

Segmental differences in the visual contribution to speech intelligibility

Kuniko Nielsen

UCLA Department of Linguistics

kuniko@humnet.ucla.edu

Abstract

It is well known that the presence of visual cues increases the overall intelligibility of a speech signal [1,2]. Although much is known about segmental differences in both audio-only and visual-only perception, little is known about segmental differences in terms of visual contribution to auditory-visual perception. The purpose of this study was to examine whether segments differ in their visual contribution to speech intelligibility, and whether the presence of visual cues always increases speech intelligibility. Forced-choice word-identification experiments were carried out under auditory-visual (AV) and auditory-only (A) conditions with varying S/N ratios. The experimental results reveal significant differences in the visual contribution for the different consonants, with visual cues greatly improving speech intelligibility for most segments. Surprisingly, the results also suggest that the presence of visual cues can reduce intelligibility. In particular, the intelligibility of [r] decreased significantly in the AV condition, being perceived as [w] in most cases.

1 Introduction

Looking at the speaker's face influences the way we perceive speech: it can change the phonemic perception when the auditory and visual information are incongruent [3] ('McGurk effect'), and it improves overall intelligibility when the information from the two modalities is congruent [1,2]. Although the latter effect is particularly prominent when the auditory signal is less than optimal, perceivers appear to use the visual signal regardless of auditory signal quality.

Despite the powerful influence of visual information in face-to-face communication, there are important open questions remaining in the field of audio-visual speech perception. Although the acoustical and articulatory properties of individual segments have been studied extensively [4,5] – producing knowledge that has been applied to development of speech production/perception theories – the visual contribution to audio-visual (AV) speech intelligibility has been mostly discussed in terms of its overall increase between audio-only and audio-visual speech intelligibility. For example, [1] showed that the addition of visual information could increase overall speech intelligibility by an amount equivalent to that produced by increasing the level of an auditory signal by 15 dB. On the other hand, segmental differences have been found and studied in both audio-only and visual-only perception [4-6] studies. However, little is known about the combination of the two modalities: namely, segmental differences in the visual contribution to audio-visual speech intelligibility. An investigation of segmental differences in audio-visual speech perception seems beneficial for a better understanding of the

way we perceive speech. Given that listeners seem to use the visual signal regardless of auditory signal availability, and that segments are different in their salience both visually and auditorily, we expect the visual contribution to speech intelligibility to differ across segments. In particular, the segments with salient visual cues and relatively poor acoustic cues (e.g. /t/, /θ/) are expected to display a greater visual contribution to speech intelligibility than those segments with relatively poor visual and acoustic cues (e.g. /r/). It is also our interest to examine the possible range of visual contribution.

The aim of this study is to examine (1) whether segments differ in their visual contribution to speech intelligibility, (2) whether segments with relatively salient visual cues display a greater visual contribution, and (3) whether the contribution of visual cues is always to increase speech intelligibility. In addition, this study also aims to replicate the results of Sumby *et al.* [1] using up-to-date technology to determine if similar results will still be obtained. An experiment involving both audio-visual and auditory-only speech perception was carried out.

2 Method

Sixteen native speakers of American English (11 females and 5 males, age 18-25) with normal hearing and normal or corrected vision served as subjects for this experiment. The material consisted of 108 English words that met the following criteria: (1) monosyllabic (CVC), (2) the initial consonant was one of the 15 American English consonants /p, t, k, f, θ, s, b, d, g, v, ð, ʃ, tʃ, r, w/, (3) the word was listed in the CELEX corpus of frequency counts. Real words were used as opposed to nonsense syllables [4,5,6] in order to reduce segmental frequency effects. The consonants were arranged into six triplet groups such that each set contains auditorily highly confusable consonants in noise. The six triplet groups were then classified into two groups, Easy and Hard: The segments in the Easy group (p/t/k, b/d/g, f/θ/s, v/ð/b) are expected to be easy to distinguish visually, and contrast with their counterparts in place of articulation within a triplet, while the segments in the Hard group (r/w/v, s/ʃ/tʃ) are expected to be difficult to distinguish visually, and contrast in manner of articulation (and sometimes secondary place). Note that /b/, /s/ and /v/ were included in two triplets for the sake of forming suitable triplets. The degree of visual confusability was determined from [6]. Six minimal sets (e.g., *pick, tic, kick*) were then chosen for each triplet group. In order to control possible bias due to lexical frequency effects, the relative frequencies in the stimulus set were balanced.

The stimuli were recorded in a sound booth in the UCLA Phonetics Laboratory. The speaker producing the stimuli was a trained phonetician. She was seated in front of a plain dark blue background, approximately 1.5 m away from a video

camera. All utterances were recorded onto audiotape and videotape (audio signal - 48 kHz/16-bit; tape speed - 28.2 mm/sec), and then transferred onto a computer. The movie clips were edited into 72 three-second clips using Apple iMovie. The audio tracks were extracted from the edited movie files (sampling rate: 22000 Hz) so that Audio-Visual and Auditory-Only tokens had exactly the same sound tracks. Signal level was determined in terms of the peak RMS amplitude over a 30 ms window. All the audio files were then equated for the peak RMS amplitude at 80 dB (nominal). Noise was then added to the speech at several S/N ratios. The noise used in this experiment is flat shape, band-pass filtered at 200-6500 Hz using Kay Elemetrics' MultiSpeech. This noise spectrum was chosen from [4] because it seemed to induce confusions related to place of articulation in their study. The audio files and noise were mixed in five different S/N ratios: (-10, -5, 0, 5, 10, 15 dB) using the program NOISE (Tehrani, 2002). S/N was defined by keeping the signal level constant at 80dB and manipulating the noise level from 65dB to 90dB. The experimental audio stimuli were calibrated at a fixed 85 dB SPL (Larson Davis 800B) for all the S/N ratios and presented binaurally over headphones. The visual displays were presented on a 20-inch computer screen.

The stimuli were presented using Psyscope 1.2.5 [7]. Each subject was seated in front of a computer in a sound booth. A three-button button box was placed directly below the computer screen. Each main session was divided into two blocks: auditory-only (A) and auditory-visual (AV). In the A block, subjects were asked to listen to a word while they saw three response options (a minimal triplet) on a screen, and to press the button which corresponded to the word they thought they heard (forced-choice). In the AV block, the same subjects were asked to watch and listen to the video clips of the speaker on the screen, again with a forced-choice response from three words, and to press the button which corresponds to the word they thought they heard. The subjects were told to guess when unsure, or if they could not detect the stimulus in the noise. Figure 1 shows an example of what the subjects saw on the computer screen in the AV block. The screen in the A block looked the same regarding the response choices, but it was otherwise blank with no video display.

Due to experimental time limitations, the 72 words were divided into two lists of 36 words each. The program randomized 36 words and 6 S/N ratio settings within a S/N setting and a block, respectively, and recorded both the key response and the reaction time. Each subject went through 432 total trials: 2 blocks (A and AV), 6 S/N ratio settings in each block, 36 trials in each setting. Each word was presented once for each trial, and one session lasted about 45 minutes, including the initial practice session. The order of blocks and the word lists within blocks were counterbalanced across subjects.

3 Results

The independent variables in this study are 1) Signal-to-Noise (S/N) ratio, 2) presence of visual (V) information in addition to audio (A) (A/AV), 3) segment class (=consonant), 4) Easy/Hard (visibility of cues). (Note that lexical frequency, order of A/AV blocks presentation, and word list were counterbalanced across subjects). The effect of each of these variables on the correct response rate was determined by cross-section time-series logistic regression [8]. Note that the data for /s/, as well as /v/ when paired with /r/ and /w/ were excluded from the analysis.

It was found that /s/, as well as /v/ when paired with /r/ and /w/, were identified perfectly both with and without visual information. Given that the main question of interest is the additional information provided by visual presentation and items which were not degraded by noise cannot contribute to answering the question, the data for these sounds were excluded from the analysis.

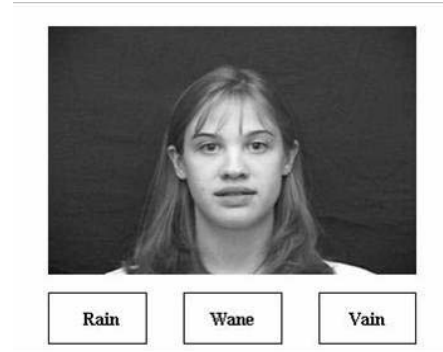


Figure 1: An example stimulus presentation of the AV block.

3.1 S/N ratio and presence of visual information

Figure 2 shows the speech intelligibility under A and AV conditions as a function of (auditory) S/N ratio. All 15 segments were averaged to produce the curve in Figure 2. We see from Figure 2 that speech intelligibility under both A and AV conditions increases as the signal-to-noise ratio is increased, and the effect of the signal-to-noise ratio on speech intelligibility is statistically significant in both conditions (A: Wald $\chi^2(14) = 194.96$, $p < 0.001$, AV: Wald $\chi^2(14) = 8.57$, $p = 0.0034$). Figure 2 also shows that the visual contribution to oral speech intelligibility (the difference between the A and AV curves) increases as the signal-to-noise ratio is decreased. These results are as expected and in agreement with [1].

The effect of the presence of visual information was found to be statistically significant (Wald $\chi^2(1) = 316.45$, $p < 0.001$). The presence of visual information increases the overall intelligibility of speech perception. This result agrees with previous studies [1,2].

3.2 Segment Class

The effect of segment (the 15 consonants) on speech intelligibility was found to be statistically significant under both A and AV conditions (A: Wald $\chi^2(14) = 262.88$, $p < 0.001$, AV: Wald $\chi^2(14) = 397.07$, $p < 0.001$). Figure 3 shows speech intelligibility under A and AV conditions as a function of S/N ratio across the 15 segments. As can be seen, the locations of the A curves differ across segments, and this confirms the results from previously reported acoustic confusion matrices [4.5]. There are two segments that show a lower curve for AV condition, namely, /r/ and /s/. However, statistical tests showed that only for /r/ is this negative visual contribution significant. They are both in the group which contrast manner of articulation (Hard), and thus a relatively small visual contribution was expected. However, a negative contribution was not expected since no previous study has shown this type of visual effect. This result will be further discussed below.

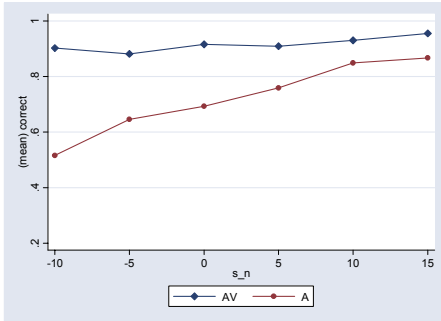


Figure 2: Speech intelligibility under AV (audio-visual) and A (auditory only) conditions as a function of the S/N ratio

3.3 Easy/Hard

A significant effect of Easy vs. Hard on correct response was found under the AV condition, but not under the A condition (A: Wald $\chi^2(1) = 0.48$, $p = 0.4888$, AV: Wald $\chi^2(1) = 233.26$, $p < 0.001$). This result was predicted given that this Easy/Hard classification was made solely based on visual confusability. There is no significant difference for Hard group between the conditions, while there is more than a 25% increase of intelligibility for the Easy group (A:72%, AV:98%).

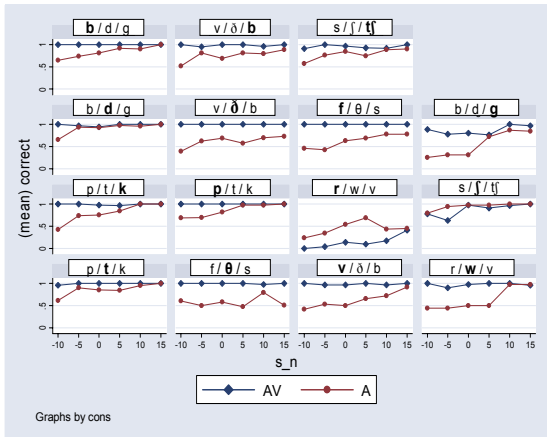


Figure 3: Speech intelligibility under AV (audio-visual) and A (auditory only) conditions as a function of S/N ratio. Bold face consonants were presented as stimuli.

4 Discussion

4.1 Segmental Difference

Our results show that the contribution of visual cues to auditory-visual perception differs significantly across segments. The Easy group showed larger visual effects than the Hard group. Among them, the consonants with relatively poor acoustic cues (such as /f/, /θ/, /v/, /δ/) improved more than the ones with relatively salient acoustic cues (such as /d/, /t/). Also as expected, those segments which contrast in manner of articulation with their counterparts showed a very small (or even negative) visual contribution to speech intelligibility. Given this result, we expect that the contribution of visual cues depends (at least partially) on the segmental composition of the stimuli. Thus, the notion of ‘visual cues = +15 dB’ should not hold for many cases, and there should indeed be no such magic number. Our data suggest that the addition of visual cues does not simply

increase overall intelligibility: it actually changes the pattern of intelligibility as well, in the sense that some segments become more intelligible but some do not.

This segmental difference might explain cross-linguistic McGurk effect variations found in the literature. [9][10] examined the strength of the effect between Japanese and American listeners, and reported that the effect differs between the two languages. By conducting McGurk-type experiments, [11] examined the effect of language and culture (Japanese, Spanish, and English) on speech perception in face-to-face communication. Although they reported that there is no difference in the nature of processing across language groups, their data showed a relatively weaker effect of visual cues among Japanese speakers. The results from this study suggest that the segmental difference in visual contribution may account for this cross-linguistic difference. Every language has a different phoneme inventory, and some languages have more visually distinctive segments (e.g. labials, interdental) than others. If the contribution of visual cues depends on the segmental composition of the stimuli, there must be cross-linguistic differences due to inventory variation. If a language has few visually distinctive segments (such as Japanese) and/or has many visually indistinctive segments (such as Yupik), even if the nature of processing is the same as a language which has many visually distinctive segments (such as English), their weight of visual information in processing speech may be much smaller. Additional cross-linguistic studies testing audio-visual perception are needed to answer this question.

4.2 Negative Effect

Perhaps the most striking finding of this study is that the presence of visual cues can actually decrease speech intelligibility. We found one such case, /r/. According to the auditory confusion matrix in [12], the intelligibility of /r/ and /w/ are quite similar at -5 dB S/N and were both often misheard as /j/, although /w/ is much more intelligible than /r/ at +5 dB S/N. Our results confirm this acoustic difference between the two consonants (see Figure 3). On the other hand, the visual confusion matrix in [6] shows that /r/ was often mistaken as /w/ and /f/ (26%, 26%, and 35%, respectively) while /w/ was rarely perceived incorrectly (89% correct response, the highest among 23 consonants), indicating that visual cues for /w/ are more salient than those for /r/. Taken together, it appears that /r/ has very weak auditory and visual cues, while /w/ has weak auditory cues yet very strong visual cues (even more than other labials). According to [6], the optical signals of /r/ and /w/ are quite different from each other. If that is the case in general, then, why did all the subjects in this study, as well as the subjects in Jiang’s study, often respond with /w/ when /r/ was presented as the stimulus?

One possibility is that this effect in this study is due to the speaker: her speech actually exhibits very distinctive lip-rounding in general. None of the subjects had seen her talking before the experiment and had no chance to normalize for her speech. If their image of /r/ does not involve much lip-rounding, it seems reasonable that they perceived the visual cues of the speaker’s very rounded /r/ as those of /w/. As mentioned earlier, /r/ has relatively weak acoustic cues, and thus auditory signals would not help in determining the segment in lower S/N ratio settings. If the facial movement of our speaker’s /r/ was different from our subjects’, what is the source of this variation? According to the data from [6], one of his four speakers (M2) showed a very similar pattern to our speaker: among his 360 /r/ tokens,

271 (75%) were perceived (lip-read) as /w/ while only 77 (23%) were perceived as /r/ (summed over five deaf adults). In terms of visemes, two speakers (M2 and F1) formed {r,w} clusters, and the other two formed {r,f,v} clusters. Note that unlike speaker M2, speaker F1's /r/ was perceived correctly 59% of the time. Given that Jiang's speakers were all from California, it does not seem to be a dialect difference.

Another possibility that could account for the occasional negative effect of visual cues is that perceivers know that the visual properties for /r/ have much wider variation compared to visually more salient consonants like /w/, and thus they do not tune into visual cues of /r/ in general. When we produce /r/, there are three places of constriction (the lip, palatal and pharyngeal regions) which all contribute to its low F3, and thus there are many ways to achieve this acoustic goal. In particular, lip-rounding is useful, yet not necessarily crucial. However, lip-rounding is absolutely crucial for /w/ and thus speakers round their lips without exception.

As mentioned earlier, the physical measurements of speech production by Jiang [6] showed a large difference between /r/ and /w/. However, his visual confusion matrices show that two speakers' speech (out of four he tested) formed a viseme (cluster) {w, r}, indicating that the two were perceptually indistinguishable. The same viseme was also obtained in [13, 14]. Note that these visemes are formed mainly due to the confusion of /r/ as /w/, for /w/ was rarely misperceived as /r/. It might also be the case that the perceivers have little conscious knowledge of what /r/ is supposed to look like (since it varies so much across speakers), but they do know what /w/ should look like. Therefore, when they are *forced* to pay attention to visual cues of /r/ as in this experiment, other segments with similar yet more salient visual cues (in this case, /w/) are chosen. This scenario seems to fit the study by [14] which showed that /r/ was relatively undefined in the pre-training testing, and yet it demonstrated the largest lip-reading training effect. Of course, this type of process would take place only in those special cases where neither its auditory nor visual cues are sufficient for reliably identifying the stimulus.

One of the key issues in audio-visual integration is whether the integration happens pre-phonetically or post-phonetically. Investigating the nature of bimodal integration is not the aim of this study, and our results do not provide any conclusive support for either views. However, simulations by the model which assumes pre-phonetic integration (FLMP) [15] successfully predicted the negative effect obtained in this study, providing additional credibility to the pre-phonetic integration. Lastly, although the current model weighs the input from two modalities equally, it may be enabled to predict the cross-linguistic variation reported in previous McGurk studies by assigning different weights to each modality.

5 Conclusions

The current study examined segmental differences in their visual contribution to speech intelligibility, by conducting intelligibility tests for 15 English consonants with, and without, visual observation of the speaker's facial movements. As expected, there were significant differences in the visual contribution for the different consonants, with visual cues greatly improving speech intelligibility for most segments. The segments with more salient visual cues (e.g. /f/, /θ/) displayed greater improvement than segments with less salient cues. Surprisingly, the results also suggest that the presence of visual cues can reduce intelligibility. In particular, the intelligibility of /r/ decreased significantly in

the AV condition, being perceived as /w/ in most cases. These results are relevant to explaining 1) the inconsistency in terms of the magnitude of visual-gain found in the previous audio-visual perception literature, and, possibly, 2) the attested cross-linguistic variability in McGurk effect. This study also aimed to replicate [1] using more up-to-date methods. Our result agrees with their finding that the visual contribution to oral speech intelligibility increases as the signal-to-noise ratio is decreased, although it does not provide support for their finding that its contribution relative to its possible contribution is independent of S/N ratio.

6 References

- [1] Sumby, W. and Pollack, I. "Visual Contribution to Speech Intelligibility in Noise", *J. Acoust. Soc. Amer.*, Vol. 26, 212-215, 1954.
- [2] Erber, N. "Interaction of Audition and Vision in the Recognition of Oral Speech Stimuli", *JSHR*, Vol. 12, 423-425, 1969.
- [3] McGurk, H. and MacDonald, J. "Hearing lips and seeing voices", *Nature*, Vol. 264, 746-748, 1976.
- [4] Miller, G. and Nicely, P. "An Analysis of Perceptual Among Some English Consonants", *J. Acoust. Soc. Amer.*, Vol. 27, 338-352, 1955.
- [5] Wang, M. and Bilger, R. "Consonant confusion in noise: a study of perceptual features", *J. Acoust. Soc. Amer.*, Vol. 54, 1248-1266, 1973.
- [6] Jiang J. "Relating Optical Speech to Speech Acoustic and Visual Speech Perception", *Ph.D. diss., Dept. of Electrical Engineering, UCLA, 2003*.
- [7] Cohen, J. D., MacWhinney, B., Flatt, M., & Provost, J. "PsyScope: a new graphic interactive environment for designing psychology experiments", *Behavioral Research Methods, Instruments, and Computers*, 25, 257-271, 1993.
- [8] Hardin, J. and Hilbe, J. *Generalized Estimating Equations*. Boca Raton, FL: Chapman & Hall/ CRC, 2002.
- [9] Sekiyama, K., & Tohkura, Y. "McGurk effect in non-English listeners: Few visual effects for Japanese subjects hearing Japanese syllables of high auditory intelligibility", *J. Acoust. Soc. Amer.*, Vol. 90, 1797-1805, 1991.
- [10] Sekiyama, K., & Tohkura, Y. "Inter-language differences in the influence of visual cues in speech perception", *Journal of Phonetics*, Vol. 21, 427-444, 1993.
- [11] Massaro, D. W., Tsuzaki, M., Cohen, M., Gesi, A., & Heridia, R. "Bimodal speech perception: An examination across languages", *Journal of Phonetics*, Vol. 21, 445-478, 1993.
- [12] Luce, P. A. *Neighborhoods of words in the mental lexicon. Research on Speech Perception Progress Report, No. 6. Bloomington: Indiana University, Psychology Department, Speech Research Laboratory, 1986*.
- [13] Woodward, M. F. and Barber, C. G. "Phoneme perception in lipreading", *JSHR*, Vol. 3, 212-222, 1960.
- [14] Walden, B.E., Prosek, R.A., Montgomery, A.A., Scherr, C.K., and Jones, C.J. "Effect of training on the visual recognition of consonants", *JSHR*, Vol. 20, 130-145, 1977.
- [15] Massaro, D. W. *Perceiving Talking Faces*, MIT Press. Cambridge, MA, 1998.