clear [k].” After their evaluation of the 5 practice items, the listeners were given a chance to ask for clarification of the instructions before commencing their evaluation of the 130 test stimuli.

The first author evaluated the total set of 3627 productions following the same procedure as the 10 recruited listeners. A randomly extracted 3% subset of these productions (120 items) was then re-annotated to enable calculation of intra-rater reliability. Intra-rater reliability was calculated by means of an intra-class correlation (ICC), showing high reliability: ICC(2,2) = 0.98 (95% CI: 0.97–0.98). Intra-rater reliability was also calculated for each of the participating listeners, revealing high reliability figures overall, except for one of the participants (listener 7 in Table II). This participant was therefore excluded from further analyses. The agreement within the listener group, i.e., among the remaining 9 listeners, was also high: ICC(2,9) = 0.97 (95% CI: 0.96–0.97). In order to estimate the agreement between the listener group and the single annotator, the responses of the listener group were averaged and then compared to the single annotator’s response. The inter-rater agreement between the group average (with listener 7 excluded) and the single annotator was then ICC(2,2) = 0.93 (95% CI: 0.90–0.95), p < 0.001.
were estimated by means of the following formula:

\[
p(C|V) = \frac{p(C \cap V)}{p(V)},
\]

where \(p(V)\) and \(p(C \cap V)\) were both estimated within the set of all words in the newspaper corpus beginning with /kV/ or /tV/. Table III presents a list of transitional probabilities for all CV sequences among the recorded items.

As indicated in the table, /t/ would be considerably more expected than a /k/ before an /a:/, whereas the expectation would be driven in the opposite direction before an /o/.

Another potential perceptual bias was also explored, namely, whether listeners exhibit a lexical preference for interpretations that make words. This analysis would reveal a bias in the VAS evaluations towards, for example, interpreting intermediate /t/-/k/ productions as /k/ in the CV-syllable /Cu:/, since /ku:/ is a word in Swedish (“ko,” English: cow), whereas /tu:/ is not. Note, however, that whereas /tu:/ does not form a word alone, it is not a rare CV-sequence in Swedish. The reason why it is not presented in Table III is that it never occurs in the recorded data. This motivates an exploration of a potential word-forming bias that is independent from potential biases driven by phonotactic frequency. In order to explore the potential existence of a word-forming bias, three different vowel contexts were contrasted.

(1) BothWords (n = 18): In cases where the plosive is followed by /o:/, the /t/-/k/-decision is expected not to be biased in any direction, since both /ko:/ and /to:/ make words in Swedish (“k” and “tå,” respectively). English: the letter k and toe, respectively). This context serves as a baseline.

(2) T_Word (n = 71): In cases where the plosive is followed by /a:/, a lexical preference would predict a bias in the /t/-/k/-decision towards /t/, since /tu:/ makes a word in Swedish (“ta”; English: take), whereas /ku:/ does not.

(3) K_Word (n = 11): In cases where the plosive is followed by /u:/, a lexical preference would predict a bias in the /t/-/k/-decision towards /k/, since /ku:/ makes a word in Swedish (“ko”; English: cow), whereas /tu:/ does not.

The numbers following the category label indicate the number of recorded items per category. As the vowels /o:/, /a:/ and /u:/ are the only ones that form words if preceded by /k/ or /t/, this is a subset of all items evaluated as Km or Tm.) In the case of a word-forming bias, the T_Word cases would be expected to fall closer towards /t/ (numerically represented as 0) than the BothWord (baseline) cases, and the K_Word cases would be expected to fall closer towards /k/ (numerically represented as 100) than the BothWord cases.

### C. Exploration of perceptual biases

In order to allow the exploration of perceptual biases driven by phonotactic expectations, a Swedish newspaper corpus of around 14 \(\times\) 10^6 word tokens, each associated with a phonological transcription, was used. Based on this corpus (provided by the Swedish Agency for Accessible Media), word frequencies and phonotactic frequencies were calculated. A bias driven by phonotactic (or transitional) probabilities would, for example, drive the VAS evaluations closer to the /t/ end of the scale in cases where the initial plosive is followed by /i:/, as /ti:/ is a much more frequent initial CV-sequence than /ki:/ in Swedish. Transitional probabilities were estimated by means of the following formula:

\[
p(C|V) = \frac{p(C \cap V)}{p(V)},
\]

### D. Acoustic evaluation

The goal of the acoustic analysis was to characterize the acoustic properties of the word-initial voiceless plosives. In order to do this, two methods for feature extraction and two modeling methods were used. In all cases, the analyzed segment stretched across the entire burst-to-vowel interval, the endpoints of this interval having been identified in the automatic alignment procedure described above. For the first feature extraction method a DCT analysis was used, following the same procedure as described in Marin et al. (2010). From this analysis, three time-varying spectral coefficients...
are obtained, representing the spectrum’s mean, slope and curvature, respectively. (For more details, see Marin et al., 2010.) The second feature extraction method was based on spectral moments (SM). The SM analysis followed the procedure described in Forrest et al. (1988), with the modification that all four first moments were included in the analysis. These moments represent spectral mean, variance, skewness and kurtosis.

For both analysis procedures, the first step was to compute a discrete Fourier transform (DFT) over windows spaced at constant time intervals. In the SM analysis, the first four spectral moments were evaluated on the DFT, whereas in the DCT analysis, the DFT was first converted to the Bark scale, and the DCT was then applied to the resulting Bark spectrum. Three DCT coefficients were considered in this case. As the segments were of varying length, both analysis procedures resulted in a variable number of feature vectors per segment. To obtain a fixed length representation, a DCT analysis was applied along the time axis to each feature. Again, three DCT coefficients in time were used, resulting in a representation with 12 elements in the case of SM and 9 elements for the DCT.

Modeling the perceptual categories on the basis of the acoustic features was performed in two different ways. In both cases, a statistical model of prototypical utterances of /t/ and /k/ was built, and used as a reference against which untrained utterances were evaluated on a continuous scale between /t/ and /k/. This enabled using the model both for classification into perceptual categories and for prediction of the gradient perceptual evaluations. For building the statistical models, the dataset was divided into a training set and a test set. The training set consisted of error-free tokens of /t/ (973 items) and /k/ (1024 items), i.e., the categories T and K, respectively. All these items were produced by children with typical speech. (These categories appear in the first two rows, in the column representing children with TD, in Table I.) The test set included only tokens by children with SSD; the distribution of these tokens across the six different categories is displayed in Table I. The first modeling method makes use of Gaussian distributions: mean vectors and covariance matrices for /t/ and /k/ utterances were estimated on the training set. Given a new test utterance, a proximity index, \( P \), was defined by Marin et al. (2010) as

\[
P = \log(M_k) - \log(M_t),
\]

where \( M_k \) and \( M_t \) are Mahalanobis distances of the current example from the prototypical distributions for /k/ and /t/, respectively. To ensure consistency in the manner in which the visual analog scale was implemented in the perceptual tests (where \( /t/ = 0 \) and \( /k/ = 100 \)), the opposite of the proximity index (\( \text{PROX} = -P \)) is used in the presentation of results. Hence, when \( \text{PROX} = 0 \), the token is equidistant between /t/ and /k/. Positive PROX-values indicate tokens closer to /k/ than to /t/, whereas negative PROX-values indicate tokens closer to /t/ than to /k/.

The second modeling method was based on support vector machines (SVMs). SVMs perform classification by selecting representative samples at the boundary between the classes. Although the SVM implements a linear classifier, kernels can be used to project the feature space into a high dimensional space where the classes are linearly separable. For this, a radial basis function kernel was used, as in Marin et al. (2010). The predictions of the SVM model are interpreted as the probability that the observation belongs to class /k/. The SVM method was tested on DCT and SM features separately, as well as on the combination of DCT and SM features. Finally, a linear discriminant analysis (LDA) was performed in order to verify the importance of nonlinearities modeled by the SVM. Table IV summarizes the different models used in the analysis.

III. RESULTS

A. Perceptual evaluation

The VAS ratings of the single annotator were defined as a dependent variable in a one-way analysis of variance (ANOVA), where stimulus category was entered as an independent variable. In order to compensate for unequal variances, a Welch variance-weighted ANOVA was conducted. This determined that there were significant differences between the stimulus categories in how they were rated with regards to the visual-analog scale, \( F(5, 485.82) = 31530.38, p < 0.001 \). Figure 1 displays the VAS ratings for each stimulus category, with outliers removed. A Games-Howell post hoc analysis revealed significant differences between all stimulus categories \( (p < 0.01) \). It can be noted that although the 3627 rated items were distributed across the entire range of the VAS, the majority were rated as either 0 or 100 (40% rated as 0, and 38% rated as 100).

An analogous ANOVA was conducted on the perceptual evaluations performed by the listener panel, in order to explore the extent to which their evaluations of the 120 item subset exhibited the same pattern as the evaluations performed by the single annotator on the full dataset. In parallel to the former analysis, this analysis also revealed significant differences between the stimulus categories in how they were rated with regards to the VAS, \( F(5, 492.03) = 725.08, p < 0.001 \). A Games-Howell post hoc analysis revealed significant differences between all stimulus categories \( (p < 0.05) \), except between K and KS \( (p = 0.054) \).

In order to explore the presence of perceptually measurable differences between the children with SSD and their TD

| TABLE IV. Overview of the different methods used in the acoustic analysis, and of the range used to describe the similarity to /k/. |
|-----------------|-----------------|-----------------|-----------------|-----------------|
| Method          | Features        | Model           | Range           |
| PROX(DCT)       | DCT             | Proximity index | [−∞, ∞]         |
| PROX(SM)        | SM              | Proximity index | [−∞, ∞]         |
| SVM(DCT)        | DCT             | Support vector machines | [0, 1] |
| SVM(SM)         | SM              | Support vector machines | [0, 1] |
| LDA(DCT + SM)   | DCT + SM        | Linear discriminant analysis | [0, 1] |
| PROX(DCT + SM)  | N/A             | Support vector machines | [0, 1] |
| SVM(DCT + SM)   | DCT + SM        | Support vector machines | [0, 1] |

*The PROX index could not be modeled with the combined DCT + SM features, because the amount of training data was not sufficient to estimate the covariance matrices for such high dimensional vectors.*
peers among their productions categorized as correct, three speaker groups were defined: SSD, TD_young (up to 6 years old) and TD_old (older than 6 years old). (The splitting of the TD children into two groups was motivated by the fact that these children had been recruited to different age groups in earlier studies.) A one-way (Welch variance-weighted) ANOVA with VAS score as a dependent variable and speaker group (SSD, TD_young, TD_old) as an independent variable showed significant effects of speaker group for the K category: F(2, 853.00) = 10.26, p < 0.001. A Games–Howell post hoc analysis revealed significant differences (p < 0.01) between the SSD-group and the two TD-groups, but not between the two TD-groups. Hence, even sounds categorized as correct /k/’s were evaluated as being less /k/-like when produced by children with SSD, compared to when produced by children with TD. An analogous ANOVA was calculated for items categorized as correct /t/’s (i.e., the T category); this, too, showed a significant dependence on speaker group: F(2, 680.12) = 9.11, p < 0.001. As for the /k/ sounds, a Games-Howell post hoc analysis revealed significant differences (p < 0.001) between the SSD-group and the two TD-groups, but not between the two TD-groups. Hence, even sounds categorized as correct /k/’s, i.e., sounds which most of the children with SSD were not judged as having problems producing, were evaluated as being less /k/-like compared to the /t/’s produced by their TD peers. All ANOVA results reported above were verified by non-parametric Kruskal–Wallis analyses.

B. Perceptual biases

In order to explore the potential existence of a perceptual bias driven by phonotactic (or transitional) probability, a Spearman’s correlation analysis was conducted among the intermediate stimuli. A strong correlation between these factors would indicate a potential influence of phonotactic probability on the VAS-evaluation among these stimuli. However, this analysis showed no such interdependence: \( r_s = -0.10, n = 298, p = 0.09 \).

Regarding the potential existence of a perceptual bias driven by lexical expectations, a one-way ANOVA was conducted, where vowel context (T_WORD, K_WORD and BOTHWORD) was entered as an independent variable. As described earlier, a pattern where the T_WORD cases fall closer towards /t/ (numerically represented as 0) than the BOTHWORD (baseline) cases, and where the K_WORD cases fall closer towards /k/ (numerically represented as 100) than the BOTHWORD cases, would indicate a word-forming bias. However, the analysis revealed no significant differences between these vowel contexts: F(2, 97) = 0.64, p = 0.53, \( \eta^2 = 0.01 \). The expectation of a word-forming bias is further refuted by the visual inspection of the results illustrated in Fig. 2; the fact that the T_WORD average falls closer to /k/ (or 100) than the BOTHWORD average disfavors the suggestion of a word-forming bias. \(^2\)

C. Acoustic analysis

The goal of the acoustic analysis was to estimate the similarity between untrained productions of /t/ and /k/ and given prototypical productions of /t/ and /k/. A second aim was to examine the correlation between the acoustic measures of similarity and the perceptual evaluation described in Secs. III A and III B. In all methods used, the predictor increases with the similarity to /k/. As seen in Table IV, the proximity index has a range from \(-\infty\) to \(+\infty\), whereas the LDA and SVM estimate the probability of /k/, which is bounded between 0 and 1.

The classification results for all the methods are presented in Table V, for the training and the test set, respectively. (As described above, the training set only contains recorded items from the TD group, whereas the test set only contains recorded items from the SSD group.) In order to allow for comparisons, accuracy (both for the training and test set) is computed as the number of correctly classified K and T over the total number of K and T. The table also provides figures of partial accuracy, representing the classification accuracy for K and T separately. For the test set, confusion matrices also show the intermediate categories that are not included in the accuracy calculation. For these categories, a correct assignment to either T or K does not

FIG. 1. Box plot displaying median, quartiles, and range of VAS scores across the different stimulus categories, with outliers removed.

FIG. 2. Average VAS score for each of the three different vowel contexts; K_Word (n = 11) defined as forming a word only in a /k/ context, T_Word (n = 71) defined as forming a word only in a /t/ context, and BOTHWords (n = 18) defined as forming words in both /k/ and /t/ contexts. Error bars represent a 95% confidence interval.
A closer inspection of how classification accuracy varies across the two categories T and K reveals a pattern that K is not very accurately classified by methods informed by SM features alone. And in analyses informed by DCT features alone, an opposite pattern may be observed: that of T being less accurately classified than K (although this pattern is less consistent). When DCT and SM features are combined, however, the difference in classification performance between T and K reveals a pattern that K is not as pronounced. This pattern can be observed both for the training data and the test data.

The difference in accuracy between the training and test sets can be due to less prototypical productions in the SSD group compared to the TD group, or to overfitting, i.e., that the models may be describing part of the noise contained in the training data that does not generalize to the test data. To investigate the extent of overfitting, a cross-validation was performed by randomly splitting the training data into two sets, with 80% and 20% of the observations, respectively. The 80% set was then used for training, and the 20% set for testing. This procedure—random splitting and training/testing—was repeated ten times, and the average and standard deviation were registered. The resulting cross-validated accuracy figures revealed varying degrees of overfitting depending on the conditions. For the SVM(DCT), the cross-validated accuracy 83.5% (SD: 1.01) is closer to the 83.9% test accuracy than the 88.6% training accuracy. In this case, the difference between training results and test results (as presented in Table V) can largely be attributed to overfitting. For the SVM(SM), however, the cross-validated accuracy of 82.4% (SD: 1.08) is closer to the 87.0% training accuracy than to the 69.6% test accuracy. In this case, overfitting alone does not explain the performance difference. Similarly, but to a lesser extent, the cross-validated accuracy 86.1% (SD: 0.83) for the SVM(DCT + SM) shows that overfitting only partly accounts for the difference in accuracy between training (92.3%) and test (83.4%). The drop in classification accuracy from training to test is hence most apparent in the analysis based on SM features, indicating that training on SM features in TD data does not generalize very well to SSD data.

In order to explore how the acoustic similarity indices vary across the different perceptual categories, six separate Welch variance-weighted ANOVAs and follow-up Games–Howell post hoc tests were performed, one for each method (applied to the SSD data, i.e., the test set). Table VI summarizes the results of these analyses. Figure 3 illustrates these results graphically, with indications of which pairs exist; therefore, estimation of accuracy is not applicable. A first observation is that DCT features in general yield higher accuracy than SM features. Second, SVM methods usually outperform the PROX method, although the improvement is larger on the training set than on the test set. Furthermore, Table V reveals a slightly lower performance of the LDA as compared to the SVM methods, with accuracy figures revealed varying degrees of overfitting and K is not as pronounced. This pattern can be observed both for the training data and the test data.

The difference in accuracy between the training and test sets can be due to less prototypical productions in the SSD group compared to the TD group, or to overfitting, i.e., that the models may be describing part of the noise contained in the training data that does not generalize to the test data. To investigate the extent of overfitting, a cross-validation was performed by randomly splitting the training data into two sets, with 80% and 20% of the observations, respectively. The 80% set was then used for training, and the 20% set for testing. This procedure—random splitting and training/testing—was repeated ten times, and the average and standard deviation were registered. The resulting cross-validated accuracy figures revealed varying degrees of overfitting depending on the conditions. For the SVM(DCT), the cross-validated accuracy 83.5% (SD: 1.01) is closer to the 83.9% test accuracy than the 88.6% training accuracy. In this case, the difference between training results and test results (as presented in Table V) can largely be attributed to overfitting. For the SVM(SM), however, the cross-validated accuracy of 82.4% (SD: 1.08) is closer to the 87.0% training accuracy than to the 69.6% test accuracy. In this case, overfitting alone does not explain the performance difference. Similarly, but to a lesser extent, the cross-validated accuracy 86.1% (SD: 0.83) for the SVM(DCT + SM) shows that overfitting only partly accounts for the difference in accuracy between training (92.3%) and test (83.4%). The drop in classification accuracy from training to test is hence most apparent in the analysis based on SM features, indicating that training on SM features in TD data does not generalize very well to SSD data.

In order to explore how the acoustic similarity indices vary across the different perceptual categories, six separate Welch variance-weighted ANOVAs and follow-up Games–Howell post hoc tests were performed, one for each method (applied to the SSD data, i.e., the test set). Table VI summarizes the results of these analyses. Figure 3 illustrates these results graphically, with indications of which pairs exist; therefore, estimation of accuracy is not applicable. A first observation is that DCT features in general yield higher accuracy than SM features. Second, SVM methods usually outperform the PROX method, although the improvement is larger on the training set than on the test set. Furthermore, Table V reveals a slightly lower performance of the LDA compared to SVM, although the LDA appears to generalize well compared to SVM, although the LDA appears to generalize well between training and test. The best results on the test set are obtained by SVMs with DCT features.

A closer inspection of how classification accuracy varies across the two categories T and K reveals a pattern that K is not very accurately classified by methods informed by SM features alone. And in analyses informed by DCT features alone, an opposite pattern may be observed: that of T being less accurately classified than K (although this pattern is less consistent). When DCT and SM features are combined, however, the difference in classification performance between T and K is not as pronounced. This pattern can be observed both for the training data and the test data.

The difference in accuracy between the training and test sets can be due to less prototypical productions in the SSD group compared to the TD group, or to overfitting, i.e., that the models may be describing part of the noise contained in the training data that does not generalize to the test data. To investigate the extent of overfitting, a cross-validation was performed by randomly splitting the training data into two sets, with 80% and 20% of the observations, respectively. The 80% set was then used for training, and the 20% set for testing. This procedure—random splitting and training/testing—was repeated ten times, and the average and standard deviation were registered. The resulting cross-validated accuracy figures revealed varying degrees of overfitting depending on the conditions. For the SVM(DCT), the cross-validated accuracy 83.5% (SD: 1.01) is closer to the 83.9% test accuracy than the 88.6% training accuracy. In this case, the difference between training results and test results (as presented in Table V) can largely be attributed to overfitting. For the SVM(SM), however, the cross-validated accuracy of 82.4% (SD: 1.08) is closer to the 87.0% training accuracy than to the 69.6% test accuracy. In this case, overfitting alone does not explain the performance difference. Similarly, but to a lesser extent, the cross-validated accuracy 86.1% (SD: 0.83) for the SVM(DCT + SM) shows that overfitting only partly accounts for the difference in accuracy between training (92.3%) and test (83.4%). The drop in classification accuracy from training to test is hence most apparent in the analysis based on SM features, indicating that training on SM features in TD data does not generalize very well to SSD data.

In order to explore how the acoustic similarity indices vary across the different perceptual categories, six separate Welch variance-weighted ANOVAs and follow-up Games–Howell post hoc tests were performed, one for each method (applied to the SSD data, i.e., the test set). Table VI summarizes the results of these analyses. Figure 3 illustrates these results graphically, with indications of which pairs exist; therefore, estimation of accuracy is not applicable. A first observation is that DCT features in general yield higher accuracy than SM features. Second, SVM methods usually outperform the PROX method, although the improvement is larger on the training set than on the test set. Furthermore, Table V reveals a slightly lower performance of the LDA compared to SVM, although the LDA appears to generalize well between training and test. The best results on the test set are obtained by SVMs with DCT features.

A closer inspection of how classification accuracy varies across the two categories T and K reveals a pattern that K is not very accurately classified by methods informed by SM features alone. And in analyses informed by DCT features alone, an opposite pattern may be observed: that of T being less accurately classified than K (although this pattern is less consistent). When DCT and SM features are combined, however, the difference in classification performance between T and K is not as pronounced. This pattern can be observed both for the training data and the test data.

The difference in accuracy between the training and test sets can be due to less prototypical productions in the SSD group compared to the TD group, or to overfitting, i.e., that the models may be describing part of the noise contained in the training data that does not generalize to the test data. To investigate the extent of overfitting, a cross-validation was performed by randomly splitting the training data into two sets, with 80% and 20% of the observations, respectively. The 80% set was then used for training, and the 20% set for testing. This procedure—random splitting and training/testing—was repeated ten times, and the average and standard deviation were registered. The resulting cross-validated accuracy figures revealed varying degrees of overfitting depending on the conditions. For the SVM(DCT), the cross-validated accuracy 83.5% (SD: 1.01) is closer to the 83.9% test accuracy than the 88.6% training accuracy. In this case, the difference between training results and test results (as presented in Table V) can largely be attributed to overfitting. For the SVM(SM), however, the cross-validated accuracy of 82.4% (SD: 1.08) is closer to the 87.0% training accuracy than to the 69.6% test accuracy. In this case, overfitting alone does not explain the performance difference. Similarly, but to a lesser extent, the cross-validated accuracy 86.1% (SD: 0.83) for the SVM(DCT + SM) shows that overfitting only partly accounts for the difference in accuracy between training (92.3%) and test (83.4%). The drop in classification accuracy from training to test is hence most apparent in the analysis based on SM features, indicating that training on SM features in TD data does not generalize very well to SSD data.

In order to explore how the acoustic similarity indices vary across the different perceptual categories, six separate Welch variance-weighted ANOVAs and follow-up Games–Howell post hoc tests were performed, one for each method (applied to the SSD data, i.e., the test set). Table VI summarizes the results of these analyses. Figure 3 illustrates these results graphically, with indications of which pairs exist; therefore, estimation of accuracy is not applicable. A first observation is that DCT features in general yield higher accuracy than SM features. Second, SVM methods usually outperform the PROX method, although the improvement is larger on the training set than on the test set. Furthermore, Table V reveals a slightly lower performance of the LDA compared to SVM, although the LDA appears to generalize well between training and test. The best results on the test set are obtained by SVMs with DCT features.
TABLE VI. Results from the Welch variance-weighted one-way ANOVAs describing the variance of the acoustic predictors (for the six different methods) across the perceptual categories (T, Ts, Tm, Km, Ks, K), for recorded items from the SSD group.

<table>
<thead>
<tr>
<th>Method</th>
<th>F</th>
<th>Degrees of freedom (numerator)</th>
<th>Degrees of freedom (denominator)</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>PROX(DCT)</td>
<td>152.51</td>
<td>5</td>
<td>308.41</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>PROX(SM)</td>
<td>84.90</td>
<td>5</td>
<td>320.09</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>SVM(DCT)</td>
<td>206.28</td>
<td>5</td>
<td>306.17</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>SVM(SM)</td>
<td>95.27</td>
<td>5</td>
<td>312.08</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>LDA(DCT + SM)</td>
<td>199.55</td>
<td>5</td>
<td>308.65</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>SVM(DCT + SM)</td>
<td>195.51</td>
<td>5</td>
<td>313.87</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

features, and between Ks and K, where it is inverted for methods based on SM features. Note, however, that in all these cases the differences are not statistically significant.

Exploration of the association between perceptual VAS ratings and acoustic predictors reveals that most of the data are concentrated at the extreme ends of the VAS (0 and 100). It is, therefore, difficult to appreciate the distribution of the data from a normal scatter plot. In order to illustrate this problem, Fig. 4 displays a scatter plot of the raw data [SVM(DCT + SM) predictor vs VAS], and the two marginal histograms. In response to this problem, the VAS evaluations were grouped into intervals and presented as boxplots instead of scatter plots, see Fig. 5. Similar to the graphs in Fig. 3, the graphs in Fig. 5 all show a trend of the predictor values increasing from the /t/ end of the axis (represented by 0) to the /k/ end (represented by 100). However, exceptions to this trend can be observed, especially in the center of the graphs, where data are also scarcer. Table VII presents the results of six separate Spearman’s correlation analyses (one for each model, calculated on raw data), describing the associations between acoustic predictors and perceptual VAS ratings. As the table shows, the SVM model based on combined DCT and SM features results in the strongest of

FIG. 3. Boxplots showing median, quartiles and range of acoustic predictors across the six different /t/-/k/ categories, for all six methods on SSD data, i.e., the test set. (Non-significant differences, i.e., where p > 0.01, are indicated with “ns.”) TD data (i.e., the training set) for T and K is included for visual reference, highlighted in gray; note, however, that these data are not included in the statistical analysis.
these associations; however, none of the correlations are very strong.

IV. DISCUSSION

The first research question addressed in this study was whether perceptual evidence of covert contrast could be found in productions categorized as “clear substitutions.” For both contrasts examined, the results are affirmative. Thus, correct productions of /t/ are rated as being more prototypical (or /t/-like) than substitutions of [t] for intended /k/, and correct productions of /k/ are rated as being more prototypical (or /k/-like) than substitutions of [k] for intended /t/. Regarding the distinction between correct [t]s and [k]s as substitutions for intended /k/, these findings are inconsistent with the findings in Munson et al. (2012), where listeners were not able to distinguish [t] for /k/ substitutions from correct productions of /t/. These divergent findings may be attributed to differences between the languages explored (Swedish in the present study, and American English in the study of Munson et al.), or to differences in data size. Compared to the 88 stimuli in the study of Munson et al., the 3627 evaluated stimuli in the present study contribute to a high reliability of the results. On a general level, the findings presented here are consistent with the suggestion that listeners are indeed sensitive to fine-grained acoustic-phonetic differentiation in children’s correct productions of a speech sound, and productions where “the same” speech sound is inaccurately produced as a substitution for another speech sound (Munson et al., 2012). (Note, however, that whereas this perceptual sensitivity is evidenced in trained listeners, the question of whether clinically untrained listeners respond equivalently falls outside the scope of this investigation.) Considering that this perceptual sensitivity to acoustic-phonetic detail is masked in descriptions based on phonetic transcription, this finding motivates the use of perceptual instruments that reflect fine-grained acoustic variation within sound categories, e.g., the VAS.

Regarding the second research question, of whether any perceptual differences could be found between /t/’s and /k/’s produced by children with SSD compared to those produced by their peers, the results were also affirmative: for items categorized as correct productions of /t/ and /k/, the plosives produced by the children with SSD were evaluated as being less prototypical in comparison to those produced by TD peers. This is in congruence with the results in Forrest et al. (1994), where acoustic differences were found between target-appropriate [t]’s and [k]’s produced by children with SSD who are in the process of acquiring the /t/-/k/-contrast, compared to peers with typical speech and language development. The present study extends these findings to demonstrate that these acoustic traces of a newly acquired contrast (or, of a contrast that is in the process of being acquired) are also reflected in the perceptual domain. Considering that the children with SSD had specific problems producing /k/ (and other velar sounds), these perceptual responses may not come as a surprise. More unexpected is that the same pattern was observed for the children’s productions of /t/, even...
though this was not a documented problem for the children with SSD. That is, for the children with SSD, their correct [t]’s were perceived as being less /t/-like than the [t]’s produced by their TD peers. However, given that these children may signal and distinguish the /t/-/k/-contrast by acoustic-phonetic features other than those used in typical speech production (Forrest et al., 1990; Tyler et al., 1993), it makes sense that both /t/ and /k/ are affected, even though only one of the sounds—most often /k/—is categorized as being mis-articulated. This finding not only contributes to our growing understanding of phonetic detail in misarticulating children’s speech production but also highlights the potential value of using fine-grained perceptual measures, as these patterns would have passed undetected if using categorical, or otherwise coarse-grained, measures.

The third research question concerned the relation between the perceptual evaluations and the acoustic measures of the same items, and to what extent acoustic models based on prototypical correct /t/’s and /k/’s produced by children with TD could predict the perceptual evaluations performed by a human listener. In addressing this question, two different methods for feature extraction were used: spectral moments analysis and DCT analysis—and three different modeling methods: Gaussian distributions, support vector machines, and linear discriminant analysis. The results

<table>
<thead>
<tr>
<th>Method</th>
<th>( r_s )</th>
<th>( p )</th>
</tr>
</thead>
<tbody>
<tr>
<td>PROX(DCT)</td>
<td>0.56</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>PROX(SM)</td>
<td>0.45</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>SVM(DCT)</td>
<td>0.54</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>SVM(SM)</td>
<td>0.48</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>LDA(DCT+SM)</td>
<td>0.56</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>SVM(DCT+SM)</td>
<td>0.58</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

FIG. 5. Boxplots showing median, quartiles and range of acoustic predictors across the gradient perceptual evaluations with regards to the VAS, for all six methods on SSD data, i.e., the test set.
shown that, although the models were built on correct pronunciations of /k/ and /t/ alone, they could provide meaningful scores for misarticulated utterances. The difference between Ts and T on the one hand and Ks and K on the other was usually not significant, although the categories Tm and Km were usually distinguishable from each other and from their T and K counterparts. Both these results are in agreement with the perceptual results. In spite of these promising results, acoustic differentiation within these gross categories was not demonstrated to the same extent as in the perceptual evaluation. It is clear—especially from the associations between acoustic similarity measures and perceptual evaluations with regards to the VAS—that none of the presented analytic approaches reflect the same perceptual sensitivity as the human listeners. Not only are the correlations weak between the acoustic similarity measures and the perceptual evaluations, the acoustic analysis approaches also result in much higher confusability between the more fine-grained categories within the gross categories /t/ and /k/, i.e., in the distinction between target-appropriate items and substitutions and intermediate productions. So, whereas this study provides perceptual evidence of covert contrasts between children’s production of [t] for intended /t/ and [t] as a substitution for /k/, as well as for [k] for intended /k/ and [k] as a substitution for /t/, this evidence is only partly accompanied by acoustic evidence. It should be noted, however, that the human listeners had access to more acoustic information (entire CV syllables) compared to what was used as input to the acoustic analysis (the burst-to-vowel portion of the CV syllables). Regardless of whether this is why the acoustic instruments fail to show the same sensitivity as human listeners, or whether the explanation lies in how the available acoustic signal is analyzed, these results illuminate the need for continued efforts at optimizing current acoustic measures.

The findings presented here do not provide a definite answer to the question of which of the two different acoustic feature sets is most appropriate for the acoustic evaluation of children’s successful and unsuccessful efforts at producing /t/ and /k/. In terms of performance, the method based on DCT analysis achieved higher classification accuracy than the method based on spectral moments analysis, regardless of which acoustic similarity index was used (PROX or SVM). This difference in performance cannot be explained by differences in sensitivity to background noise, as both DCT and SM perform similarly on the TD data, as shown in the cross-validation results. A closer inspection of the classification of the two different consonants /t/ and /k/ reveals that the poorer classification accuracy in analyses based on SM features alone is largely driven by poorer classification of /k/ compared to that of /t/. This pattern is consistent with that reported by Forrest et al. (1990), in their spectral moments analysis of children’s production of /t/ and /k/. Notably, an opposite pattern is observed in the analysis informed by DCT features alone, where classification of /t/ is less accurate than that of /k/. However, this category-dependent difference is leveled out if the analysis involves a combination of SM and DCT features. Moreover, the choice of acoustic similarity model (PROX or SVM) was found to have only marginal effects; models based on SVM predictions generally perform slightly better. To conclude, both the spectral moments analysis and the DCT analysis perform on the same level as reported by others—a global accuracy of 80%–83% in the present study is comparable to a corresponding value of 82% in Forrest et al. (1990). However, to ensure a balanced classification accuracy—with comparable performance irrespective of category—the results presented here suggests that the acoustic analysis of the alveolar-velar contrast is best approached through a combination of SM and DCT features.

In order to control for the influence of a perceptual bias in the listener evaluations, only intermediate productions (i.e., items categorized as being in between /t/ and /k/, but either more /t/-like or more /k/-like) were examined, as these items are the only ones in the data that could be considered phonologically ambiguous. These analyses did not reveal evidence of any perceptual bias driven by lexical and/or phonotactic expectations. However, it should be noted that the acoustic-phonetic variation among these items was not controlled (e.g., as when generating synthesized continua between endpoints along some identified acoustic dimension), and that the distribution among the intermediate items (Km and Tm) was skewed, such that Tm were twice as common as Km. Hence, it is reasonable to question whether these items were indeed phonologically ambiguous. Admittedly, for studies focusing on the exploration of effects of lexical and/or phonotactic expectations on speech perception these data are not optimal; in the present study, however, the purpose of performing these analyses was secondary to the main research questions.

Another potential source of bias stems from the fact that the same person collected the recordings and then later performed perceptual evaluations. It cannot be ruled out that this person may have recognized the children’s voices and that this influenced her evaluations. However, this suspicion is disfavored by the fact that at least one year—and more often two years—had passed between the collection of the recordings and the perceptual evaluation of these recordings. Moreover, the high inter-annotator agreement in both the categorical and the continuous evaluations serves to further alleviate this concern. A possible limitation is the number of tokens selected for re-annotation, both for the listener panel evaluation (a 3% subset of the tokens) and for enabling estimation of intra- and inter-annotator agreement (a 5% subset of the tokens). Although these tokens were randomly selected, there is a risk that they are somehow not representative of the total dataset. Admittedly, larger subsets would have addressed this concern. However, these subset sizes (120 or 180 tokens in each set) were considered a maximum for what recruited listeners could be expected to evaluate during one session, at a high and stable level of attention and alertness. And indeed, the high intra-annotator agreement is an indication that the listeners did remain consistent in their evaluations. Another potential concern is the information available regarding the listeners’ hearing, as this was limited to self-reported normal hearing at the time of participation and routine audiology screening tests prior to acceptance into
their educational program. However, here too, the high level of intra-rater agreement for all included listeners serves to strengthen the confidence in the perceptual evaluations.

The recordings used in the present study were collected from previous studies where the focus had been on ecological validity (in this context: representing an expected clinical environment). This gives us the opportunity to test if acoustic instruments developed mainly on studio recordings, often performed in sound-treated booths (as, for example, in Forrest et al., 1990), are as accurate when tested on recordings more closely reflecting clinical conditions. The observation that the overall classification accuracy in the present study is comparable to that of Forrest et al. (1990) confirms that these instruments are indeed useful even with more challenging acoustic material. The recordings analyzed in the present study were made using a 16 kHz sampling rate. This can involve a risk of disregarding valuable spectral information at high frequencies (Nittrouer, 1995). However, in previous studies where higher sampling rates were used in recordings, a comparable frequency band was used in the analysis (e.g., Marín et al., 2010).

The observation that human listeners are perceptually sensitive to covert contrasts in the speech of children with SSD implies that there is indeed acoustic information signaling this contrast present in the recordings. However, this finding involving clinically trained listeners may not necessarily extend to listeners without training. Examining the perceptual effects in untrained listeners when exposed to misarticulated speech is an interesting venue for future research, in order to better capture reactions of assumed listeners in the child’s everyday environment. Another important path for future study is to relate the presented findings of the associations between perceptual and acoustic measures of children’s misarticulations to the articulatory domain, in order to better understand the articulatory patterns that underlie these perceptual and acoustic consequences. Although an articulatory pattern involving undifferentiated tongue gestures has been described in children presenting with a backing pattern, i.e., the substitution of velar sounds [k, g, θ] for alveolar sounds /t, d, n/ (Gibbon, 1999), this pattern does not necessarily describe what underlies the patterns produced by children exhibiting the reversed misarticulation pattern. In light of the articulatory-acoustic asymmetries reported for adult speakers’ misarticulated /h/’s and /k/’s (e.g., Marín et al., 2010; Pouplier and Goldstein, 2005), similar studies on children’s misarticulated speech can be expected to provide new and valuable insights both for the theoretical understanding of children’s misarticulations and for the clinical management of this large group of children.

To conclude, we have explored human listeners’ ability to gradiently evaluate children’s successful and unsuccessful efforts at producing /k/ and /t/, within an unusually large set of recordings, collected from an unusually large number of speakers. Contrary to what would be expected if listeners’ perception were purely categorical, the listeners make use of the whole range of the gradient scale from “clear [t]” to “clear [k],” differentiating not only between “clear substitutions” of [t] for /k/ and [k] as correct productions of /t/,

but also for the opposite pattern. Furthermore, even among target-appropriate productions of /t/ and /k/, items produced by children with SSD are evaluated as being less prototypic than those produced by their TD peers. These findings illustrate the value of using fine-grained perceptual measures in descriptions of children’s speech production. Moreover, the present study serves as a call for continued research into improving methods of acoustic evaluation of children’s misarticulations, to better reflect the sensitivity to phonetic detail exhibited by human evaluators.

ACKNOWLEDGMENTS

This research was supported by the Promobilia foundation, with support from The Swedish Graduate School of Language Technology (GSLT). We thank Christina Tännander at the Swedish Agency for Accessible Media, for generously providing access to frequency data from the newspaper corpus. Many thanks also to Stephanie Marín for providing access to the documentation of the analyses performed in Marín et al. (2010), and to associate editor Laura Koenig and three anonymous reviewers for their valuable comments.

APPENDIX: RECORDED WORDS

<table>
<thead>
<tr>
<th>Orthography</th>
<th>Transcription</th>
<th>In English</th>
</tr>
</thead>
<tbody>
<tr>
<td>k</td>
<td>/k/</td>
<td>the letter k</td>
</tr>
<tr>
<td>kaka</td>
<td>/ka:ka/</td>
<td>cake</td>
</tr>
<tr>
<td>kam</td>
<td>/ka:m/</td>
<td>comb</td>
</tr>
<tr>
<td>karta</td>
<td>/ka:ta/</td>
<td>map</td>
</tr>
<tr>
<td>katt</td>
<td>/ka:t/</td>
<td>cat</td>
</tr>
<tr>
<td>kavel</td>
<td>/ka:val/</td>
<td>rolling pin</td>
</tr>
<tr>
<td>ko</td>
<td>/ku/</td>
<td>cow</td>
</tr>
<tr>
<td>kopp</td>
<td>/kp/</td>
<td>cup</td>
</tr>
<tr>
<td>korg</td>
<td>/kn/</td>
<td>basket</td>
</tr>
<tr>
<td>kula</td>
<td>/ku:la/</td>
<td>ball</td>
</tr>
<tr>
<td>kulle</td>
<td>/kola/</td>
<td>hill</td>
</tr>
<tr>
<td>kung</td>
<td>/kon/</td>
<td>king</td>
</tr>
<tr>
<td>tak</td>
<td>/ta:k/</td>
<td>roof</td>
</tr>
<tr>
<td>tant</td>
<td>/tan/</td>
<td>lady</td>
</tr>
<tr>
<td>tavla</td>
<td>/ta:vla/</td>
<td>picture</td>
</tr>
<tr>
<td>tidning</td>
<td>/ti:nθ/</td>
<td>newspaper</td>
</tr>
<tr>
<td>tiger</td>
<td>/ti:ɡe/</td>
<td>tiger</td>
</tr>
<tr>
<td>tomte</td>
<td>/tomte/</td>
<td>Santa Claus</td>
</tr>
<tr>
<td>topp</td>
<td>/tɔp/</td>
<td>top</td>
</tr>
<tr>
<td>tumme</td>
<td>/tumθe/</td>
<td>thumb</td>
</tr>
<tr>
<td>tunga</td>
<td>/tʊŋa/</td>
<td>tongue</td>
</tr>
<tr>
<td>tupp</td>
<td>/top/</td>
<td>rooster</td>
</tr>
<tr>
<td>tag</td>
<td>/tɔɡ/</td>
<td>train</td>
</tr>
</tbody>
</table>

1 As outliers have been removed, the significant difference between categories K and KS is not evident from the graphical presentation in Fig. 1.
2 In the baseline condition ‘BornWords, the two words “tå” and “ß” are not equally frequent in the newspaper corpus; “ß” is more than 30 times more common than “tå.” In case of a lexical bias in the baseline condition, evaluations would be driven towards the /k/ end of the VAS, i.e., closer to 100. However, this is not evidenced in the analysis.
In comparison to other similar studies, the present is not only based on a large number of recordings \( n = 3627 \), but these recordings are also collected from a large number of children \( (117, \text{whereof 35 with SSD}) \). In Forrest et al. (1990) and Tyler et al. (1993), datasets contain about 300–400 recordings, retrieved from less than ten speakers.


