Native and non-native listeners' judgement on the overall speech quality of hearing-impaired children

Authors

Nathalie Boonen (corresponding author)

Computational Linguistics & Psycholinguistics Research Centre, University of Antwerp, Antwerp, Belgium

E-mail: nathalie.boonen@uantwerpen.be

University of Antwerp

Prinsstraat 13

2000 Antwerp

Belgium

Phone: +3232655234

Fax: +3232655898

Hanne Kloots

Computational Linguistics & Psycholinguistics Research Centre, University of Antwerp, Antwerp, Belgium

E-mail: hanne.kloots@uantwerpen.be

Steven Gillis

Computational Linguistics & Psycholinguistics Research Centre, University of Antwerp, Antwerp, Belgium

E-mail: steven.gillis@uantwerpen.be

Abstract

This study focuses on the influence of listeners' native language on their judgement of the overall speech quality of normally hearing (NH) and hearing-impaired (HI) children. Studies have shown that listeners' native language influences their judgements on linguistic aspects of a foreign language. Since judging speech quality does in principle not require any knowledge of the language, the question arose if the native language influences listeners' judgements. For this purpose, the overall speech quality of seven-year-old Dutch speaking children (n = 21) with an acoustic hearing aid (HA), a cochlear implant (CI) and normal hearing (NH) was judged by four listener groups (native speakers of Italian, German, French and Dutch). Listeners completed a comparative judgement task in which stimuli were presented in pairs. For each pair, they selected the better sounding stimulus. This procedure ultimately led to a ranking of the stimuli according to their speech quality. The ranking showed that NH children had a significantly higher speech quality than HI children. Interestingly, there was no significant effect of language background. Both native and non-native listeners perceived a significant difference in speech quality. Also, within the group of HI children, all listener groups preferred the speech of CI children when comparing them to HA children. These results indicated that the differences were purely speech related since the non-native listeners had no linguistic knowledge of Dutch. Considering that all listeners perceived a similar qualitative difference, we conclude that there was no transfer of native language in this type of judgement.

Keywords: Comparative judgement; Non-native listeners; Children with a cochlear implant; Children with an acoustic hearing aid

Introduction

Acoustic and perceptual studies on the speech of hearing-impaired children

Hearing loss affects the speech and language development of children (Gillis, 2017; Osberger & McGarr, 1982). Based on the place and the severity of this hearing loss, a hearing device such as a cochlear implant (CI) or an acoustic hearing aid (HA) may be provided to these children (Korver et al., 2017). The period of time before receiving a device, the fact that their hearing cannot be fully restored and/or a deteriorated incoming signal may lead to an (initial) delay in their speech and language development (Korver et al., 2017).

But although children with a hearing impairment (HI) have an initial delay in their speech and language development, some studies report that children who receive a device at an early age eventually catch up with their normally hearing (NH) peers. For example in the study of Geers and Nicholas (2013), 73% of the 10-year-old children implanted before the age of 18 months reached language perception and production scores on a par with their NH peers. Concerning their speech development, it was found that some children catch up nearly completely on the speech of their NH peers (Chin & Kuhns, 2014), whereas others do not reach this level despite long term device use (Fang et al., 2014; Freeman et al., 2017; Tomblin et al., 1999; Verhoeven et al., 2016).

Comparing the speech outcomes of children with CI and HA, children with CI tend to reach better scores (Baudonck et al., 2010a; Baudonck et al., 2010b; Spencer et al., 1998; Tomblin et al., 1999; Yoshinaga-Itano et al., 2010). Children with CI and HA also seem to differ in how length of device use affects their speech. For children with CI, speech continues to improve after implantation (Bakhshaee et al., 2007; De Raeve, 2010; Peng et al., 2004). For children with HA, length of device use is often not reported or the effect for this group is minimal or non-existent (Lejeune & Demanez, 2006; Tomblin et al., 1999).

Non-native perception

The speech of HI children is either assessed in acoustic studies or perceptual studies. Up to date, the participating listeners in perceptual studies either had the same native language as the children whose speech they judged (i.a. Chin & Kuhns, 2014; Ching et al., 2015; Montag et al., 2014) or the native language of the listeners was not specified (i.a. AlSanosi & Hassan, 2014; Ertmer, 2007; Fang et al., 2014; Kloiber & Ertmer, 2015; Tobey et al., 2004). Perceptual studies with non-native listeners already exist, but they often have a completely different goal and usually involve adult speech. Often, these experiments focus on the classification or identification of specific sounds (Escudero et al., 2009; Escudero et al., 2014; Flege et al., 1997; Holden & Nearey, 1986), in which listeners classify speech sounds of a non-native language into (phoneme) categories. Interestingly, listeners with a different native language tend to label the same sounds differently. Thus, many studies show a crosslinguistic influence, i.e., an influence of the native language on the perception of a foreign language (Escudero et al., 2014; Gandour et al., 2000; Gottfried & Beddor, 1988; Williams & Escudero, 2014).

In principle, crosslinguistic influence can take place in all linguistic domains, leading to e.g., phonological, semantic and syntactic transfer (Jarvis & Pavlenko, 2008). For example, phonological transfer "refer[s] to the ways in which a person's knowledge of the sound system of one language can affect that person's perception and production of speech sounds in another language" (Jarvis & Pavlenko, 2008: 62). In other words: the perception of sounds is influenced by the native phonological system of listeners.

Although crosslinguistic transfer has been well documented in all linguistic domains or *language* levels, e.g., phonology, grammar, vocabulary, it is relatively unclear whether transfer also takes place on the *speech* level. For a thorough understanding of our study, the distinction between language and speech is of crucial importance. Speech can be considered as a *medium* of language and is in this respect completely language independent (Abercrombie, 1967). Speech quality, or rather sound quality was already judged by non-native listeners in Tang (2009). In that study Dutch and English listeners judged the sound quality of Chinese speech samples in terms of "how intelligible the fragment would be if they were a native listener of the same language that was spoken in the fragment" (Tang, 2009: 66). Without having any knowledge of Chinese, the listeners were able to reliably differentiate between good and poor sound quality of the Chinese speech samples. The present study focuses on the speech of HI children. To the best of our knowledge, this type of speech has never been judged by non-native listeners without knowledge of the native language of the children in the study in order to detect the possible influence of the listeners' native language. Our experiment will – at least partially – fill this gap.

Overall speech quality

In this study, listeners judged the overall speech quality of HI children. Overall speech quality does not refer to the content or the message of an utterance, but to the general impression of the utterance itself (Kondo, 2012). Various characteristics of the spoken message such as its "naturalness", "clarity", "pleasantness", "brightness", "hoarseness" are considered to determine together the general impression of speech quality (Loizou, 2011: 624). These voice qualities can in principle be judged without assessing the message that is intended to be conveyed by the speaker. In a previous perceptual study, native speakers of Belgian Dutch heard a difference in the overall speech quality between Belgian Dutch speaking NH and HI children: NH children appeared to sound "better" than the HI children (Boonen et al., 2020). Thus, the question turns up if this is also the case when the speech is judged by non-native listeners of Dutch.

From a methodological perspective, perceptual studies on the speech of HI children mostly use scales (Fang et al., 2014; Van Lierde et al., 2005). One of the most frequently used scales is ordinal, e.g., a 7-point scale. On such a scale, a particular characteristic of children's language or speech is given a score between one and seven (Colman, 2009). For instance, one means that the overall speech quality is very low, and seven means that it is very high. However, these scales are perceived as cumbersome, especially for listeners who are unfamiliar with judging speech on a rating scale (Ellis & Beltyukova, 2008; Munson et al., 2012). In order to score on a scale, a judge needs to have an internal reference point as to what constitutes "high/low speech quality", or what constitutes the difference between a score of three versus four on the scale. This reference point is purely mental, probably implicit, and difficult to keep consistent throughout rating. That is why rating scale tasks – especially with inexperienced listeners – are associated with low reliability scores (Munson et al., 2012; Schiavetti, 1992).

The present study explores an alternative approach, viz. comparative judgement. In such a task, stimuli are not presented individually but in pairs. More specifically, participants do not score individual stimuli by means of an ordinal rating but rather compare stimuli in pairs. For each pair of two randomly selected stimuli, the listener indicates the stimulus with the higher overall speech quality. Ultimately, these judgements lead to a ranking of the stimuli according to their speech quality. A pairwise comparison task has several advantages. First, judging stimuli in pairs, relative to one another, is less difficult than judging a single stimulus on a scale (Bejar, 2012). Independent of the type of task, participants will search for a reference point to orientate themselves when rating a stimulus. In a comparative judgement task, reference points are explicitly present, since the reference point of stimulus A is stimulus B and vice versa. This explicit reference point in pairwise comparisons not only makes the task easier, but the results are also more consistent and reliable (Lesterhuis et al., 2017).

Native perception of overall speech quality

When native listeners judge the overall speech quality of children, the question is whether they can judge and compare the speech of children who have the same native language in an unprejudiced way. Can they focus exclusively on the children's speech quality or are they influenced by their knowledge and experience with their native language? Are they, for example, influenced by the regional accent of the children? These questions are particularly relevant for Belgian Dutch, i.e., Flemish, listeners. On a daily basis, Flemish people – and therefore also Flemish children – usually speak *tussentaal* (literally: in-between language), i.e., a variety that is situated between standard Dutch and the local dialects. By definition, this spoken variety exhibits regional characteristics and thus reveals the regional background of the speaker (Lybaert & Delarue, 2017). Flemish listeners are able to correctly identify some of these regional varieties (Grondelaers & Lybaert, 2017).

In an experimental context, Belgian Dutch listeners are influenced by their regional background when judging speech. More specifically, they tend to perceive and evaluate speech using their own regional variety as a personal point of reference. In other words, "Belgian listeners evaluate the samples as Limburgians, Brabantians, Antwerpians and East and West Flemings (instead of Dutch-speaking Belgians)" (Grondelaers et al., 2011: 221). Moreover, Belgian Dutch listeners consider their own regional variety to be more beautiful than the varieties of other Dutch speaking regions in Belgium and the Netherlands (Impe, 2010). Following this reasoning, when the regionally accented variety spoken by a particular child does not match the region of origin of a particular listener, that listener may appreciate the sample of the child less favourably. Thus, when Belgian Dutch listeners hear the speech of children from different regions of Flanders, it is reasonable to expect that (1) these children will speak *tussentaal* (because their parents do), (2) the regional background of these children will be detectable in their speech, and (3) listeners will evaluate children's speech using their own regional variety as a point of reference. When investigating speech quality, listeners will thus not only hear and evaluate the quality of speech, but they may also take into account their (lack of) appreciation for the regional characteristics of other speakers. In contrast, listeners without any knowledge of – in this case – Dutch cannot take into account the regional background of a speaker in their judgements.

Aims of this study

In this study, a comparative judgement task is administered to investigate native and non-native listeners' appreciation of the overall speech quality of Dutch speaking (NH and HI) children. Concerning the listener groups, the main question is whether non-native listeners' judgements on the speech quality of NH and HI children are comparable to those of native listeners. There has been a large tradition of studying the crosslinguistic transfer in linguistic domains, yet only very little is known about whether crosslinguistic transfer also takes place on the speech level. Moreover, regional characteristics affect the speech quality judgements of native listeners, yet do not play a role in non-native listeners without knowledge of the language that is judged. Therefore, there is a need to replicate the experiment with non-native listeners. These listeners are influenced by their own linguistic background, but they cannot take into account the (negative) connotations that Belgian Dutch listeners might have concerning regional aspects. Until now, no other perceptual study involving native and non-native listeners has focused on aspects such as overall speech quality.

Research has shown that the typological distance between two languages influences how these languages are perceived by their respective native speakers (van Heuven, 2008; Wang & van Heuven, 2004). Therefore, two listener groups with a Germanic language background, viz. Dutch and German, and two listener groups with a Romance language background, viz. French and Italian, participated. We hypothesize that at least the results of the French and the Italian participants, resp. the Dutch and German participants will be comparable. Moreover, since Dutch and German are typologically closely related (Dalby, 2006; Elmentaler, 2009; Gooskens et al., 2018), the distance between them is smaller than that between Dutch and the Romance languages French and Italian. Hence, we hypothesize on typological grounds that the patterns in the judgements of the German listeners will be most similar to those of the Dutch speaking listeners.

Concerning the overall speech quality of NH and HI children, we hypothesize that listeners will consistently prefer the speech of NH children over that of HI children. Between children with CI and HA, we hypothesize that a perceivable qualitative difference will be in favour of the children with CI, since research comparing the two groups has revealed a slight advantage of children with CI (Baudonck et al., 2010a; Baudonck et al., 2010b; Tomblin et al., 1999). We also hypothesize that the effect of length of device use will be more prominent in children with CI than in children with HA (Peng et al., 2004; Tomblin et al., 1999). Moreover, it will be of interest whether the listener groups exhibit similar patterns in their judgements. When native and non-native listeners both hear a clear difference between NH and HI children and judge the general quality of NH and HI samples differently, we can conclude that the difference between NH and HI children is indeed part of the *speech* level (Tang, 2009). On the other hand, when native listeners hear a difference between both groups and judge their speech differently, whereas the non-native listeners do not, we can conclude that knowledge of the *language* system – i.c. phonology – is necessary to differentiate between children with and without a HI (Escudero et al., 2014; Jarvis & Pavlenko, 2008).

Method

In this study, listeners with a different language background judged the overall speech quality of Dutch speaking children with NH, children with an acoustic HA as well as children with a CI in a comparative judgement task. The main aim was to investigate whether native speakers of German, French and Italian, appreciated the overall speech quality of Dutch speaking children's speech in a similar way as native Dutch speaking listeners. This study was approved by the Ethics Committee for the Social Sciences and Humanities (SHW_15_37) of the University of Antwerp.

Listeners

Four groups of adult listeners (n = 81) with different native languages participated in this perceptual study (native speakers of Dutch (n = 20), Italian (n = 21), French (n = 20) and German (n = 20)). The latter three groups did not report any knowledge of the Dutch language. No participant reported any hearing problems, and no participant reported any particular experience or familiarity with the speech of HI children. All participants were either acquaintances of one of the authors or they were recruited through the help of (foreign) colleagues. Prior to the experiment, they were informed about the goal and the procedure of the study.

Stimuli

The stimuli used in the present study were selected from recordings made for a previous study on the speech of hearing-impaired (HI) children (Hide, 2013). For that study, speech of one hundred-eleven children was collected: 90 NH children, 10 children with HA and 11 children with CI. They were all monolingual native speakers of Dutch living in Flanders, the northern Dutch speaking part of Belgium. All participants attended mainstream education. The task administered consisted of an imitation task: the children imitated several times short utterances "Ik heb X gezegd" ("I have said X"), where X represented /lVlV/ (with V = /a/, /e/ or /o/), resulting in stimuli like "Ik heb lolo/lala/lele gezegd". All recordings were made with the same recording equipment and in comparable ambient circumstances in the comfort of the children's homes or schools.

For the present study, seven NH, seven HA and seven CI children were selected at random. Of each child, six utterances were again randomly selected, two for each of the vowels /a, e, o/. This resulted in a total of 126 stimuli in the present study. Detailed information on the hearing-impaired children can be found in Boonen et al. (2019) and Hide (2013).

Children with CI

At the time of the recording, the mean chronological age of the children with CI (four girls, three boys) was 7;10 (years;months) (SD = 1;1). They were implanted at a mean age of 12 months (SD = 6 months). The average length of device use was 6;9 years at the time of the recording (SD = 1;5). All but one were implanted bilaterally and had a mean of 3;11 years of bilateral device experience (SD = 1;11). The average unaided hearing threshold was 116 dB HL (decibel hearing level, SD = 7 dB HL). After implantation, the mean hearing threshold was 29 dB HL (SD = 7 dB HL). Other than their hearing loss, no additional disabilities were reported.

Children with HA

The children with bilateral HA (four girls, three boys) were on average 7;9 years old (SD = 0;11) at the time of the recording. This was not significantly different from the age of the children with CIs (Wilcoxon Rank Sum Test: $z = 0.00$, $p = 1.0$). Hearing aids were on average provided at 0;11 (SD = 0;7), meaning that they had 6;10 years of device use at the time of the recording (SD = 1;6). Their mean unaided hearing threshold was 66 dB HL (SD = 15 dB HL). With a mean of 33 dB HL (SD = 7 dB HL), their aided hearing loss was comparable with that of the children with CI (Wilcoxon Rank Sum Test: $z = 0.91$, $p = 0.37$). Other than the hearing loss, no additional disabilities were apparent at the time of testing.

Children with NH

Seven NH controls (four girls, three boys) participated in this study. They attended the same school as the children with CI. Moreover, they were matched with the HI children on gender, age and regional background. As a part of the Universal Neonatal Hearing Screening, their hearing was checked with an automated auditory brainstem response test (AABR) or otoacoustic emissions (OAE). Similar to the HI children, no health-related problems were reported.

Procedure

The comparative judgement task was implemented in the online tool D-PAC (Digital Platform for the Assessment of Competences, accessible via https://www.d-pac.be/english/) (Lesterhuis et al., 2017). Participants received a personal login and completed the experiment at home.

In their own native language, the listeners were informed that they would hear short speech samples of Dutch speaking children. They also knew that the presented utterances were all the same carrier sentence containing a nonsense word with varying vowels. The nonsense words in each pair could contain the same or different vowels. The listeners were informed that children with a different hearing status were included, i.e., children with NH, HA or CI. However, the listeners did not receive any information about the technical differences of both hearing devices, nor the implications this could have on the speech of these children. They were also instructed not to pay attention to potential differences in the loudness or the quality of the recordings. Although the same equipment was used for all recordings and the recordings were made in similar circumstances, there appeared to be some differences between the recordings. However, in order to avoid that speech quality would be affected, the speech signal was not manipulated post recording. Other than these general guidelines, participants did not receive any restrictions with respect to the aspects they should consider in their judgements.

In the present study, the overall speech quality of 126 stimuli was examined. This number of stimuli would result in a total of 7,875 possible pairs. Obviously, it would be impossible for a single listener to rate this number of pairs. Fortunately, an exhaustive pairing of all the stimuli is not necessary for a reliable ranking in a comparative judgement task using D-PAC. In order to obtain reliable results, stimuli are judged in a number of rounds. With each round, the reliability of the stimulus' perceived speech quality increases. Previous research has shown that this reliability steadily improved up to 20 rounds (Verhavert, 2018). At that point, a plateau was reached and the reliability barely increased anymore. Therefore, in this study as well, each stimulus was judged 20 times by each listener group. In order to reach this number of comparisons, each participant completed 63 pairs (rounded off to 65, see formula (1)). Thus, in a comparative judgement task, not all possible comparisons were made. Rather, all stimuli were judged 20 times by the listeners of each listener group. The selection of the 65 pairs differed for each listener. For each pair of stimuli, the same procedure was followed. Participants listened to both stimuli through headphones as many times as they wished and then selected whether stimulus A or stimulus B had the higher speech quality by pushing a button on the screen. This process was repeated for all pairs at each participant's own pace.

(1) number of comparisons per listener = $\left. \dfrac{number\ of\ stimuli \times 20}{number\ of\ listeners} \right/ 2$

Data analysis

After collecting the data on which stimulus was selected as the one with the higher speech quality, a misfit analysis was conducted in order to verify whether certain participants deviated from the participants' overall consensus (Lesterhuis et al., 2017). Moreover, the Bradley-Terry-Luce (BTL) model was applied (Bradley & Terry, 1952; Lesterhuis et al., 2017). The BTL model was used to compute for each stimulus how likely it was to be selected as the better sounding stimulus, or in other words, how likely it was to "win a competition" in a comparative judgement. These scores increased with increasing speech quality. Therefore, the scores were used to rank the stimuli according to increasing overall speech quality. The BTL score was computed according to the formula in (2):

(2) $p\left(x_{ij} = 1 | v_i, v_j\right) = \dfrac{e^{(v_j - v_i)}}{1+e^{(v_j - v_i)}}$

where x_{ij} = 1 if stimulus *j* was considered better sounding than stimulus *i*. v_i and v_j were the estimated logits scores of the respective stimuli (Verhavert et al., 2018)

The likelihood scores were expressed in logits (Verhavert et al., 2018). Thus, the individual logit score of each stimulus was determined, representing the overall speech quality perceived by a listener group. Consequently, for each listener group, all stimuli were ordered in a ranking according to their logits. Since the scores for the four listener groups differed in terms of their range, the logit scores of the four rankings were normalised by applying a z-score conversion. Each stimulus thus had a z-score representing its speech quality: the stimulus with the highest perceived speech quality was marked with the highest z-score, and the stimulus with the lowest speech quality had the lowest z-score. By means of these scores, a ranking going from the lowest speech quality to the highest speech quality was determined for each listener group.

The statistical analysis was conducted in two consecutive steps. In the first part, the overall speech quality of NH and HI children was examined by means of the z-score of each stimulus. These z-scores were the predicted (or dependent) variable in the first analysis. The second analysis examined all the pairwise comparisons. The pairs that were presented to the listeners consisted of two randomly paired stimuli, so six combinations of hearing statuses were possible. Either two stimuli produced by children with the same hearing status were combined (i.e., two utterances of children with NH, CI or HA), or two stimuli of different hearing statuses, resulting in a combination of either (1) a CI and a NH child, (2) a HA and a NH child, or (3) a CI and a HA child. An analysis of the pairs in which the two stimuli originated from children with the same hearing status was beyond the scope of this study. Hence, only the pairs of two different hearing statuses were included in the second analysis. For the analysis of the pairs, the dependent variable was binomial: (not) being selected as the better sounding stimulus in a pair.

Multilevel mixed-effect modeling (MLM) was used for the statistical analyses of this study (Baayen, 2008). These analyses were performed in the open source software R (R Core Team, 2016). In MLM, a model consists of fixed and/or random effects. Fixed effects are repeatable variables, whereas random effects are specific to the study. In an iterative process, random and fixed effects are entered separately into models until the best fitting model is found. Each fixed effect has a reference category which is mentioned in the tables. Unless otherwise stated, only the best fitting model for each analysis is reported in the results section. A significance level of $p < 0.05$ was set.

For both analyses, the fixed effects were Hearing status (with values NH, HA and CI), Length of device use, i.e., the period of time in which the children used their assistive device, and Listener group (with values German, Italian, French and Dutch). The random effects depended on the analysis. In the analysis on the separate pairs, the random effects were the individual utterances, children and listeners. The random effects in the analysis of the rankings were the individual children and utterances.

Results

This study investigated how the overall speech quality of NH and HI children (which consisted of children with a CI and children with an acoustic HA) was perceived by listener groups with a different language background. Originally, 81 listeners participated in this study. However, a misfit analysis (Lesterhuis et al., 2017) showed that four participants (one Italian, one French and two Dutch speaking listeners) exhibited substantially divergent judgements in comparison to the other participants. Their responses deviated more than two standard deviations from the group means and were therefore excluded from the analyses (Lesterhuis et al., 2017).

Place in ranking

Place in ranking for NH and HI children

In the first instance, it was investigated whether listeners with a different language background rated the overall speech quality of NH and HI children differently. Children with CI and HA were thus considered as one group. Since the stimuli's place in the ranking was determined by the (z-score converted) logit values resulting from the application of the BTL model, the dependent variable of this MLM analysis was numerical.

For each listener group, the ranking of each stimulus as determined by its z-score is shown in Figure 1. How were NH and HI children positioned in the rankings? At first sight, it appeared that NH and HI children formed two distinct groups. On the left side, i.e., the upper places in the ranking, the majority of NH children's stimuli were situated, whereas the stimuli of the HI children were roughly on the right. This indicated that the overall speech quality of both groups differed: compared to NH children, HI children's speech quality was perceived as less high. Moreover, all rankings looked fairly similar. This suggested that all listener groups, independent of their native language, perceived a similar qualitative difference between the speech of NH and HI children.

These observations were confirmed by the MLM analysis of the z-scores. Table 1 shows that the position in the ranking for utterances of NH children was significantly higher ($p < 0.0001$) than that of HI children. In terms of z-scores, HI children's estimated position in the ranking (derived from their estimated z-score) was –0.40 (SE = 0.17), whereas NH children were positioned at 0.79 (SE = 0.27). Thus, children with NH were attributed a higher overall speech quality than children with HI. In constructing the best fitting model, the factor Listener group, i.e., listeners with Dutch, German, French or Italian as their native language, was added as a fixed effect. Interestingly, this factor did not contribute to a significantly better model fit, as reported in Table 1. This result indicated that the position of NH and HI children in the ranking was indeed comparable for all listener groups. In other words: despite having a different language background and in the case of the German, Italian and French listeners, not knowing

Dutch, all listeners had a similar concept of the overall speech quality of Dutch speaking NH and HI children.

	Estimate	Std. Error	t-value	p
Intercept	–0.396	0.169	–2.340	< 0.05
Hearing status [NH]	1.187	0.274	4.334	< 0.0001
Listener group [French]	1.240 e–15	0.073	1.698 e–14	1.000
Listener group [German]	–3.968 e–11	0.073	–5.436 e–10	1.000
Listener group [Italian]	–1.587 e–11	0.073	–2.174 e–10	1.000

Table 1: Z-scores representing the place on the ranking for NH and HI children (fixed effects = hearing status (NH of HI (= reference category) and Listener group (Dutch (= reference category), French, German or Italian); random effects = individual children and individual utterances)

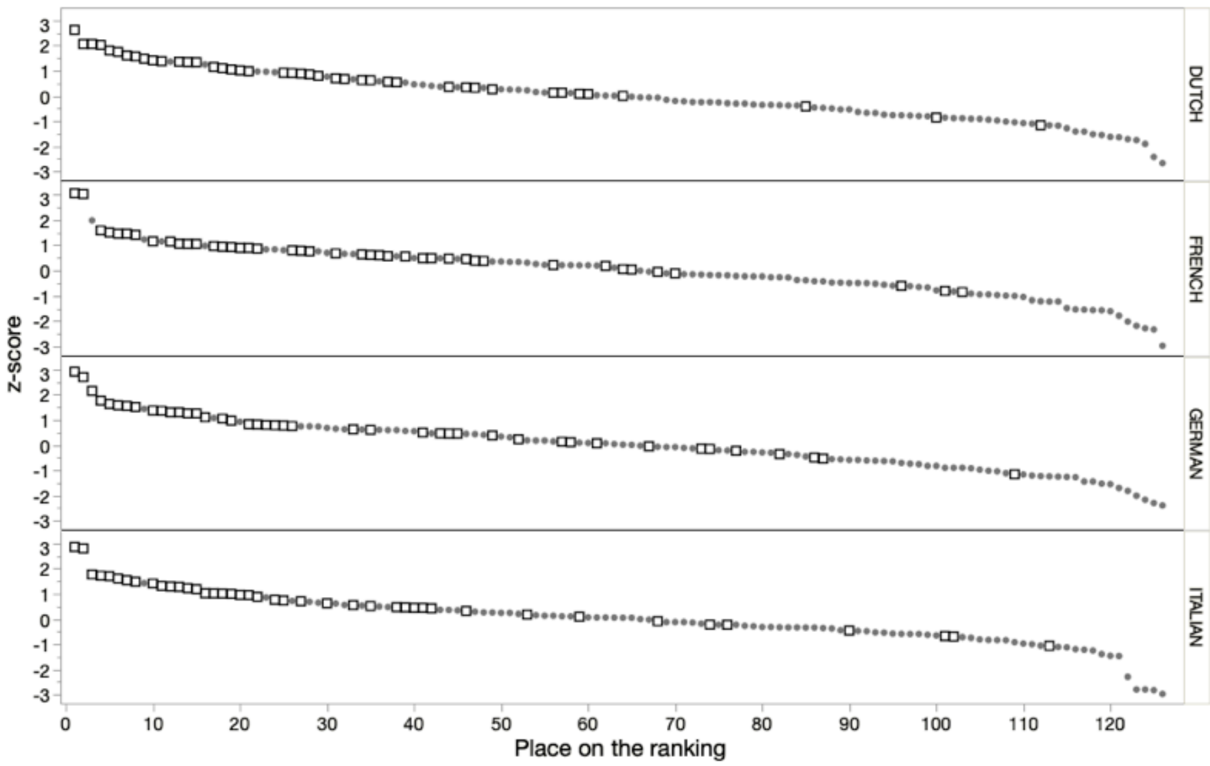

Figure 1: Z-scores representing the place in the ranking for NH and HI children (NH: squares; HI: dots (each symbol represents a stimulus (n = 126))) (estimated values in z-scores (y-axis), position in the ranking as an ordinal score (x-axis))

Place in ranking for children with CI and HA

In the previous analysis, HI children were treated as one group. It was shown that all listener groups perceived a clear difference in the overall speech quality of NH and HI children. But did the listeners also perceive a qualitative difference between the speech of children with CI and HA? In order to answer this question, an additional analysis was performed. The dependent variable, i.e., z-score, was again numerical.

In HI children, various factors may influence their speech, such as the type of device, age at implantation/receiving HA, length of device use, chronological age, (aided) hearing thresholds etc. The best fitting model, displayed in Table 2, consisted of the factors Hearing status, Length of device use, and their interaction. At the intercept, i.e., 85 months of device use, the position in the ranking of children with CI and HA did not differ significantly ($p > 0.05$), meaning that the participants did not perceive a marked qualitative difference in the speech of children with CI and HA. However, the factor Length of device use had a significant effect ($p = 0.0005$): children who had used their device for a longer time were perceived as having better overall speech quality. However, the effect of Length of device use differed significantly for children with CI and children with HA. More specifically, as children's length of device use increased, the difference in speech quality between both hearing statuses enlarged. As visualised in Figure 2, increasing length of device use resulted in higher speech quality for children with CI. For children with HA, this evolution was barely noticeable, which indicated that the speech quality of these children remained fairly constant over time.

Again, adding the factor Listener group to the model did not lead to a better fit. This suggested that the judgements of the native and non-native listeners were comparable. The native language background of the listeners did not significantly influence their appreciation of the overall speech quality of the Dutch speaking HI children. In other words: how listeners perceived the speech quality of CI and HA children and the effect of length of device use was not influenced by their language background.

	Estimate	Std. Error	t-value	p
Intercept	–0.171	0.185	–0.925	0.355
Hearing status [HA]	–0.322	0.248	–1.298	0.194
Length of device use	0.039	0.011	3.494	0.0005
Hearing status [HA] * Length of device use	–0.043	0.015	–2.800	0.005

Table 2: Z-scores representing the place on the ranking for CI and HA children (fixed effects = Hearing status (CI (= intercept) or HA), Length of device use and the interaction of these variables); random effects = individual children and utterances

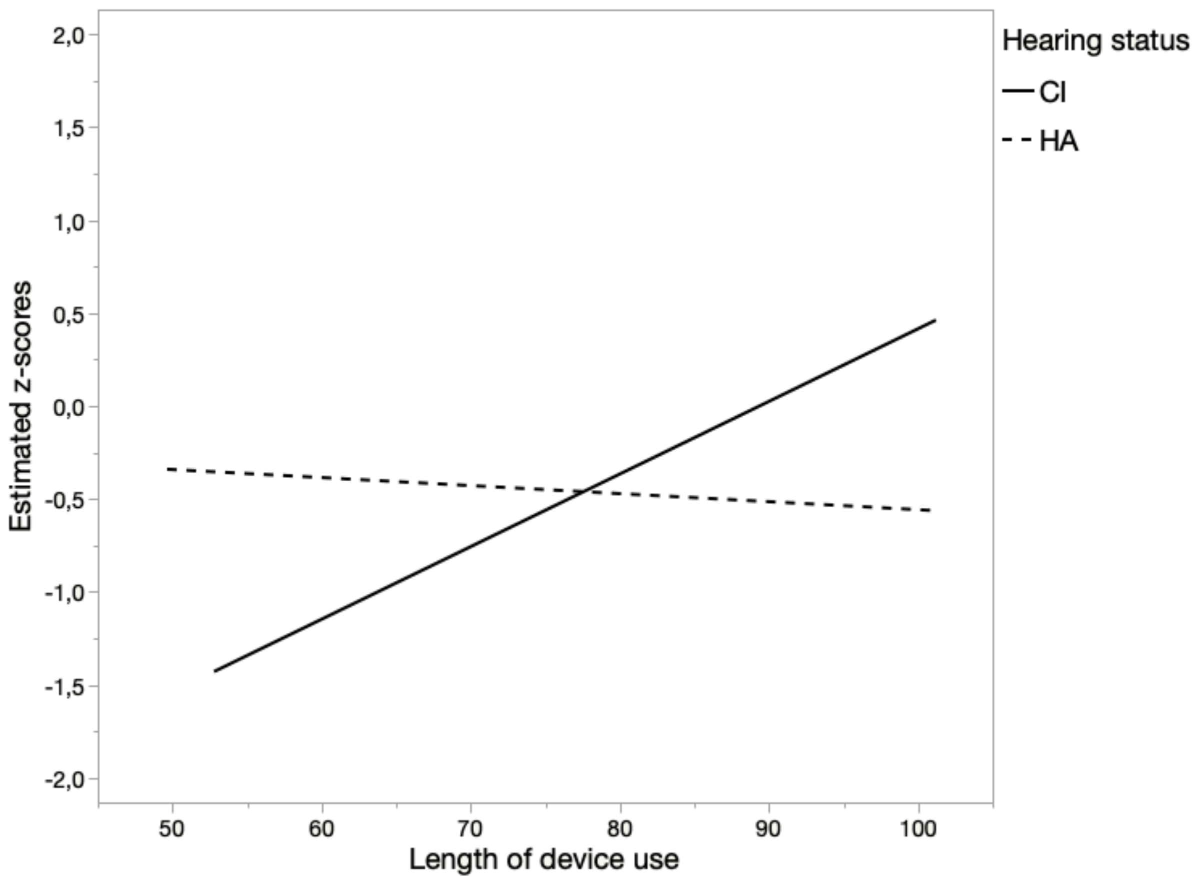

Figure 2: Z-scores representing the place on the ranking for CI and HA children considering their length of device use (estimated values in z-scores, length of device use in months)

Pairwise comparisons

In the previous section, only the final place of each stimulus in the ranking was considered. The final ranking of a stimulus was based on the number of pairs in which it was selected as the better sounding stimulus. This section will have a closer look at these pairs. Of the 5,005 pairs that were judged by the four groups of listeners, the present analysis considers the pairs in which two different hearing statuses were combined. These were either a combination of (1) a CI and a NH child, (2) a HA and a NH child or (3) a CI and a HA child. Thus the question to be addressed for these comparison types is: which hearing status was more likely to be selected as the better sounding one? Moreover, this analysis investigates whether native and non-native listeners reacted differently to specific comparison types.

In this section, all comparison types will be discussed separately, yet the best fitting models consisted of the same effects: Hearing status, Listener group and the interaction between those variables. Separate tables containing the best fitting models for each type of comparison were added as Supplementary materials 1-3. Results are visualised in Figure 3. The dependent variable of this MLM analysis was binomial: (not) being selected as the better sounding stimulus in a pair. The results of the analysis were expressed in logits in the tables, yet discussed in terms of probabilities.

General findings

In contrast to the overall analysis of the children's speech quality, the analysis on the separate pairs showed some subtle, yet significant, differences between the listener groups. Before discussing these results in detail, some general findings for all listener groups are highlighted. For the comparison types NH-CI and NH-HA, all listener groups exhibited a very clear preference for NH children. More specifically, the probability that a NH child was selected as better sounding in the comparison type NH-CI and the comparison type NH-HA was over 70% for all listener groups. This result indicated that for all listener groups, the overall speech quality of NH children was considerably better than that of children with CI or HA. For the comparison type with a stimulus of a child with CI and a child with HA, all listener groups – except the German listeners – had a slight preference (50% to 65%) for children with CI. This suggests that the speech of children with CI, independent of the native language of the listener, was considered better than the speech of children with HA (or in the case of the German listeners: was considered as equally good).

Effect of listener group

Concerning the differences between the listener groups, it was striking that the judgements of the German listeners often differed from the judgements of the other listener groups (see Table 3 and Figure 3). In pairs containing the speech of a NH child and a child with CI, the child with

CI had a significantly higher chance of being selected as the better sounding by German listeners than by Dutch speaking listeners ($p < 0.05$). Conversely, NH children were less often selected as better sounding by German than by Dutch speaking participants ($p < 0.05$).

In pairs of NH and HA children, a similar pattern arose. Children with HA had a significantly higher chance of being selected as the better sounding stimulus by German listeners than by Dutch speaking ($p < 0.0001$) and French speaking listeners ($p < 0.05$). Again, these results had a consequence on the judgements of NH children. These children were preferred significantly less than in Dutch speaking ($p < 0.0001$) and French listeners ($p < 0.05$). In pairs of NH and HA children, the judgements of Italian and German listeners did not differ significantly ($p > 0.05$).

In comparisons of children with CI and HA, German and Dutch judgements again differed significantly. The German's preference for children with CI was less strong than in Dutch speaking participants. In German participants, children with CI were thus significantly less often preferred ($p < 0.005$). Consequently, children with HA were more often preferred ($p < 0.05$).

		Estimate	Std. Error	z-value	p
Comparison type NH-CI	CI_German - CI_Dutch	0.654	0.213	3.070	0.026
	NH_German - NH_Dutch	–0.619	0.214	–2.888	0.047
Comparison type NH-HA	HA_German - HA_French	0.639	0.211	3.035	0.029
	HA_German - HA_Dutch	1.095	0.232	4.722	< 0.0001
	NH_German - NH_French	–0.628	0.211	–2.977	0.035
	NH_German - NH_Dutch	–1.095	0.230	–4.750	< 0.0001
Comparison type CI-HA	CI_German - CI_Dutch	–0.664	0.185	–3.581	0.004
	HA_German - HA_Dutch	0.567	0.178	3.186	0.017

Table 3: Post hoc pairwise comparison analyses with Bonferroni adjustment, only significant results

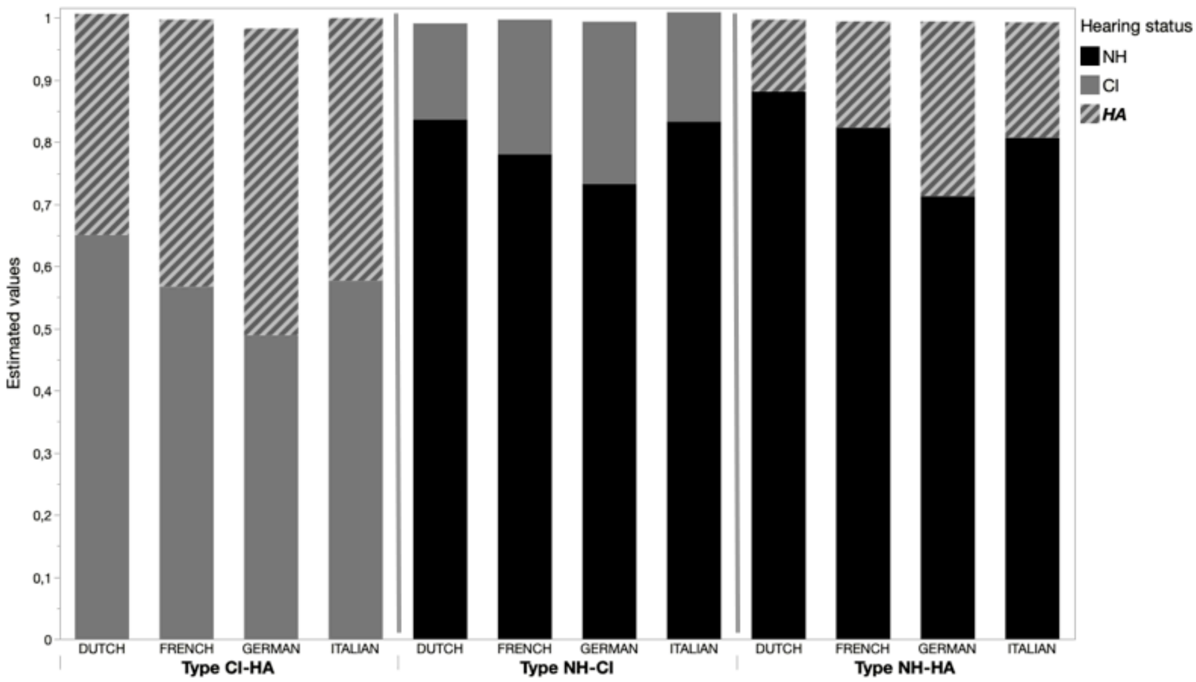

Figure 3: Being selected as the better sounding hearing status for each type of comparison and each listener group (expressed in probabilities)

Discussion

Overall speech quality of NH and HI children

The overall speech quality of NH and HI Dutch speaking children was assessed in a comparative judgement task completed by native and non-native listeners. The aim of this study was to investigate whether listener groups with a different language background (native speakers of Dutch, German, French and Italian) judged speech quality differently from native listeners.

In general, i.e., without taking into account the different listener groups, the results showed a clear preference for the speech of NH children. The listeners attributed higher overall speech quality to the majority of NH children's utterances in comparison to HI children's utterances. This result suggested that, even after several years of device use, something in the speech of HI children was still not on a par with NH children's speech. The result thus seemed to be in line with previous studies of speech development in HI children: these children may not completely catch up with their hearing peers (Fang et al., 2014; Freeman et al., 2017; Tomblin et al., 2015). However, these studies investigated speech intelligibility rather than overall speech quality, and speech quality and intelligibility are obviously not identical. Theoretically, a child may be (nearly) perfectly intelligible while his/her speech quality may still be rather far from that of other children. Further work is required to establish the correlation between these two aspects.

Within the group of HI children, this study showed some differences. Children with CI received higher speech quality scores than children with HA, especially with increasing length of device use. Thus, this study provides new evidence that the effect of length of device use is more prominent in children with CI than in matched children with HA (Lejeune & Demanez, 2006; Tomblin et al., 1999). Children with CI were also overall preferred in the pairwise comparisons. These results are in line with previous studies in which children with CI obtained better scores than children with HA (Baudonck et al., 2010a; Baudonck et al., 2010b; Tomblin et al., 1999).

Native vs. non-native listeners

The main research question of this study was whether native and non-native listeners rated the speech quality of NH and HI children similarly. Interestingly, no significant effect of listener group was found in the position of NH and HI children in the ranking. In other words: the difference in the overall speech quality of NH and HI children was salient for Dutch speaking listeners as well as listeners without knowledge of Dutch. This result is in line with Tang (2009), in which Dutch and English listeners without knowledge of Chinese successfully differentiated between good and poor sound quality of Chinese speech samples. Thus, these results strongly indicated that listeners take into account language independent speech elements in their judgements.

Moreover, the four non-native listener groups did not differ significantly. Since the judgements were comparable for all non-native listener groups, there did not seem to be a transfer of an individual's native language when judging merely the speech of children with a different native language background. This implied that phonological transfer did not seem to play a role in listeners' judgements. However, the question as to which elements contribute to the similarity in the judgements of native and non-native listeners remains. The judgements were not influenced by language, because the non-native listeners did not know Dutch. Incontestably, speech elements were the contributing factor to the judgements, but it is still unknown which speech elements exactly contributed: voice, intonation, speaking rate, pitch, etc.? Further research should be undertaken to investigate the possible correlation between acoustic (spectral and temporal) parameters and the perceived overall speech quality. This would show which acoustically measurable speech characteristics are perceptually salient for (non-)native listeners.

Note: German participants

No statistically significant difference between the listener groups was found in the ranking. Also in the pairwise comparisons, listeners had very similar judgements. They were all more likely to prefer the speech of NH children and in pairs of children with CI and HA, the former were slightly preferred. However, for the pairwise comparisons, there was an effect of listener group. More specifically, in all comparison types, the judgements of the German participants differed from the judgements of the other listener groups. Whereas the other listener groups clearly preferred the speech of NH children, this preference of German listeners was less strong. This observation seemed to suggest that the German listeners had a slightly different approach when judging the speech of NH and HI children. In the formulation of the aims and hypotheses of this study, it was already mentioned that German is closely related to Dutch. Possibly, the close relationship between the two languages explains this result. Not only are the languages closely related, they are also mutually intelligible to a certain extent (Gooskens et al., 2015; Gooskens et al., 2018). Therefore, the German participants possibly used their own linguistic system as a reference point to what they considered "good quality speech". In other words: German listeners were possibly searching for German-like features in the Belgian Dutch children's speech, which did not match German (children's) speech. Thus, we assume that, since German listeners compared the Belgian Dutch children's speech directly to the children's speech of their own native language, their judgements were slightly different than those of the other listener groups.

Limitations and clinical implications

In this study, listeners rated the speech of NH children as qualitatively better than the speech of HI children. More interestingly, this result was obtained by native as well as non-native listeners without any knowledge of the language that was spoken by the children in the stimuli. Since the non-native listeners could not rely on linguistic knowledge, this result strongly indicated that deviations between NH and HI children were purely speech related. Moreover, since the judgements of the native listeners did not differ significantly from those of non-native listeners, this suggests that all listeners primarily considered language unspecific (phonatory and articulatory) elements in their judgements. Therefore, this study emphasizes that speech and language therapy should continue to focus on linguistic as well as speech related aspects.

The fact that judging speech quality appears to be language independent also has practical and clinical implications. First, it offers an alternative approach for perceptual studies on aspects such as speech quality. Usually, these studies only include native speakers of the language that is judged in the samples. In this study, non-native listeners gave reliable judgements that were comparable to the judgements of the native Dutch speaking listeners. This result seemed to suggest that these listeners could thus act as valuable judges in future studies. Including non-native listeners would make it easier to collect a vast number of participants. Secondly, if judging speech quality is language independent, this could have an impact on clinical practices. Often, speech and language therapists treat children whose native language differs from the ambient language. These therapists obviously cannot judge the linguistic skills of these children's native language, but our study suggests that they can make a reliable judgement of their speech quality.

Further work is needed to fully confirm the implication discussed in the previous paragraph. The current study only contained the speech of Dutch speaking children. In order to provide robust evidence, speech samples of French, Italian and German speaking (HI) children should be judged by the same groups of listeners. By doing so, all listeners would judge the speech quality of HI children in their own native language as well as several non-native languages. If this design would show similar results, we could conclude that (1) non-native listeners are able to reliably judge native children's speech quality and (2) non-native children's speech quality can reliably be judged by native listeners.

Moreover, in a future study, it should be investigated whether the participating Belgian Dutch listeners took into account the children's regional variety in their judgements. In other words, if listeners were to judge the children's speech in terms of the degree of regional variety, would this score correlate with their overall perceived speech quality? Previous research showed that Belgian Dutch listeners were particularly sensitive to regional characteristics: they easily identified one's regional background and found their own regional accent the most beautiful (Grondelaers & Lybaert, 2017; Grondelaers et al., 2011; Impe, 2010). A follow-up study should examine whether listeners were actually affected by the regional background of speakers.

Finally, a number of limitations with respect to the participating children and the stimuli need to be considered. Since we controlled for chronological age, gender, geographical background and several hearing related variables, the present study was limited by the small number of participating children. We recommend a replication of this study with a larger number of children while still controlling for these variables. Also, the length of device use in this study was limited to approximately seven years, but the overall speech quality – especially for children with CI – continued to improve. Therefore, it would be interesting to include older children with longer device use in the sample. Concerning the stimuli, it should be noted that the same carrier sentence was used for all children and only the vowel in the nonsense word varied. Hence, the speech samples of the children only contained a limited number of phonemes. Future studies should contain lexically different stimuli that represent the full phonological inventory of each child.

Conclusion

Previous research showed that native listeners' general impression of the overall speech quality of NH and HI children differed. However, these listeners were possibly influenced by their linguistic knowledge. Therefore, the study was replicated with listeners without any knowledge of Dutch: native speakers of Italian, French and German. Since the non-native listeners had no knowledge of the linguistic structure of Dutch, their judgements were purely based on the speech characteristics of the children. The result of the comparative judgement task, i.e., a ranking, showed that the judgements on NH and HI children's overall speech quality did not differ significantly for native and non-native listeners. Both groups indicated that they heard a qualitative difference between NH and HI children's speech. The pairwise comparisons themselves showed similar results: the utterances of NH children were mostly preferred by all listeners. Finally, the fact that the rating of speech quality was language independent suggests that future studies should not exclude non-native listeners in perceptual (language independent) speech experiments and that speech therapists can reliably judge the speech quality of non-native children.

Supplementary materials

	Estimate	Std. Error	z-value	p
Intercept	–1.6962	0.3304	–5.134	< 0.0001
Hearing status [NH]	3.3252	0.4673	7.116	< 0.0001
Listener group [French]	0.4073	0.2208	1.845	0.0651
Listener group [German]	0.6538	0.2130	3.070	0.0021
Listener group [Italian]	0.1538	0.2279	0.675	0.4996
Hearing status [NH] * Listener group [French]	–0.7648	0.3138	–2.438	0.0148
Hearing status [NH] * Listener group [German]	–1.2729	0.3022	–4.212	< 0.0001
Hearing status [NH] * Listener group [Italian]	–0.1775	0.3234	–0.549	0.5831

Supplementary materials 1: Parameter estimates of the MLM model estimating the preference of listeners in the comparison type CI-NH (fixed effects = Hearing status (NH or CI), Listener group (Dutch, French, German or Italian) and the interaction of these two; random effects = individual children)

	Estimate	Std. Error	z-value	p
Intercept	–2.0347	0.2835	–7.178	< 0.0001
Hearing status [NH]	4.0400	0.4003	10.092	< 0.0001
Listener group [French]	0.4562	0.2457	1.856	0.0634
Listener group [German]	1.0955	0.2320	4.722	< 0.0001
Listener group [Italian]	0.5542	0.2378	2.330	0.0198
Hearing status [NH] * Listener group [French]	–0.9228	0.3472	–2.658	0.0079
Hearing status [NH] * Listener group [German]	–2.1901	0.3272	–6.694	< 0.0001
Hearing status [NH] * Listener group [Italian]	–1.1261	0.3356	–3.355	0.0008

Supplementary materials 2: Parameter estimates of the MLM model estimating the preference of listeners in the comparison type NH-HA (fixed effects = Hearing status (NH or HA), Listener group (Dutch, French, German or Italian) and the interaction of these two; random effects = individual children)

	Estimate	Std. Error	z-value	p
Intercept	0.6201	0.3036	2.042	0.0411
Hearing status [HA]	–1.2123	0.4272	–2.838	0.0045
Listener group [French]	–0.3490	0.1891	–1.846	0.0650
Listener group [German]	–0.6639	0.1854	–3.581	0.0003
Listener group [Italian]	–0.3102	0.1907	–1.627	0.1038
Hearing status [HA] * Listener group [French]	0.6580	0.2612	2.519	0.0118
Hearing status [HA] * Listener group [German]	1.2306	0.2569	4.790	< 0.0001
Hearing status [HA] * Listener group [Italian]	0.5901	0.2633	2.241	0.0250

Supplementary materials 3: Parameter estimates of the MLM model estimating the preference of listeners in the comparison type CI-HA (fixed effects = Hearing status (CI or HA), Listener group (Dutch, French, German or Italian) and the interaction of these two; random effects = individual children)

Acknowledgements

This project was funded by a predoctoral research grant of the Research Foundation – Flanders (FWO) to the first author (1100316N). We would like to thank M. Kilani-Schoch, B. Grandon and G. Cassani for their help with the translations of the instructions and/or the recruitment of the participants.

Statement of interest

The authors declare no conflict of interest.

References

Abercrombie, D. (1967). *Elements of general phonetics*. Edinburgh: Edinburgh University Press.

AlSanosi, A., & Hassan, S. M. (2014). The effect of age at cochlear implantation outcomes in Saudi children. *International Journal of Pediatric Otorhinolaryngology,* 78(2), 272-276. https://doi.org/10.1016/j.ijporl.2013.11.021

Baayen, H. (2008). *Analyzing linguistic data. A practical introduction to statistics using R.* Cambridge: Cambridge University Press.

Bakhshaee, M., Ghasemi, M. M., Shakeri, M. T., Razmara, N., Tayarani, H., & Tale, M. R. (2007). Speech development in children after cochlear implantation. *European archives of oto-rhino-laryngology : official journal of the European Federation of Oto-Rhino-Laryngological Societies (EUFOS) : affiliated with the German Society for Oto-Rhino-Laryngology - Head and Neck Surgery,* 264(11), 1263-1266. https://doi.org/10.1007/s00405-007-0358-1

Baudonck, N., Dhooge, I., D'haeseleer, E., & Van Lierde, K. (2010a). A comparison of the consonant production between Dutch children using cochlear implants and children using hearing aids. *International Journal of Pediatric Otorhinolaryngology,* 74(4), 416-421. https://doi.org/10.1016/j.ijporl.2010.01.017

Baudonck, N., Dhooge, I., & Van Lierde, K. (2010b). Intelligibility of hearing impaired children as judged by their parents: A comparison between children using cochlear implants and children using hearing aids. *International Journal of Pediatric Otorhinolaryngology,* 74(11), 1310-1315. https://doi.org/10.1016/j.ijporl.2010.08.011

Bejar, I. I. (2012). Rater cognition: Implications for validity. *Educational Measurement: Issues and Practice,* 31(3), 2-9. https://doi.org/10.1111/j.1745-3992.2012.00238.x

Boonen, N., Kloots, H., & Gillis, S. (2020). Rating the overall speech quality of hearing-impaired children by means of comparative judgements. *Journal of Communication Disorders,* 83, 105969. https://doi.org/10.1016/j.jcomdis.2019.105969

Boonen, N., Kloots, H., Verhoeven, J., & Gillis, S. (2019). Can listeners hear the difference between children with normal hearing and children with a hearing impairment? *Clinical Linguistics & Phonetics,* 33(4), 316-333. https://doi.org/10.1080/02699206.2018.1513564

Bradley, R. A., & Terry, M. E. (1952). Rank analysis of incomplete block designs. *Biometrika,* 39(3-4), 324-345. https://doi.org/10.1093/biomet/39.3-4.324

Chin, S. B., & Kuhns, M. J. (2014). Proximate factors associated with speech intelligibility in children with cochlear implants: A preliminary study. *Clinical Linguistics & Phonetics,* 28(7-8), 532-542. https://doi.org/10.3109/02699206.2014.926997

Ching, T. Y. C., Rattanasone, N. X., Macdonald, G., Zhang, V. W., Button, L., & Demuth, K. (2015). Intelligibility of speech produced by children with hearing loss: Conventional amplification versus nonlinear frequency compression in hearing aids. *Journal of Communication Disorders, Deaf Studies & Hearing Aids,* 3, 135. https://doi.org/10.4172/2375-4427.1000135

Colman, A. M. (2009). *A dictionary of psychology* (3rd ed.). Oxford: Oxford University Press.

Dalby, A. (2006). *Dictionary of languages: The definitive reference to more than 400 languages*. London: A&C Black.

De Raeve, L. (2010). A longitudinal study on auditory perception and speech intelligibility in deaf children implanted younger than 18 months in comparison to those implanted at later ages. *Otology & Neurotology,* 31(8), 1261-1267. https://doi.org/10.1097/MAO.0b013e3181f1cde3

Ellis, L. W., & Beltyukova, S. A. (2008). Effects of training on naive listeners' judgments of the speech intelligibility of children with severe-to-profound hearing loss. *Journal of Speech, Language, and Hearing Research,* 51(5), 1114-1123. https://doi.org/10.1044/1092-4388(2008/06-0217)

Elmentaler, M. (2009). *Deutsch und seine Nachbarn*. Frankfurt am Main: Peter Lang.

Ertmer, D. J. (2007). Speech intelligibility in young cochlear implant recipients: Gains during year three. *The Volta Review,* 107(2), 85-99.

Escudero, P., Benders, T., & Lipski, S. C. (2009). Native, non-native and L2 perceptual cue weighting for Dutch vowels: The case of Dutch, German, and Spanish listeners. *Journal of Phonetics,* 37(4), 452-465. https://doi.org/10.1016/j.wocn.2009.07.006

Escudero, P., Sisinni, B., & Grimaldi, M. (2014). The effect of vowel inventory and acoustic properties in Salento Italian learners of Southern British English vowels. *Journal of the Acoustical Society of America,* 135(3), 1577-1584. https://doi.org/10.1121/1.4864477

Fang, H. Y., Ko, H. C., Wang, N. M., Fang, T. J., Chao, W. C., Tsou, Y. T., & Wu, C. M. (2014). Auditory performance and speech intelligibility of Mandarin-speaking children implanted before age 5. *International Journal of Pediatric Otorhinolaryngology,* 78(5), 799-803. https://doi.org/10.1016/j.ijporl.2014.02.014

Flege, J. E., Bohn, O. S., & Jang, S. (1997). Effects of experience on non-native speakers' production and perception of English vowels. *Journal of Phonetics,* 25(4), 437-470. https://doi.org/10.1006/jpho.1997.0052

Freeman, V., Pisoni, D. B., Kronenberger, W. G., & Castellanos, I. (2017). Speech intelligibility and psychosocial functioning in deaf children and teens with cochlear implants. *Journal of Deaf Studies and Deaf Education,* 22(3), 278-289. https://doi.org/10.1093/deafed/enx001

Gandour, J., Wong, D., Hsieh, L., Weinzapfel, B., Van Lancker, D., & Hutchins, G. D. (2000). A crosslinguistic PET study of tone perception. *Journal of Cognitive Neuroscience,* 12(1), 207-222. 10.1162/089892900561841

Geers, A. E., & Nicholas, J. G. (2013). Enduring advantages of early cochlear implantation for spoken language development. *Journal of Speech, Language, and Hearing Research,* 56(2), 643-655. https://doi.org/10.1044/1092-4388(2012/11-0347)

Gillis, S. (2017). Speech and language in congenitally deaf children with a cochlear implant. In A. Bar-On & D. Ravid (Eds.), *Handbook of communication disorders: Theoretical, empirical, and applied linguistic perspectives* (pp. 763-790). Berlin: Mouton De Gruyter.

Gooskens, C., van Bezooijen, R., & van Heuven, V. J. (2015). Mutual intelligibility of Dutch-German cognates by children: The devil is in the detail. *Linguistics,* 53(2), 255-283. https://doi.org/10.1515/ling-2015-0002

Gooskens, C., van Heuven, V. J., Golubović, J., Schüppert, A., Swarte, F., & Voigt, S. (2018). Mutual intelligibility between closely related languages in Europe. *International Journal of Multilingualism,* 15(2), 169-193. https://doi.org/10.1080/14790718.2017.1350185

Gottfried, T. L., & Beddor, P. S. (1988). Perception of temporal and spectral information in French vowels. *Language and Speech,* 31(1), 57-75. https://doi.org/10.1177/002383098803100103

Grondelaers, S., & Lybaert, C. (2017). Bepaalt wat we denken en voelen over taal ook wat we doen in taal? In G. De Sutter (Ed.), *De vele gezichten van het Nederlands in Vlaanderen* (pp. 163-181). Leuven/Den Haag: Acco.

Grondelaers, S., van Hout, R., & Speelman, D. (2011). A perceptual typology of standard language situations in the Low Countries. In T. Kristiansen & N. Coupland (Eds.), *Standard Languages and Language Standards in a Changing Europe* (pp. 199-222). Oslo: Novus Press.

Hide, Ø. (2013). *Acoustic features of speech by young cochlear implant users. A comparison with normal-hearing and hearing-aided age mates.* (Unpublished doctoral dissertation), University of Antwerp, Antwerp, Belgium.

Holden, K. T., & Nearey, T. M. (1986). A preliminary report on three Russian dialects: Vowel perception and production. *Russian Language Journal,* 40(136/137), 3-21.

Impe, L. (2010). *Mutual intelligibility of national and regional varieties of Dutch in the Low Countries.* (Unpublished doctoral dissertation), KU Leuven, Leuven, Belgium.

Jarvis, S., & Pavlenko, A. (2008). *Crosslinguistic influence in language and cognition*. New York/London: Routledge Taylor and Francis.

Kloiber, D. T., & Ertmer, D. J. (2015). Can children substitute for adult listeners in judging the intelligibility of the speech of children who are deaf or hard of hearing? *Language, Speech, and Hearing Services in Schools,* 46, 56-63. https://doi.org/10.1044/2014_lshss-13-0043

Kondo, K. (2012). *Subjective quality measurement of speech*. Berlin/Heidelberg: Springer.

Korver, A. M., Smith, R. J., Van Camp, G., Schleiss, M. R., Bitner-Glindzicz, M. A., Lustig, L. R., Usami, S. I., & Boudewyns, A. N. (2017). Congenital hearing loss. *Nature Reviews Disease Primers,* 3, 16094. https://doi.org/10.1038/nrdp.2016.94

Lejeune, B., & Demanez, L. (2006). Speech discrimination and intelligibility: Outcome of deaf children fitted with hearing aids or cochlear implants. *B-ENT,* 2(2), 63-68.

Lesterhuis, M., Verhavert, S., Coertjens, L., Donche, V., & De Maeyer, S. (2017). Comparative judgment as a promising alternative to score competences. In E. Cano & G. Ion (Eds.), *Innovative Practices for Higher Education Assessment and Measurement* (pp. 119-138). Hershey: IGI Global.

Loizou, P. C. (2011). Speech quality assessment. In W. Lin, D. Tao, J. Kacprzyk, Z. Li, E. Izquierdo, & H. Wang (Eds.), *Multimedia Analysis, Processing and Communications* (Vol. 346, pp. 623-654). Berlin/Heidelberg: Springer.

Lybaert, C., & Delarue, S. (2017). 'k Spreek ekik ver altijd zo. Over de opmars van tussentaal in Vlaanderen. In G. De Sutter (Ed.), *De vele gezichten van het Nederlands in Vlaanderen* (pp. 142-162). Leuven/Den Haag: Acco.

Montag, J. L., AuBuchon, A. M., Pisoni, D. B., & Kronenberger, W. G. (2014). Speech intelligibility in deaf children after long-term cochlear implant use. *Journal of Speech, Language, and Hearing Research,* 57(6), 2332-2343. https://doi.org/10.1044/2014_JSLHR-H-14-0190

Munson, B., Johnson, J. M., & Edwards, J. (2012). The role of experience in the perception of phonetic detail in children's speech: A comparison between speech-language pathologists and clinically untrained listeners. *American Journal of Speech-Language Pathology,* 21(2), 124-139. https://doi.org/10.1044/1058-0360(2011/11-0009)

Osberger, M. J., & McGarr, N. S. (1982). Speech production characteristics of the hearing impaired. In N. Lass (Ed.), *Speech and Language: Advances in Basic Research and Practice* (Vol. 8, pp. 221-283). New York: Academic Press.

Peng, S.-C., Spencer, L. J., & Tomblin, J. B. (2004). Speech intelligibility of pediatric cochlear implant recipients with 7 years of device experience. *Journal of Speech, Language, and Hearing Research,* 47(6), 1227-1236. https://doi.org/10.1044/1092-4388(2004/092)

R Core Team. (2016). R: A language and environment for statistical computing. Retrieved from www.R-project.org

Schiavetti, N. (1992). Scaling procedures for the measurement of speech intelligibility. In R. D. Kent (Ed.), *Intelligibility in Speech Disorders: Theory, measurement and management* (pp. 11-34). Amsterdam: John Benjamins Publishing Company.

Spencer, L. J., Tye-Murray, N., & Tomblin, J. B. (1998). The production of English inflectional morphology, speech production and listening performance in children with cochlear implants. *Ear & Hearing,* 19(4), 310-318. https://doi.org/10.1097/00003446-199808000-00006

Tang, C. (2009). *Mutual intelligibility of Chinese dialects.* (Doctoral dissertation), University of Leiden, Leiden, The Netherlands.

Tobey, E. A., Rekart, D., Buckley, K., & Geers, A. E. (2004). Mode of communication and classroom placement impact on speech intelligibility. *Archives of Otolaryngology - Head and Neck Surgery,* 130(5), 639-643. https://doi.org/10.1001/archotol.130.5.639

Tomblin, J. B., Harrison, M., Ambrose, S. E., Walker, E. A., Oleson, J. J., & Moeller, M. P. (2015). Language outcomes in young children with mild to severe hearing loss. *Ear and Hearing,* 36, 76S-91S. https://doi.org/10.1097/aud.0000000000000219

Tomblin, J. B., Spencer, L., Flock, S., Tyler, R., & Gantz, B. (1999). A comparison of language achievement in children with cochlear implants and children using hearing aids. *Journal of Speech, Language, and Hearing Research,* 42(2), 497-509. https://doi.org/10.1044/jslhr.4202.497

van Heuven, V. J. (2008). Making sense of strange sounds: (Mutual) intelligibility of related language varieties. A review. *International Journal of Humanities and Arts Computing,* 2(1-2), 39-62. https://doi.org/10.3366/e1753854809000305

Van Lierde, K. M., Vinck, B. M., Baudonck, N., De Vel, E., & Dhooge, I. (2005). Comparison of the overall intelligibility, articulation, resonance, and voice characteristics between children using cochlear implants and those using bilateral hearing aids: A pilot study. *International Journal of Audiology,* 44(8), 452-465. https://doi.org/10.1080/14992020500189146

Verhavert, S. (2018). *Beyond a mere rank order: The method, the reliability and the efficiency of comparative judgment.* (Unpublished doctoral dissertation), University of Antwerp, Antwerp, Belgium.

Verhavert, S., De Maeyer, S., Donche, V., & Coertjens, L. (2018). Scale separation reliability: What does it mean in the context of comparative judgment? *Applied Psychological Measurement,* 42(6), 428-445. https://doi.org/10.1177/0146621617748321

Verhoeven, J., Hide, Ø., De Maeyer, S., Gillis, S., & Gillis, S. (2016). Hearing impairment and vowel production. A comparison between normally hearing, hearing-aided and cochlear implanted Dutch children. *Journal of Communication Disorders,* 59, 24-39. https://doi.org/10.1016/j.jcomdis.2015.10.007

Wang, H., & van Heuven, V. J. (2004). Cross-linguistic confusion of vowels produced and perceived by Chinese, Dutch and American speakers of English. *Linguistics in the Netherlands,* 21, 205-216. https://doi.org/10.1075/avt.21.22wan

Williams, D., & Escudero, P. (2014). Native and non-native speech perception. *Acoustics Australia,* 42(2), 79-83.

Yoshinaga-Itano, C., Baca, R. L., & Sedey, A. L. (2010). Describing the trajectory of language development in the presence of severe-to-profound hearing loss: A closer look at children with cochlear implants versus hearing aids. *Otology & Neurotology,* 31(8), 1268-1274. https://doi.org/10.1097/MAO.0b013e3181f1ce07