4 Discussion and conclusions

The retracting tongue body manoeuvre from [y] to [ʀ], seen in Figure 1 (left) and in profiles 2300 to 2304 in Figures 2 and 3, was depressed slightly. Consequently it passed below the uvula and continued into the pharynx. For this instance of [ʀ], the subject did not elevate the tongue dorsum towards the uvula. Similar behaviour was exhibited by the West Greenlandic subject for [ʀ q χ], and by the second Swedish subject whose results will be presented at the conference.

The target of the tongue body gesture of [ʀ] was the upper pharynx, as hypothesized. This was also the case in the other data to be reported at the conference. The upper pharynx is also the region that is constricted for [o] and [ɔ]-like vowels, which means that this one place of articulation is shared by all these consonants and vowels.

The upper pharynx is a more suitable place than the uvula for producing "uvular" stops, fricatives and trills. The soft smooth elastic surfaces of the posterior part of the tongue and the opposing posterior pharyngeal wall allow perfect occlusion, or the creation of apertures narrow enough for the generation of turbulence.

References

Jones, D., 1964. *An Outline of English Phonetics*. Cambridge: W. Heffer & Sons Ltd. (9th amended edition).

Mrayati, M., R. Carré & B. Guérin, 1988. Distinctive regions and modes: a new theory of speech production. *Speech Communication 7*, 257-286.

Stevens, K.N., 1972. The quantal nature of speech: Evidence from articulatory-acoustic data. In P.B. Denes & E.E. David, Jr. (eds.), *Human Communication: A Unified View*. New York: McGraw Hill, 243-255.

Stevens, K.N., 1989. On the quantal nature of speech. In J. Ohala (ed.), *On the Quantal Nature of Speech. Journal of Phonetics 17*, 3-45.

Wood, S.A.J., 1974. A spectrographic study of allophonic variation and vowel reduction in West Greenlandic Eskimo. *Working Papers 4*, Dept. of Linguistics and Phonetics, University of Lund, 58-94.

Wood, S.A.J., 1979. A radiographic analysis of constriction locations for vowels. *Journal of Phonetics 7*, 25-43.

Wood, S.A.J., 1996a. Temporal coordination of articulator gestures: an example from Greenlandic. *Journal of the Acoustical Society of America* (A). *Poster presented at 131st meeting of the Acoustical Society of America*, Indianapolis.

Wood, S.A.J., 1996b. The gestural organization of vowels: a cinefluorographic study of articulator gestures in Greenlandic. *Journal of the Acoustical Society of America 100*, 2689 (A). *Poster presented at the Third Joint Meeting of the Acoustical Societies of America and Japan*, Honolulu.

Wood, S.A.J., 1997. A cinefluorographic study of the temporal organization of articulator gestures: Examples from Greenlandic. In P. Perrier, R. Laboissière, C. Abry & S. Maeda (eds.), *Speech Production: Models and Data (Papers from the First ESCA Workshop on Speech Modeling and Fourth Speech Production Seminar, Grenoble 1996). Speech Communication 22*, 207-225.

Lund University, Centre for Languages & Literature, Dept. of Linguistics & Phonetics
Working Papers 52 (2006), 149–152

Acoustical Prerequisites for Visual Hearing

Niklas Öhrström and Hartmut Traunmüller
Department of Phonetics, Stockholm University
{niklas.ohrstrom|hartmut}@ling.su.se

Abstract

The McGurk effect shows in an obvious manner that visual information from a speaker's articulatory movements influences the auditory perception. The present study concerns the robustness of such speech specific audiovisual integration. What are the acoustical prerequisites for audiovisual integration to occur in speech perception? Auditory, visual and audiovisual syllables (phonated and whispered) were presented to 23 perceivers. In some of the stimuli, the auditory signal was exchanged for a schwa syllable, a dynamic source signal and a constant source signal. The results show that dynamic spectral information from a source signal suffice as auditory input for speech specific audiovisual integration to occur. The results also confirm that type (and absence) of lip rounding are strong visual cues.

1 Introduction

Visual contribution to speech comprehension was for a long time ignored by theorists and only accounted for when the auditory speech signal was degraded (Sumby & Pollack, 1954). However, McGurk and MacDonald (1976) showed that auditory speech perception could be altered by vision even when the auditory stimulus lacked ambiguity. They used [baba] and [gaga] syllables dubbed on visual stimuli with different consonants. A visual [gaga] dubbed on an auditory [baba] evoked the percept of /dada/. A visual [baba] dubbed on an auditory [gaga] was often perceived as /gabga/ or /bagba/. This demonstrated ordinary speech perception to be a bimodal process in which optic information about a speaker's articulatory movements is integrated into auditory perception. Traunmüller and Öhrström (in press) have demonstrated that this also holds for vowels. It has been shown experimentally that perception of features such as labiality and lip rounding is dominated by the visual signal. In addition, it is worth mentioning that crossmodal illusions are not necessarily restricted to speech perception: Shams et al. (2000) demonstrated that the visual perception of the numerosity of flashes can be altered by simultaneous auditory presentation of clicks.

Bimodal speech perception normally involves synchrony between the auditory and the visual information from a speaker's articulatory movements. Visual information can, therefore, be expected to have a substantial influence on auditory speech perception but visual hearing might require presence of a more or less authentic acoustic speech signal. This study aims at exploring acoustical prerequisites for visually influenced auditory perception to occur. How much information from the original acoustic signal can we remove and still evoke visual hearing? In this study the four long Swedish vowels /i/, /u/, /ɛ/ and /ɒ/ will be tested (appearing both phonated and whispered in a [b_d] frame). In the first condition the formant frequencies of the vowel will be changed (in this case a [ə] will be used). In the second condition the formant peaks will be flattened out, whereby an approximate source signal will be obtained. In the third condition, the formant peaks will be flattened out and the acoustic signal will be kept in a steady state. It can be expected that at least the visible type of lip rounding will have an influence on auditory perception.

2 Method

2.1 Speakers and speech material

One male and one female lecturer from the Department of Linguistics served as speakers. The recordings took place in an anechoic chamber using a video camera Panasonic NVDS11 and a microphone Brüel&Kjær 4215. The speakers produced the more or less meaningful Swedish syllables /bid/, /bud/, /bɛd/ and /bɒd/ in both phonated and whispered fashion. They were also asked to produce [bəd]. The video recordings were captured in DV format and the acoustic signal was recorded separately (sf = 44.1 kHz, 16 bit/sample, mono). The acoustic recordings were subsequently manipulated in different ways using Praat (Boersma & Weenink, 2006): Firstly all acoustic syllables were resynthesized (sf = 11 kHz). The resynthesis was carried out using the Praat algorithm "LPC-burg". The [bəd] syllable was also resynthesized with formant bandwidths expanded to $B_n = 2\ F_n$. The spectrally flattened schwa in this syllable is most similar to a source signal. Finally, to obtain a constant spectrally flattened signal, one glottal period of this schwa was selected and iterated. To obtain a constant whispered spectrally flattened signal, a window of 25 ms of the spectrally flattened whispered schwa was subjected to LPC analysis and resynthesized with increased duration.

The final audiovisual stimuli were obtained by synchronizing the video signals with the manipulated audio signals. The synchronization was based on the release burst of the first consonant and performed in Premiere 6.5. The constant spectrally flattened signals were made equal in duration with the whole visual stimuli (approximately 2s). Each optic stimulus (except [bəd]) was presented together with its corresponding auditory one, the acoustic schwa vowel [ə], the spectrally flattened signal (SF) and the constant spectrally flattened signal (CSF). Each visual syllable (except [bəd]) and each auditory stimulus was also presented alone. In this way, 54 stimuli were obtained for each speaker. The total perception task consisted of two blocks. Block one consisted of 92 audio (A) and audiovisual (AV) stimuli in which each stimulus was presented once in random order. Block two consisted of 16 visual (V) stimuli, each presented twice in random order.

2.2 Perceivers

23 subjects who reported normal hearing and normal or corrected-to-normal vision (11 male, aged 17 to 52 years, and 12 female, aged 20 to 50 years) served as perceivers. All were phonetically naïve native listeners of Swedish.

2.3 Perception task

The perceivers wore headphones AKG 135 and were seated with their faces at approximately 50 cm from a computer screen. All the stimuli were presented using Windows Media Player.

During block 1 (which contained AV and A stimuli), the perceivers were instructed to report what they had heard. Nevertheless, they were instructed to always look at the speaker when shown. The perceivers were allowed to repeat each stimulus as many times as they wished. If they had heard a [bVd] syllable (which could appear in a very distinct or vague manner) they were asked to report which one of the nine long Swedish vowels it resembled the most. They gave their responses by clicking on orthographic symbols of the Swedish vowels (a /ɒ/, e /e/, i /i/, o /u/, u /ʉ/, y /y/, å /o/, ä /ɛ/, ö /ø/) arranged in the right half of the screen in manner of an IPA-chart. There was a response alternative "hör ingen vokal" right under the chart. This was to be used when no syllable was heard or when the sound was not heard as a human vowel.

During block 2 (which contained optic stimuli only) the perceivers were instructed to report the vowel perceived through lipreading. As before, they were allowed to repeat each stimulus as many times as they wished. The whole experiment lasted for approximately 30 minutes.

3 Results

The responses of all subjects to each stimulus combination, whispered and phonated versions pooled, are shown in Table 1. It can be seen that the responses to congruent AV stimuli and auditorily presented vowels are in accord with the speaker's intention. With vowels presented visually only, there were many confusions. The unrounded /i/ and /ɛ/ were mostly confused with other unrounded vowels. The in-rounded /u/ was predominantly confused with other in-rounded vowels (/ʉ/ and /o/)). The out-rounded /ɒ/ was mostly confused with other out-rounded vowels (in this case with /ø/) and, to some extent, with in-rounded vowels. The auditory [ə] was almost exclusively categorized as an out-rounded vowel (/ɒ/ or /ø/) and incongruent visual cues, such as absence of lip rounding, contributed only marginally to the auditory perception.

Table 1. Responses from all subjects (in %) to all stimuli, whispered and phonated versions pooled. Boldface: most frequent response. A: acoustic cues, V: optic cues, SF: spectrally flattened [bəd], CSF: constant spectrally flattened [ə]. "*": no audible vowel.

Stimuli		Responses									
A	V	/i/	/y/	/ʉ/	/u/	/e/	/ø/	/o/	/ɛ/	/ɒ/	*
/i/	/i/	**99**				1					
/u/	/u/				**89**			11			
/ɛ/	/ɛ/					11	1		**86**	2	
/ɒ/	/ɒ/									**100**	
/i/	-	**99**	1								
/u/	-				**86**			14			
/ɛ/	-					10	2		**88**		
/ɒ/	-									**100**	
-	/i/	**70**				24	1		4	2	
-	/u/	1	1	7	**87**			4			
-	/ɛ/	6				18	1		**66**	10	
-	/ɒ/			1	1		19	5		**74**	
[ə]	-						48			**52**	
[ə]	/i/					5	38		3	**53**	
[ə]	/u/						**49**	2		**49**	
[ə]	/ɛ/					1	35		3	**60**	1
[ə]	/ɒ/						**50**			**50**	
SF	-	7	1	1	9	1	22	1	2	**48**	9
SF	/i/	29			4	2	18		3	**39**	3
SF	/u/			2	**48**		9	4	1	27	9
SF	/ɛ/	10			4	4	13		13	**48**	8
SF	/ɒ/		2		9		21	2	1	**61**	4
CSF	-				2		9	1	1	7	**80**
CSF	/i/	5				2	5	1		3	**83**
CSF	/u/			1	7		4	2		4	**82**
CSF	/ɛ/					2	5		4	2	**86**
CSF	/ɒ/		1	1			9			10	**79**

When presented in auditory mode alone, the spectrally flattened vowel (SF) was mostly categorized as out-rounded (in 70% of the cases as /ɒ/ or /ø/) but also, to some extent, as an unrounded or in-rounded vowel. When the auditory source signal was presented with different visual cues, type of rounding was very often perceived in accord with the visual stimulus. The constant source signal (CSF) was not very often identified as a human vowel or syllable, but there were traces of influence from the visual signal.

4 Discussion

These experiments have shown that auditory perception of an approximate acoustic source signal (SF) is sensitive to visual input. In this case, the type of rounding was often perceived in accord with the visual signal. Interestingly, there was a perceptual bias towards /ø/ and /ɒ/ concerning the stimuli containing acoustical [bəd] syllables, (SF) signals and (CSF) signals, while [ə] is undefined with respect to its roundedness. It is obvious that the (SF) and (CSF) still contain some acoustic traces from the [bəd]. In this study the [ə]s produced by the two speakers were categorized as rounded vowels. A possible explanation is that the Swedish phonological system does not offer any unrounded vowels, except possibly /ɛ/, in this region of the vowel space. Therefore, it cannot be excluded that subjects actually heard an unrounded vowel for which they lacked a response alternative, but coarticulation effects from the initial labial consonant might also have caused a bias in favor of rounded vowels.

The auditory perception of the acoustic schwa vowel was not much influenced by the visual signal. This could be due to the fact that a naturally articulated schwa has a definite place in auditory vowel space since the formant peaks are distinct. This makes the acoustic cues quite salient and leaves just a small space for visual influence. On the other hand, the approximate acoustic source signal with preserved dynamics (SF) evoked the perception of a human vocalization, although the place of the vowel in auditory vowel space was only vaguely suggested by its acoustic cues, which were much less salient. This gives opportunity for the visual information about a speaker's articulatory movements to be integrated into the auditory percept. The constant source signal (CSF) lacked both dynamic properties and a distinct place in vowel space. It also lacked the temporal alignment with the visual signals that was present in the other stimuli. It was, therefore, perceived as a sound that was separate from the human utterance that the visible articulatory movements suggested. Thus, it appears that visual hearing of speech requires the presence of an acoustic signal that can easily be interpreted as belonging together with the visual signal.

Acknowledgements

This investigation has been supported by grant 2004-2345 from the Swedish Research Council.

References

Boersma, P. & D. Weenink, 2006. Praat – a system for doing phonetics by computer. *http://www.fon.hum.uva.nl/praat/*.

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

Shams, L., Y. Kamitani & S. Shimojo, 2000. What you see is what you hear. *Nature 408*, 788.

Sumby, W.H. & I. Pollack, 1954. Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America 26*, 212-215.

Traunmüller, H. & N. Öhrström, in press. Audiovisual perception of openness and lip rounding in front vowels. *Journal of Phonetics*.