HOW CAN SPEECH PRODUCTION SKILLS BE PREDICTED FROM VISUAL, AUDITORY, AND HAPTIC PERCEPTION SKILLS?

Caroline Gluth, Philip Hoole

Institute of Phonetics and Speech Processing, Ludwig Maximilian University of Munich
cgluth@phonetik.uni-muenchen.de, hoole@phonetik.uni-muenchen.de

ABSTRACT

How does speech production relate to speech perception? To investigate the most suitable perceptual factors to predict the production skills of a speaker, we determined the visual acuity for speech movements, the auditory acuity for speech sounds, and the haptic acuity of speech organs, following equivalent protocols in all three modalities. These abilities were linked to the individual's distinctness of articulation. We tested 26 German cochlear implant wearers and a normal hearing control group of matching age and gender.

Our data suggest that in normal hearers, visual, auditory, and haptic skills are suitable to predict the production performance. However, this result could not be confirmed for cochlear implant wearers. Compared with the control group, cochlear implant wearers produced significantly less distinct sibilants s and ʃ, but some of them performed surprisingly well in the test for auditory acuity for sibilants.

Keywords: Acuity, multimodal perception, production, cochlear implant, German

1. INTRODUCTION

The widely used DIVA model [4] suggests that auditory and somatosensory feedback loops play a crucial role in speech production. The validity of the theory has been demonstrated, e.g. in [3], by testing young adults' auditory and haptic abilities and linking them to skills in speech production.

Our first aim is to additionally investigate the influence of the perceptive ability to visually distinguish utterances of spoken language, based on the assumption that individuals can benefit from this ability in their speech production, not by feedback loops but by observation and imitation. To make these abilities accessible for quantitative analyses we constructed a quasi-continuum of visual stimuli by video morphing and applied it within an established procedure [3].

Additionally, by including cochlear implant wearers with different hearing abilities at speech acquisition into our study, we were hoping to be able to investigate the influence of different stages of perturbation of the auditory skill on the visual performance.

2. CONTINUA FOR PERCEPTION TASKS

For the investigation of the acuity for visual recognition of speech movements we used video morphing [7] to create a continuum between temporal high resolution (540 Hz) video recordings of two spoken utterances that only differed in the degree of liprounding of the target vowel. We recorded the nonsense words /baˈdiːdə/ and /baˈdyːdə/. The videos were split into single frames, stabilized, morphed and then resynthesized to obtain a quasi-continuum of 200 videos.

To measure the auditory acuity for speech perception we used both a sibilant and a vowel continuum. An s-ʃ-continuum was synthesized by [1], from the tokens 'Asse' [ˈasə] and 'Asche' [ˈaʃə]. Additionally, we generated an iː-yː-continuum with [5] from the tokens 'Beagle' [ˈbiːgl̩] and 'Bügel' [ˈbyːgl̩].

The haptic acuity of speech organs was measured similarly as in [9] by manual application of JVP Domes with a specially designed applicator to ensure a contact pressure of approximately 0.5 N for about 0.5 s. The domes had 16 grating spacings from 5.0 down to 0.2 mm.

3. TEST PROCEDURES

In the visual and auditory acuity experiments, tokens from the quasi-continua were presented to the participants in computer-based 4-interval 2-alternative forced choice adaptive staircase discrimination tasks [3]. If the participant succeded in the discrimination, the distance between the tokens was decreased, if they failed, the distance was increased, and this was repeated until we expected a steady-state to be reached. To determine the individual categorial boundaries and to avoid categorial effects, a label task was carried out before each experiment, so that the presented pairs of tokens could be chosen

equidistantly from the determined boundary. In case a subject failed to label the tokens at all, an average value obtained from preliminary experiments was used as a substitute.

For the haptic acuities, we followed a 4-alternative forced choice protocol. The grated domes were applied manually onto the participants' speech organs labially and apically. Depending on whether the participant correctly identified the orientation (vertical, horizontal, diagonal-rising, or diagonal-falling), the trial was repeated with a smaller or wider grating. The maximum number of trials was 26. During the test, the lower half of the participants' field of view was restricted with blurred goggles to prevent visual recognition of the dome's orientation.

For all perception tests, we recorded the values of the distances between tokens or sizes of gratings, as a sequence. Using the method of least squares, we determined the most stable segment from each sequence and associated its root-mean-square with the just-noticeable difference of the respective perceptive pathway [3].

To determine the distinctness of articulation, the participants read the words 'Tasse' /tasə/ and 'Tasche' /taʃə/ within a carrier sentence. Each sentence was repeated ten times and embedded randomly in a 20 minute reading task. The speech material was recorded in an anechoic chamber, using the SpeechRecorder software [2]. We determined the first moments of the DCT-smoothed spectra of the sibilants s and ʃ and used their difference as a measure for the distinctness of articulation.

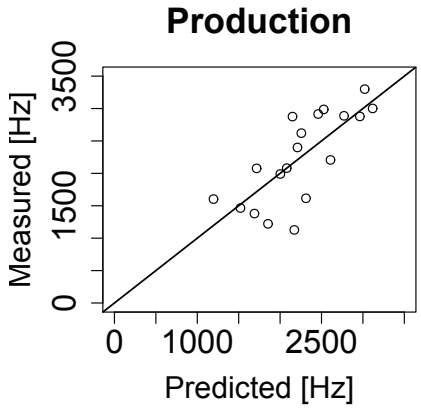
4. PARTICIPANTS

All participants of the study were speakers of German, who had not regularly used sign language and who were mainly raised and living in Southern Germany. We were able to recruit 26 cochlear implant wearers from 8y to 84y, 8 of which were male and 18 were female. We then selected a control group of 26 normal hearing subjects of matching gender and corresponding age, one for each CI wearer. The age differences were less than 1;6y for adults and less than 0;9y for minors. Depending on the hearing abilities during the age of speech acquisition we divided the cochlear implant wearers into four groups. The first group (CI, n = 3) was implanted within the first year of age and underwent relatively normal speech acquisition with the aid of a CI. The second group (LT, n = 7) also gained full speech abilities with the aid of a CI, but had a delayed phase of speech acquisition, in that none of them began to speak before 2;6y. The third group (HL, n = 6) was hard of hearing with profound hearing loss during the full phase of speech acquisition and was implanted later than 3;0y. The fourth group (NH, n = 10) underwent a normal process of speech acquisition with none, mild or one-sided impairment, with deafening and implantation later than 3;0y. The control group (CG) reported no known hearing impairments. Participants who reported other limitations of the tested senses, except from common vision aids, were excluded from the study.

5. RESULTS

A multiple linear regression model calculated for the normal hearers revealed a significant relation between the visual, haptic (labial) and auditory (iː-yː) factors and the distinctness of production. Surprisingly, the auditory acuity (s-ʃ) and the haptic acuity (apical) did not contribute significantly to the model.

Figure 1: Linear Model: Predicted and measured distinctness of articulation for normal hearers, predicted by visual acuity, auditory acuity (iː-yː) and haptic acuity (labial). (n = 23)



The three youngest (8y, 9y) and oldest (84y) normal hearing participants performed strikingly badly in the computer-based tests, which we assumed was because of unfamiliarity with the procedures. We therefore did not include them in this evalutation. The predictors and their significance levels can be found in table 1.

Cochlear implant wearers usually perceive frequencies up to 8000 Hz. Given that the relevant distinctive frequencies for the tested vowels are located more to the center of their audible spectrum than the frequencies relevant for the tested sibilants, we were surprised that we were able to find a stable segment of the measured staircase sequence in the iː-

Table 1: Significance levels for Linear Model, Mult. R^2 = 0.57, F[3,15] = 6.50, p = 0.0049 (**)

Predictor	Significance level
Visual acuity	p = 0.01538 (*)
Auditory acuity (iː-yː)	p = 0.00650 (**)
Haptic acuity (labial)	p= 0.00878 (**)

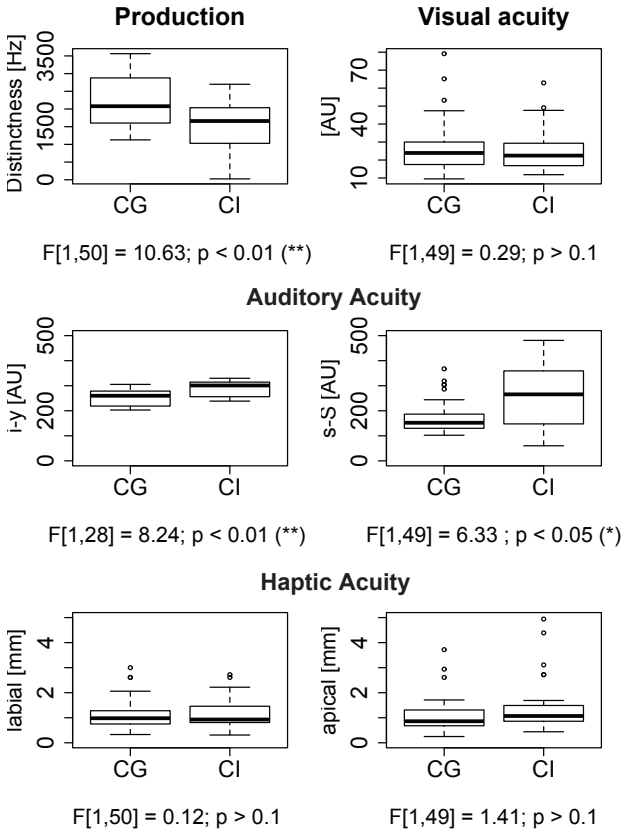
yː-perception for only ten out of 26 cochlear implant wearers. In the s-ʃ-perception, a stable discrimination threshold could be determined for all but one cochlear implant wearer.

Due to the large proportion of missing data for the auditory acuity (iː-yː), the model illustrated in figure 1 was unfortunately not directly applicable for cochlear implant wearers, however a comparison of two models accounting for visual, auditory (s-ʃ) and haptic (apical) perception in cochlear implant wearers and the control group shows that in cochlear implant wearers neither of these three perceptual measures predict the production skills in a significant way.

We also found that the hearing status during speech acquisition (i.e. the key feature by which we distinguish our four groups) did not contribute significantly to the Linear Model to predict the distinctness of articulation. We also tested the influence of the group distribution on the perception skills. Particularly for visual perception we had expected to find the groups with hearing impairment during the crucial phase of speech aquisition (LT and HL) to perform notably well. However, this assumption could not be confirmed.

As illustrated in figure 2, cochlear implant wearers produced significantly less distinct sibilants than the control group (top left panel). Their perceptual performance is comparable to the control subjects in the visual and haptic modalities. In the tests of auditory acuity, cochlear implant wearers do show overall significantly poorer s-ʃ-discrimination than controls, but they also show a striking amount of variability with quite a few subjects falling well within the normal range (indeed even better than the mean control value). For the iː-yː-perception, the poorer performance of the cochlear implant wearers is statistically more robust than for s-ʃ; in fact, the differences between cochlear implant wearers and control group can be assumed to be considerably larger than shown in the middle left panel of figure 2 in view of the large number of missing values mentioned above for the implantees.

Figure 2: Production and perception skills of control group [CG] and cochlear implant wearers [CI]. For the acuity measures, higher values correspond to poorer acuity. The arbitrary units [AU] refer to the steps of our continua and represent just-distinguishable differences in frequency for the auditory acuities and differences in the lip-surrounded area for visual acuity, respectively. One step corresponds to 3.4 Hz spectral centre-of-gravity difference for s-ʃ, 1.4 Hz front cavity resonance difference for iː-yː and 30.8 px lip-area difference (relative to a head size of approx. 700×1000 px) for the video continuum.



6. DISCUSSION

We found that for hearers of ages from 12y to 69y the visual acuity for speech movements, together with the auditory acuity for vowels, and the haptic acuity at the lip are good predictors for the distinctness of articulation. This indicates that speakers benefit not only from somatosensory and auditory feedback loops, but also from their ability to visually observe and imitate others when building up representations of stable production targets (see also [6] for evidence from blind speakers).

For our setup, we could not reproduce the relation

that the acuity of sibilant perception together with the haptic acuity was a predictor for distinctness of sibilant production, which had been shown in [3]. We found this relation neither for cochlear implant wearers, nor for the control group. Additionally, it had not been confirmed in a preliminary experiment carried out with elderly participants [8]. The main difference between these studies was that [3] tested mainly young adult subjects, so we conclude that for mixed age or elderly participant groups the linear relation must have become masked by non-linear aging effects which need to be elucidated in further experiments.

Analogous experiments to investigate the predictability of iː-yː-production are planned to shed more light on the links between vowel production and perception and sibilant production and perception.

The most immediate task is to refine the analysis of the auditory acuity experiments. The results to date were surprising from two points of view: in the control speakers vowel discrimination proved to be a better predictor of sibilant production ability than did sibilant discrimination itself. Moreover, sibilant discrimination ability did not distinguish controls and cochlear implant wearers very clearly, despite clear differences in sibilant contrast in production. On the other hand, the fact that many cochlear implant wearers were essentially untestable with the vowel discrimination test indicates that they do, as expected, have impaired auditory abilities. Nevertheless, for both the vowel and sibilant perception tasks we also have extensive categorization (labelling) data from the preliminary phase of the experiments. Accordingly, we intend to supplement the discrimination scores with measures based, for example, on the sharpness of category boundaries, and thus hopefully obtain performance measures that are both more sensitive and less affected by ceiling effects.

Once the perceptual measures have been refined we will then focus more closely on individual characteristics of the cochlear implant wearers. For example, for cochlear implant wearers showing comparable auditory acuity but divergent production skills, it is of considerable interest to understand better to what extent these differences can be explained by differences in visual and haptic acuity, and/or by differences in sensory status during speech acquisition.

7. ACKNOWLEDGEMENTS

This research was funded by DFG HO 3271/5-1 to Philip Hoole. We thank Katharina Schmidt and Julia Utter for their help conducting the experiments and segmenting the data. We also thank Christian Kroos for providing the video material.

8. REFERENCES

[1] Brunner, J., Ghosh, S., Hoole, P., Matthies, M., Tiede, M., Perkell, J. 2011. The influence of auditory acuity on acoustic variability and the use of motor equivalence during adaptation to a perturbation. *Journal of Speech, Language, and Hearing Research* 54, 727–739.

[2] Draxler, C., Jänsch, K. 2004. SpeechRecorder – a universal platform independent multi-channel audio recording software. *Proceedings of the IV. International Conference on Language Resources and Evaluation* 559–562.

[3] Ghosh, S., Matthies, M., Maas, E., Hanson, A., Tiede, M., Ménard, L., Guenther, F., Lane, H., Perkell, J. 2010. An investigation of the relation between sibilant production and somatosensory and auditory acuity. *Journal of the Acoustical Society of America* 128(5), 3079–3087.

[4] Guenther, F. 2006. Cortical interactions underlying the production of speech sounds. *Journal of Communication Disorders* 39(5), 350–365.

[5] Kawahara, H., Morise, M., Takahashi, T., Nisimura, R., Irino, T., Banno, H. 2008. Tandem-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation. *Proceedings of the ICASSP 2008, Las Vegas* 3933–3936.

[6] Ménard, L., Dupont, S., Baum, S., Aubin, J. 2009. Production and perception of French vowels by congenitally blind adults and sighted adults. *Journal of the Acoustical Society of America* 126(3), 1406–14.

[7] Mullens, S. 2008. Image morphing. http://www.stephenmullens.co.uk/imagemorphing/. online.

[8] Thoma, T. 2014. Eine Untersuchung der Beziehung zwischen Produktionsgenauigkeit und somatosensorischer und akustischer Diskriminationsfähigkeit. Master's thesis at the Institute of Phonetics and Speech Processing, Ludwig Maximilian University of Munich.

[9] Wohlert, A., Smith, A. 1998. Spatiotemporal stability of lip movements in older adult speakers. *Journal of Speech, Language, and Hearing Research* 41(1), 41–50.