Vowel production in congenitally blind and sighted Australian English speakers

Milica Janić[1], Teja Rebernik[1], Pauline Veenstra[1], Emma Wissink[1], Martijn Wieling[1,2], Michael Proctor[3]

[1]*University of Groningen, the Netherlands*
[2]*Haskins Laboratories, the Netherlands*
[3]*Macquarie University, Australia*

m.janic@student.rug.nl, t.rebernik@rug.nl, j.p.veenstra@rug.nl, e.m.j.wissink@student.rug.nl, m.b.wieling@rug.nl, michael.proctor@mq.edu.au

Abstract

The influence of visual deprivation on speech production has not been studied extensively. Previous research, which focused on French, found that sighted speakers produce vowels that are more acoustically dispersed than blind speakers. However, these results have not yet been replicated in another language. The goal of our study was to investigate how the absence of visual feedback impacts vowel production in 10 congenitally blind and 10 sighted Australian English speakers. The blind speakers in this study produced vowels that are clustered closer together than those produced by sighted speakers when measured as a vowel space area; however, other acoustic measures showed varying results. We additionally compared manual and automatic formant extraction methods and found that automatic methods showed similar patterns to those obtained by manually extracting formants.

Keywords: vowel production, vowel acoustics, blind speech, Australian English, formant estimation

1. Introduction

Visual cues provided by articulators such as the lips, face and jaw can play an important role in speech perception (McGurk & MacDonald, 1976; Massaro *et al.*, 1998). In infants, perception of visual cues has been shown to affect production, with expressive language scores at 24 months being correlated with how much the infants look at their mothers' mouths at 6 months (Young *et al.*, 2009). Venezia and colleagues (2016) argue that exposure to visual cues during speech acquisition establishes neural circuitry linking visual gestures to the speech motor system. This implies that the ability to perceive visual information in infanthood may affect speech production in adulthood. Acoustic perceptual ability has already been linked to speech production. Adult cochlear implant recipients who have diminished auditory perception produce less acoustically dispersed vowels than hearing speakers (e.g., Lane *et al*., 2001). Deprivation of visual input in infants and children has also been shown to be linked to phonological disorders and developmental delays (Elstner, 1983), demonstrating the multimodal nature of links between perception and production in typical phonological development.

It is unclear how congenital visual deprivation affects adult speech production. Congenitally blind speakers have never had access to visual information, so it is expected that their speech would differ from that of sighted speakers in some ways if visual cues play a role in speech production and phonological development. However, studies on blind speech have found varying results.

A series of studies done on speakers of Canadian French (Ménard *et al.*, 2009; Ménard *et al*., 2013; Ménard *et al.*, 2014; Ménard *et al.*, 2016; Ménard *et al.*, 2017) have all found that congenitally blind speakers produce vowels that are distributed closer together in the vowel space than sighted speakers, reflected in smaller Euclidean distances (EDs) between vowel pairs and smaller average vowel spacing (AVS). Some studies (e.g., Ménard *et al.*, 2013; Ménard *et al.*, 2016) also examined within-category dispersion and found higher dispersion values for blind speakers, indicating less accuracy in producing a given vowel category. These studies also found articulatory differences between the two groups, with blind speakers having smaller lip protrusion and compensating (sub-optimally) for this through larger tongue movements.

A study of blind and sighted speakers of Dutch (Veenstra *et al.*, 2018) found the opposite pattern of results. EDs were calculated between vowel pairs differing in key phonological features. Blind speakers were found to produce greater acoustic differentiation of each vowel contrast feature compared to sighted speakers, unlike previous findings for French. Veenstra and colleagues also concluded that blind speakers were able to produce vowels using the same articulatory strategies as sighted speakers. One possible explanation for this difference is that effects may be language-specific, depending on the phonological organization of vowel spaces and their patterns of phonetic implementation. Consequently, the differences found for French speakers have yet to be replicated in another language.

Complicating the issue, another study of French also found results that differ to those found by Ménard and colleagues. Turgeon *et al.* (2020) found no significant acoustic differences between blind and sighted speakers of Canadian French in speech production. However, the blind speech was characterized by less lip protrusion. These findings suggest that blind speakers in this study were able to compensate for the effects of visual deprivation and produce vowels that acoustically were not significantly different from those of sighted speakers.

Differences in findings between these studies might also arise from methodological differences. Studies differ in vowel contexts, formant frequency value scales, the number of formants extracted for analysis, vowel contrasts compared, and the methods used for calculating vowel spacing and contrast distances. Another possible issue concerns the small number of subjects in each study (between 9 and 14 for blind speakers). According to Elstner (1983), it is difficult to study homogeneous populations of blind speakers because uncontrolled variables, such as additional motor control or language disorders, might be responsible for the difference between sighted and blind speakers.

For all these reasons, it is difficult to determine the effect of visual deprivation on speech production in adults. The aim of our study was to investigate this question in a language not previously studied in congenitally blind speech – Australian

English – and to compare several acoustic measures to ensure maximum comparability with previous studies. We expected to find a difference between blind and sighted speakers; specifically: blind speakers were predicted to produce vowels spaced closer together in the acoustic vowel space, consistent with previous studies of French (e.g., Ménard *et al.*, 2009).

Because manual estimation of vowel formants is very time-consuming, we additionally aimed to compare the effectiveness of manual and automatic formant extraction in detecting any differences between blind and sighted speakers. The small sample size usually found in studies of blind speakers means that it is important to ensure the data analysis is as accurate as possible. All of the reviewed studies on blind speech used some variation of automatic tracking to extract formant values, yet especially in cases when F0 is high or F1 is low, vowels may be incorrectly tracked using fully automatic methods (Vallabha & Tuller, 2002).. We expected to find that the difference between blind and sighted speakers would be more pronounced in manually extracted data compared to automatic, since automatic extraction is less precise and more error-prone, and thus may obscure smaller differences between the groups.

2. Methods

2.1. Participants

Participants included 10 congenitally blind (4 male, 6 female) and 10 sighted (4 male, 6 female) native speakers of Australian English born in Australia. All blind speakers were legally blind at the time of testing and had never been able to see more than shapes or light. Some were blind from birth while others became blind in their first year. All sighted speakers had normal or corrected-to-normal vision. Sighted and blind participants were roughly age matched. Blind speakers' ages ranged from 30 to 65 years (M = 47.3) while sighted participants were between 27 and 65 years old (M = 41.5). No speech disorders were reported by any of the participants. Slight hearing problems in one ear were reported by one sighted participant.

2.2. Speech Materials

We elicited 12 Australian English monophthongs and 6 diphthongs in three stressed contexts within a carrier sentence: word-initially, in a hVd context, and in isolation (e.g., *"**E**ven as in **h**eed as in **ee**"* for vowel /iː/). A subset of the vowels is analyzed in this study: 11 monophthongs, including 5 long vowels /iː/, /ʉː/, /ɜː/, /oː/, /ɐː/, and 6 short vowels /ɪ/, /ʊ/, /ɔ/, /e/, /æ/, /ɐ/. Vowel length is a distinguishing feature in Australian English – a non-rhotic variety – as some vowel pairs differ primarily in length while sharing other spectral properties, e.g., /ɐ/ *'cut'* and /ɐː/ *'cart'* (Cox & Palethorpe, 2007).

2.3. Procedure

Participants were recorded in the lab, at their workplace, or at their home, depending on their preferences and circumstances. Stimulus sentences were presented twice in random order, either on a laptop screen for sighted speakers, or Braille cue cards for blind speakers, to elicit six repetitions of each vowel per speaker (two per context). This produced an experimental corpus of 66 tokens of each target vowel per participant.

Reading of the experimental materials was self-paced. If a participant had issues with the pronunciation of any items, this was resolved by instructing them to pay attention to the sound in the first syllable of the first target word and repeat it, or by spelling a rhyming word as an example. A Shure WH20 headset microphone was placed as close to the participant's lips as possible without touching them and connected to a Roland QuadCapture external soundcard with an XLR cable. The external soundcard was connected via USB to a Lenovo T340 laptop. Speech Recorder software was used for acquiring the acoustic recordings.

2.4. Acoustic analysis

Acoustic recordings were manually checked and pre-processed in Praat (Boersma & Weenink, 2016) and the target vowels from the elicited tokens were manually segmented and labelled. The first three formant frequencies (F1, F2, F3) were extracted for each vowel, both manually and automatically, to allow for comparison of the two methods of formant estimation. Only monophthongs are considered in the current analysis.

The automatically determined formant tracks (generated by Praat under standard settings) were used as a basis for manual extraction. A visual examination of the spectrogram with superimposed formant tracks served to determine if formant points were placed correctly. If the automatically generated formant tracks were inconsistent with spectral details, formant settings were adjusted for each token until a satisfactory looking formant track was produced. A trial-and-error approach was used to find the settings (maximum frequency and number of formants) which resulted in the best alignment between the formant tracks and spectral resonances.

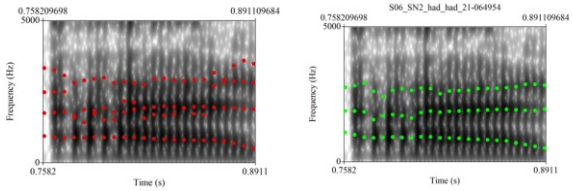

Figure 1: *Example of token /æ/ before (left) and after (right) formant settings adjustment (4200Hz, 3 formants)*

An approximate midpoint was located as the point in time at which to estimate formant frequencies, unless formants were more stable at another point in the vocalic interval, avoiding the peripheries. For each vowel, three formant frequencies were extracted using Praat keyboard shortcuts F1, F2, and F3. An example of an incorrectly tracked token and corrected formant tracks after adjusting the settings can be seen in **Figure 1**.

The automatic formant extraction procedure worked as follows. Formants were extracted automatically at temporal midpoint using a Praat formant value collection script (Lennes, 2013). Standard settings were used for female speakers. For male speakers, the settings were adjusted (5000Hz max, 5 formants, window length 0.025, dynamic range 30dB) to accommodate the characteristically lower formant frequencies.

2.5. Acoustic measurements

Prior to performing acoustic measurements, the formant frequencies were transformed from Hertz (Hz) to mel scale in R using the *phonR* package (McCloy, 2016). The mel scale is a perceptual scale of pitch that approximates the ear's integration of frequency (Menard *et al.*, 2013), and was used in a majority of studies on blind speech. Each acoustic measure was calculated separately for manual and automatic formant measurements.

Euclidean distances (EDs) between vowels differing in both place of articulation and rounding (/iː/ vs / ʉː/, /ɪ/ vs / ʊ/, /ɔ/ vs /e/, and /iː/ vs /oː/), and place of articulation only (/ʉː/ vs /oː/, and /æ/ vs /ɐ/) were calculated with the formula shown in (1). Australian English has no vowel pairs differing only in rounding, unlike French and Dutch, and backness correlates

with roundedness, which makes rounding an enhancing feature rather than a distinguishing one.

$$ED = \sqrt{(F_1 - F_1)^2 + (F_2 - F_2)^2 + (F_3 - F_3)^2} \quad (1)$$

Average vowel spacing (AVS) was calculated per speaker per context, as the mean of all EDs between all possible vowel pairs in the vowel space (55 vowel pairs for 11 vowels). AVS was calculated using both F1×F2 and F1×F2×F3 configurations (two-dimensional vs. three-dimensional). If F3 as an acoustic correlate of rounding plays a role in Australian English, differences between these two metrics may arise. Three-dimensional AVS was the main measure of AVS.

Pentagonal vowel space areas (pVSA) were calculated per speaker per context using the *phonR* R package (McCloy, 2016). The corner vowels were defined as /iː/, /æ/, /ɐː/, /oː/, /ʊ/ by visual inspection of vowel plots for each speaker and speaker group. Quadrilateral VSAs (qVSA) were also calculated with corner vowels, /iː/, /æ/, /ɐː/, /oː/ in order to compare the two VSA configurations. pVSA was the main measure of VSA. To the best of our knowledge, no other study on blind speech calculated VSA so it may be useful to compare this measure to AVS.

A measure of within-category dispersion was calculated according to Lane *et al.* (2001), as the EDs from the positions of each vowel token in the F1×F2×F3 three-dimensional vowel space to the mean position of all tokens of that vowel for each speaker separately. These distances were then averaged across repetitions to get a single dispersion value per vowel per speaker. A larger dispersion value reflects a less precise production of that vowel phoneme.

2.6. Statistical analysis

We fitted linear mixed effects models using the *lme4* R package (Bates *et al.*, 2015) for each dependent variable (VSA, AVS, EDs, dispersion) separately. We compared manual and automatic formant extraction methods by fitting a model on the combined data with type as fixed effect. Models were fitted in line with hypothesis testing, followed by an exploratory analysis of interactions and additional fixed effects, and using model comparison to determine the best model. Visual inspection of residual plots was used to confirm the absence of any obvious deviation from homoscedasticity or normality.

3. Results

3.1. Vowel Space Area – VSA

Due to small number of observations for VSA, our main focus is effect size, rather than significance. We observed a medium non-significant effect of group on pVSA (d = 0.55, p = 0.27), indicating that the congenitally blind speakers have a smaller VSA than the sighted speakers. As expected, there was a large effect of gender (d = −1.15, p = 0.03; smaller VSA for male speakers). The isolated vowel context had significantly larger pVSA values than either hVd or word initial contexts (d = 1.6, p < 0.001). Both qVSA and automatic pVSA showed the same patterns. The vowel space areas per group can be seen in **Figure 2.**

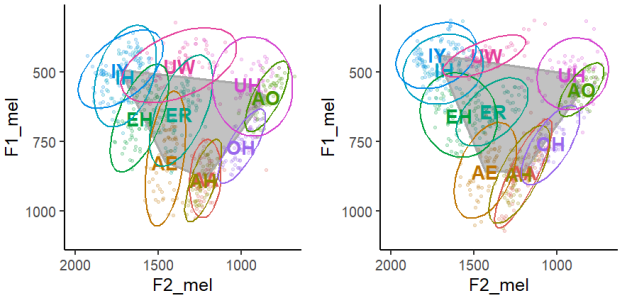

***Figure 2**: pVSA for blind (left) and sighted (right) speakers (dispersion ellipses at 95% CI)*

3.2. Average Vowel Spacing – AVS

We observed an interaction between group and context for AVS. Blind speakers' AVS was significantly higher in the isolated vowel context than either hVd (d = 0.98, p = 0.02) or word-initial contexts (d = 1.39, p < 0.001), although the difference between the hVd and word-initial contexts was not significant. Sighted speakers' AVS was lower in the hVd context than in isolated vowel (d = −1.53, p < 0.001) and word-initial contexts (d = −1, p = 0.01), but there was no significant difference between the isolated vowel and word-initial contexts. There were no significant differences between blind and sighted speakers for any of the contexts, although there was a medium effect size (d = 0.64, p = 0.13) for the difference between the two groups in the word-initial context (larger AVS for sighted speakers). There was a large effect of gender (d = −1.88, p = 0.001; smaller AVS for males). AVS computed from automatically-extracted formant measurements showed similar patterns to AVS using manual formant measurements. 2D AVS measurements also showed similar patterns. Mean AVS per group and context can be seen in **Figure 3.** Sighted speakers show more variability.

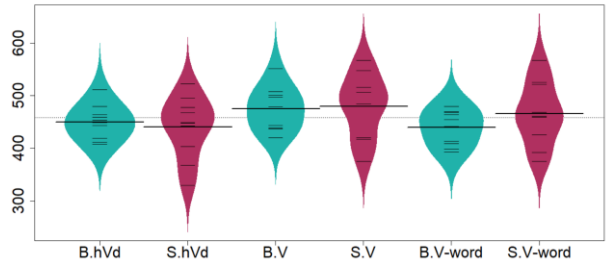

***Figure 3:** Mean AVS per group and context*

3.3. Euclidean Distances – EDs

No significant difference between blind and sighted speakers was observed in EDs between key vowel pairs (d = −0.13, p = 0.8). Type of contrast was not a significant predictor of ED. A large effect of gender was found (d = −1.89, p = 0.001; smaller EDs for male speakers). The analysis on the basis of the automatic formant measurements showed similar patterns. The difference between distances on the basis of manual vs. automatic formant measurements in blind speakers was significant (d = 0.16, p = 0.04; manual formant measurement-based distances larger than automatic formant measurement-based distances), but there was no difference between groups.

3.4. Within-category dispersion

We observed an interaction between vowel length (as a category) and group, where blind speakers produced long vowels with more dispersion than sighted speakers (d = 0.84, p

= 0.005). A large effect of gender was observed ($d = -1.72$, $p = 0.003$; smaller dispersion for males). The difference in dispersion between the manual and automatic obtained formant measurements was significant for both blind ($d = 1.17$, $p = 0.02$) and sighted speakers ($d = 1.38$, $p = 0.003$), with the automatic formant measures containing vowels that were more dispersed. However, the difference between the two groups was similar when using automatic vs. manual formant measurements.

4. Discussion

The present study investigated the production of monophthongs in congenitally blind and sighted speakers of Australian English. Based on our results, we cannot conclusively reject the null hypothesis that there is no difference between blind and sighted speakers.

While we found that blind speakers had smaller VSAs than sighted speakers, blind and sighted speakers did not differ in their production of vowel contrasts (EDs). AVS only differed between groups depending on context. It is interesting that AVS and VSA showed different results, likely due to the characteristics of each measure. VSA takes into account only F1×F2 and a subset of (individual) vowels, while AVS can include F1×F2×F3 and all vowel pairs in the vowel space. Because both F2 and F3 are important acoustic correlates to rounding, we expected to see differences between 2D and 3D AVS (2D AVS excluding F3). However, we found no difference per group, meaning that blind and sighted speakers could be compared on the basis of either measure. This indicates that F3 is not as important (at least for Australian English) as assumed for differentiating blind and sighted speakers.

Blind speakers' long vowels (/iː/, /ʉː/, /ɜː/, /oː/, /ɐː/) showed more within-category dispersion compared to sighted speakers (see Section 3.4). However, this is likely to be language-specific, as vowel length is an important phonemic distinction in Australian English, in contrast to French or Dutch. It is possible that long vowels, due to their longer duration, result in a less precise production. However, we did not have an objective measure of length. In Australian English, some long vowels show characteristics of diphthongs (Harrington e*t al.*, 1997) with more variability in their formant tracks, which could then be problematic for blind speakers who have been found to produce vowels longer in duration in general (Menard *et al.*, 2014). These results should be interpreted with caution, however. Due to the small number of tokens in our study, we did not take context into account when calculating dispersion, so these results may reflect the degree of influence of vowel context on the embedded vowel rather than actual vowel dispersion.

The two contrast types (place of articulation vs. place of articulation and rounding) did not differ in measured EDs, indicating that rounding is an enhancing feature in Australian English. This is likely why F3 was not instrumental in differentiating between the two groups. It also strengthens the hypothesis that variability across studies investigating blind speech stems partly from language-specific differences. There was also no difference between rounded and unrounded vowels in dispersion. This was not surprising, however, as rounding is not contrastive in the specific configuration of the Australian English vowel system.

Although we did find that automatically extracted formants were more dispersed and thus less precise and accurate, automatic extraction would generally be a suitable analysis strategy for this population as long as higher error rates are taken into account. The same patterns were found for VSA, AVS, EDs, and dispersion between groups on the basis of automatic formant data, as on the basis of manual formant data.

In sum, we can conclude that the differences between blind and sighted speakers seem to depend partly on the language, and partly on the methods used to study it. Future studies should include more than just one measure and language, in order to shed more light on the sources of difference between the groups.

5. References

Bates, D., Maechler, M., Bolker, D., Walker S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*(1), 1-48.

Boersma, P., & Weenink, D. (2016). *Praat: Doing phonetics by ɐcomputer* [Computer program]. http://www.praat.org/.

Cox, F., & Palethorpe, S. (2007). Australian English. *JIPA*, 37(3), 341–350.

Elstner, W. (1983). Abnormalities in the verbal communication of visually-impaired children. In A. E. Mills (Ed.), Language Acquisition in the Blind Child: Normal and Deficient. London & Canberra: Croom Helm. 18-41-

Harrington, J., Cox, F., & Evans, Z. (1997). An acoustic phonetic study of broad, general, and cultivated Australian English vowels. *Australian Journal of Linguistics, 17*, 155 – 184.

Lane, H., Matthies, M. L., Perkell, J. S., Vick, J., & Zandipour, M. (2001). The effects of changes in hearing status in cochlear implant users on the acoustic vowel space and CV coarticulation. *Journal of Speech, Language, and Hearing Research, 44*, 552-563.

Lennes, M. (2003). *Formant value collector* [Praat script]. http://www.helsinki.fi/~lennes/praat-scripts/public/collect_formant_data_from_files.praat.

Massaro, D. W., & Palmer Jr, S. E. (1998). Perceiving talking faces: From speech perception to a behavioral principle. MIT Press.

McCloy, D. (2016). *phonR: tools for phoneticians and phonologists*[R package].

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, 264(5588), 746-748.

Ménard, L., Côté, D., & Trudeau-Fisette, P. (2016). Maintaining distinctiveness at increased speaking rates: A comparison between congenitally blind and sighted speakers. *Folia Phoniatrica et Logopaedica, 68*(5), 232-238.

Ménard, L., Dupont, S., Baum, S. R., & Aubin, J. (2009). Production and perception of French vowels by congenitally blind adults and sighted adults. *The Journal of the Acoustical Society of America, 126*(3), 1406-1414.

Ménard, L., Leclerc, A., & Tiede, M. (2014). Articulatory and acoustic correlates of contrastive focus in congenitally blind adults and sighted adults. *Journal of Speech, Language, and Hearing Research, 57*(3), 793-804.

Ménard, L., Toupin, C., Baum, S. R., Drouin, S., Aubin, J., & Tiede, M. (2013). Acoustic and articulatory analysis of French vowels produced by congenitally blind adults and sighted adults. *The Journal of the Acoustical Society of America, 134*(4), 2975-2987.

Ménard, L., Trudeau-Fisette, P., Côté, D., & Turgeon, C. (2016). Speaking clearly for the blind: Acoustic and articulatory correlates of speaking conditions in sighted and congenitally blind speakers. *Plos one*, *11*(9), e0160088.

Turgeon, C., Trudeau-Fisette, P., Lepore, F., Lippé S., & Ménard, L. (2020) Impact of visual and auditory deprivation on speech perception and production in adults. *Clinical Linguistics & Phonetics, 34*(12), 1061-1087.

Vallabha, G. K., & Tuller, B. (2002). Systematic errors in the formant analysis of steady-state vowels. *Speech communication, 38*(1-2), 141-160.

Veenstra, P., Everhardt, M. K., & Wieling, M. (2018). Vision deprived language acquisition: Vowel production and ASR efficacy. Poster presented at the 16th Conference on Laboratory phonology (LabPhon), Lisbon, Portugal.

Venezia, J. H., Fillmore, P., Matchin, W., Isenberg, A. L., Hickok, G., & Fridriksson, J. (2016). Perception drives production across sensory modalities: A network for sensorimotor integration of visual speech. *NeuroImage*, 126, 196–207.

Young, G. S., Merin, N., Rogers, S. J., & Ozonoff, S. (2009). Gaze behavior and affect at 6 months: predicting clinical outcomes and language development in typically developing infants and infants at risk for autism. *Developmental science*, 12(5), 798–814.