Synthetic correction of deviant speech – children’s perception of phonologically modified recordings of their own speech

Sofia Strömbergsson

Department of Speech, Music and Hearing, School of Computer Science and Communication, KTH (Royal Institute of Technology), Stockholm, Sweden
sostr@csc.kth.se

Abstract
This report describes preliminary data from a study of how children with phonological impairment (PI) perceive automatically corrected versions of their own deviant speech. The results from 8 children with PI are compared to results of a group of 20 children with typical speech and language (nPI). The results indicate group differences only in tasks where the children make judgments of their own recorded (original or modified) speech; here, the children in the nPI group perform significantly better than the children with PI. In tasks where the children judge the phonological accuracy of recordings of other children (original or modified), however, the two groups perform equally well. Furthermore, the results indicate that sub-phonemic modifications of recordings are too subtle for the children in both groups to detect. Technical and clinical implications of these findings are discussed.

Index Terms: speech perception, children, phonological impairment, unit selection

1. Background
The most apparent feature of a phonological impairment (PI) is deviant speech production. However, the problems in speech production are often accompanied by perceptual problems, although the relation between production and perception is still not fully understood. Some children who produce a sound in error also have problems with the perception of the distinction between this sound and the sound that they substitute it for [1]. Other children with the same speech deviation do perceive this error when others produce it, but still fail to perceive the same error in their own speech production [2]. This discrepancy between external perception (i.e. perception of other people’s speech) and internal perception (i.e. perception of one’s own speech) is common in children with speech sound disorders, but also in typical speech development [3].

Presenting children with recordings of their own deviant speech has been suggested as a way of increasing the children’s awareness that their speech is deviant and to stimulate self-monitoring [4]. Shuster [5] presented children and adolescents not only with recordings of their own deviant /r/ productions, but also with corrected versions of the same recordings, with the hypothesis that the children would perceive the corrected versions as incorrect, and that this could explain why they persisted in producing deviant versions of /r/. However, the results showed that the children judged both the deviant (original) /r/-recordings and the corrected (edited) /r/-recordings as correct, suggesting that the children’s internal representations of /r/ are too wide/allowing.

Shuster’s study is an example of how corrected recordings of deviant speech can be used to gain insights into children’s speech perception in relation to their speech production. A limitation, however, of the study is that the recordings were edited by hand, by LPC manipulation and subjective auditory analysis. Presumably, synthetic correction would be more valuable – and clinically useful – if the correction was generated automatically. This would not only be time-saving, but it would also allow the speaker to receive immediate feedback to his/her speech production. Moreover, immediate and automatic modification of recorded deviant speech would allow examination of the commonly held suggestion that when children with deviant speech monitor their own speech, they hear what they intended to produce, rather than the phonetic product of their intent (e.g. [6]). A finding that the children are insensitive to modifications of their recordings would speak in favor of this suggestion.

A method for automatic re-synthesis of recorded speech has been suggested in [7]. Here, a unit selection approach is used to replace an initial voiceless plosive in a recorded speech sample by another voiceless plosive, retrieved from a recording of another child. The method has been evaluated with adult listeners [7], and preliminary results from an evaluation with children with typical speech as speakers/listeners have also been reported [8]. Although the results from these studies are promising concerning the functional aspects of the technique, an evaluation like the one presented in this report, that is carried out in a naturalistic setting with the intended target group (i.e. children with PI) as users, is more revealing of its potential clinical use.

1.1. Purpose
The goal of the present study is to explore the reactions in children with deviant speech when they are exposed to both corrected and non-corrected synthetically modified versions of their own recorded speech. The children’s performance is compared to the results of an aged-matched group of children with typical speech and language.

2. Method

2.1. Participants
At the time of writing, 8 children with PI had participated. Only children displaying patterns of either velar fronting (substituting [l], [d] and [n] for /k/, /g/ and /ŋ/, respectively) or dental backing (substituting [k], [g] and [ŋ] for /t/, /d/ and /n/, respectively) were eligible for participation. (Additional phonological deviances were not disqualifying.) The children were between 4 and 7
The children judge phonological accuracy of the stimuli as in INTMOD1, but after an intervening non-related task (non-word repetition).

### 2.2. Speech modification

A method was devised for recording the children’s production of a word, and to substitute the initial voiceless plosive in the word with a plosive retrieved from another recording. This method enables two types of synthetic modification: one where the phonological identity of the first plosive is changed (e.g. changing "topp" top into "kopp" cup), and one where it is not (e.g. changing the original "topp" into a new version of "topp", with a [t] retrieved from another recording). For the re-synthesis, a unit selection approach was used. Details are described in [7].

<table>
<thead>
<tr>
<th>Task</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>EXTIDENT</td>
<td>External word identification. The children judge phonological accuracy in both original and modified recordings of other children.</td>
</tr>
<tr>
<td>INTMOD1</td>
<td>Internal modified identification – immediate. The children judge phonological accuracy in original and synthetically modified (both corrected and non-corrected) recordings of their own voice, immediately after having produced the stimulus.</td>
</tr>
<tr>
<td>INTMOD2</td>
<td>Internal modified identification – delayed. The children judge phonological accuracy in the same stimuli as in INTMOD1, but after an intervening non-related task (non-word repetition).</td>
</tr>
<tr>
<td>INTSPKR</td>
<td>Internal speaker judgment. The children judge the speaker identity in original and synthetically modified (both corrected and non-corrected) recordings of their own voice.</td>
</tr>
<tr>
<td>INTIDENT1</td>
<td>Internal word identification – immediate. The children judge phonological accuracy of a word they have just produced, i.e. an accuracy judgment of their online, non-recorded, speech production.</td>
</tr>
<tr>
<td>INTIDENT2</td>
<td>Internal word identification – delayed. The children judge phonological accuracy of the same stimuli as in INTIDENT1.</td>
</tr>
</tbody>
</table>

### 2.3. Procedure

All children participated in two 20-minute test sessions, in which they performed a range of speech perception tasks, see Table 1. The experiment took place in a separate room with limited noise at the different preschools. For all experimental tasks, the children were fitted with a headset and the experimenter with headphones to supervise the recordings and listening tasks.

In EXTIDENT, the children were presented with 12 original (N = 4) and synthetically modified (N = 8) recordings of other children. The task for the children was to judge the phonological accuracy of each stimulus, by pointing either to a picture illustrating the stimulus word on the screen, or to a picture of a large X when they perceived the word as “wrong”, similar to the setup illustrated in Figure 1.

In INTMOD1, INTMOD2 and INTSPKR, a recording script of 8 words was used (see Appendix). In each task, the words in the script were repeated twice, resulting in 16 stimuli per task.

In INTMOD1, the children were recorded when producing the words in the script. Immediately after having recorded a word, the recording was either a) re-synthesized with a different initial consonant (N = 6), b) re-synthesized with an identical initial consonant retrieved from a recording of another child (N = 6), or c) left unchanged (N = 4). (The number of stimuli beginning with /t/ and stimuli beginning with /k/ was equal in all three conditions.) The children were told that The Parrot would imitate what they had just said, but that The Parrot would not always get it right. The task for the children was to, for each word in the script, judge if The Parrot could say it “right” or “wrong”, as illustrated in Figure 1.

In INTMOD2, the children were again presented with the same 16 stimuli that were recorded/generated in INTMOD1. Again, their task was to judge if The Parrot’s production was “right” or “wrong”.

The INTSPKR task was designed to test whether the children could detect the synthetic modification of their recordings. In the same way as in INTMOD1, the children were first recorded when producing a word, which then immediately was either a) re-synthesized with a different initial consonant, b) re-synthesized with an identical initial consonant, or c) left unchanged. This time, the children were told that The Imitation Monkey (“Härmapan”) would imitate what they had just said, but that he would only do it occasionally. The task for the children was to point to a picture of a child if they believed that the recording sounded exactly as they had produced it, or to a picture of The.
In the two INTIDENT tasks, the children were asked to judge the phonological accuracy of words they had produced, the first time immediately after having produced it (i.e. judge their online non-recorded speech, INTIDENT1), and the second time by listening to a recording of the same production (INTIDENT2). As the children with typical speech were never expected to produce deviant speech, this task was only performed by the children with PI.

2.4. Annotation and reliability

The experimenter (a certified SLT) annotated phonological accuracy of all stimuli (original and modified recordings). For each recording, the experimenter judged if the child’s production deviated phonologically from the target word. So, for example, if the word “topp” (‘top’, phonological target: /tɔp/) was produced [ˈtɔp], this would be annotated as phonologically deviant. A phonetic deviation of the type [ˈtp], however, would not be annotated as phonologically deviant. A second SLT annotated a randomly extracted subset of the stimuli (85% of the recordings, N=300). Inter-annotator agreement on this set, as measured by Cohen’s kappa, was 92% (p < 0.001).

3. Results

3.1. General results

The children’s performance on the different tasks is presented in Figure 2. The graph indicates significant differences between the two groups in INTMOD1 and INTMOD2, i.e. the tasks where the children judge phonological accuracy of their own recorded (original and modified) speech, immediately after having produced a word and after a delay. A repeated-measures ANOVA was conducted with language group as a between-subjects factor and playback condition (immediate vs. delayed) as a within-subjects factor. This established the group difference as significant (F(1,26) = 6.57, p < .01), whereas there was no significant effect of test condition (F(1,26) = .63, p =.43) or of the interaction between the factors (F(1,26) = .21, p = .65). (Average performance on INTMOD1 was 13.25 of 16 (SD = 2.2) for the PI group and 14.55 (SD = 1.2) for the nPI group, and on INTMOD2, average performance was 12.88 (SD = 1.7) for the PI group and 14.45 (SD = 1.4) for the nPI group.)

4. Discussion

The results in this study provide insights into how children with deviant speech perceive their own speech – in its online form, in its recorded form, and in synthetically modified forms. A general observation is that for children with PI, listening to one’s own
recorded speech (internal perception) – original or modified – is clearly something different from listening to someone else’s recorded speech (external perception). In the one task where the children judge recordings of other children’s utterances, the children with PI and the children in the nPI group perform on the same level, whereas in all other tasks – which all involve making judgments about one’s own recorded speech – the performance of the children with PI does not meet the performance level of the children in the nPI group. This is in line with previous research suggesting that external speech perception is often not a problem in children with PI, but that it is rather perception of one’s own speech – internal perception – that constitutes a challenge [1, 6].

For both groups of children, it is more difficult to detect when modification has taken place than to judge the phonological accuracy of the stimuli. The finding that the children often mistake modified stimuli where the phonology is not changed as having been produced by themselves, suggests that sub-phonemic modification is too subtle for the children to detect. For modified stimuli where the phonology is changed, however, the children most often perceive the change. In other words, the children generally accept the intended phonological form of the modified stimuli, seemingly without detecting the modification itself. As an evaluation of the synthesis technique, this finding is encouraging.

Previous research has found that children with deviant speech have more difficulties than children with typical speech and language development in speech perception tasks with suboptimal speech stimuli [11]. In light of this, it is reasonable to ask whether the fact that most stimuli in the present study are synthetically generated could explain why the children with PI perform worse than the children in the nPI group. Disfavoring this suggestion is the observation that for the children with PI, their performance on judging phonological accuracy of their own recorded non-modified speech (in INTIDENT2) is not better than their performance on making the same kind of decision on synthetically modified versions of their recorded speech (in INTMOD2). This suggests that their difficulty to judge phonological accuracy of their own modified speech is not an effect of the synthetic modification itself.

As mentioned earlier, when children with deviant speech judge the phonological accuracy of their own speech, it has been suggested that they actually judge their intent rather than the phonetic product of this intent [6]. In support of this account are observations that children, when exposed to recordings of their own deviant speech, actually perceive the deviance that they did not react to when producing the utterance. The present study aimed to test this suggestion, by letting the children judge both their online speech production and the recorded versions of the same utterances. However, there was no significant difference in the children’s performance between these two conditions, and therefore, we could not verify the suggested account. However, there is a large variation in the children’s performance, and it is possible that the inclusion of more children will reduce this variation and yield significant results.

An important finding in this study is that children with PI differ from the children in the nPI group only in tasks of internal perception; here, the children in the nPI group perform significantly better than the children with PI. Clinically, this can be interpreted as an indication that attending to and evaluating one’s own speech production is an important focus of intervention. In tasks of external perception, however, the two groups perform equally well. Furthermore, the results indicate that sub-phonemic modifications of recordings are too subtle for the children in both groups to detect. Regarding this study as an ecologically valid evaluation of the quality of the synthetically modified speech, these results are encouraging, as the children generally accept the intended phonological form, without detecting the modification.

5. Acknowledgements

This work was funded partly by The Swedish Graduate School of Language Technology, and partly by the Promobilia foundation.

6. References


7. Appendix

<table>
<thead>
<tr>
<th>Orthography</th>
<th>Transcription</th>
<th>In English</th>
</tr>
</thead>
<tbody>
<tr>
<td>k</td>
<td>/ˈko:/</td>
<td>(the letter k)</td>
</tr>
<tr>
<td>kaka</td>
<td>/ˈkɑːkɑ:/</td>
<td>cake</td>
</tr>
<tr>
<td>kulle</td>
<td>/ˈkuːlː/</td>
<td>hill</td>
</tr>
<tr>
<td>kung</td>
<td>/ˈkʊŋ/</td>
<td>king</td>
</tr>
<tr>
<td>tak</td>
<td>/ˈtɑːk/</td>
<td>roof</td>
</tr>
<tr>
<td>tumme</td>
<td>/ˈtʊmː/</td>
<td>thumb</td>
</tr>
<tr>
<td>tupp</td>
<td>/ˈtʊp/</td>
<td>rooster</td>
</tr>
<tr>
<td>tåg</td>
<td>/ˈtɒɡ/</td>
<td>train</td>
</tr>
</tbody>
</table>