UCLA Working Papers in Phonetics, No. 103, 106-147

Segmental differences in the visual contribution to speech intelligibility[*]

Kuniko Nielsen

University of California, Los Angeles

Abstract

It is well known that the presence of visual cues increases the intelligibility of a speech signal (Sumby & Pollack 1954). Although much is known about segmental differences in visual-only perception, little is known about the contribution of visual cues to auditory-visual perception for individual segments. The purpose of this study was to examine (1) whether segments differ in their visual contribution to speech intelligibility, and (2) whether the contribution of visual cues is always to increase speech intelligibility. One talker produced triples of real words containing 15 English consonants (p, t, k, f, θ, s, b, d, g, v, ð, ʃ, tʃ, r, w). Forced-choice word-identification experiments were carried out with these recordings under auditory-visual (AV) and auditory-only (A) conditions with varying S/N ratios, and identification accuracy for the 15 consonants was compared for A vs. AV conditions. As expected, there were significant differences in the visual contribution for the different consonants, with visual cues greatly improving speech intelligibility for most segments. Among them, labio-dentals and inter-dentals show the largest improvement. Although individual perceivers differed in their

[*] This paper was originally submitted as MA thesis, UCLA, Department of Linguistics. I would like to thank Pat Keating, Colin Wilson, and Bruce Hayes for their time and assistance with this paper. I would also like to thank Rebecca Scarborough for her willingness to assist as the speaker.

performance, the results also suggest that for some consonants, 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 (e.g. McGurk and McDonald, 1976), and it improves overall intelligibility when the information from the two modalities is congruent (O'Neill 1954, Sumby & Pollack 1954, Erber 1969). 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 – producing knowledge that has been applied to development of speech production/perception theories – the visual contribution to audio-visual speech intelligibility has been mostly discussed in terms of its overall effect, and relatively little is known about the visual properties of individual segments. An investigation into the segmental differences in audio-visual speech perception is essential for a complete understanding of the way we perceive speech.

In this thesis, I will present the results of an audio-visual speech perception experiment that was designed to examine possible segmental differences in their visual contribution to speech intelligibility. I will argue that there are significant differences in the visual contribution across segments, and surprisingly, the presence of visual cues sometimes reduces intelligibility.

The outline of this paper is as follows: Sections 1.1 and 1.2 describe in detail several previously published studies by other research groups that represent what is known about the factors influencing speech intelligibility. Section 1.1 focuses on audio-visual speech perception studies that attempt to quantify the visual contribution to speech intelligibility. Section 1.2 describes what is already known about the difference among individual segments. The discussion in Section 1.2 revolves around the description of "confusion matrices" (both auditory and visual), which describe the perceptual confusability of segments and thus indirectly the similarities and differences among them. The introduction ends with Section 1.3, where I present the specific aims and hypotheses of this study. Section 2 describes the experimental method in detail, and the results follow in Section 3. In Section 4, I discuss implications of the results.

1.1 Visual Contribution to Speech Intelligibility

Sumby and Pollack (1954) examined the visual contribution to speech intelligibility as a function of auditory signal-to-noise (S/N) ratios and vocabulary size. They tested the speech intelligibility of real words both with and without supplementary visual observation of the speaker's facial movements. They concluded that the visual contribution to overall speech intelligibility increases as the S/N ratio is decreased, such that a listener who has never been trained in lip-reading always perceives speech more accurately when she sees the speaker's face than when she does not, while the ratio of actual contribution (the intelligibility with visual observation minus the intelligibility without visual observation: AV-A) to the possible contribution (the maximum intelligibility minus the intelligibility without visual observation: 100% - A) is independent of S/N ratio. Although this effect is more robust in noise, it exists

regardless of auditory signal quality. Later audio-visual studies confirmed these points (e.g. Erber, 1969, Summerfield, 1979).

Since Sumby and Pollack (1954) showed that the addition of visual information could increase overall speech intelligibility by as much as would an increase in S/N ratio of +15 dB, their result has often been interpreted as 'visual cues = +15 dB', although they never made such a claim. That is, the presence of a visual signal has been taken to be equivalent to increasing the level of an auditory signal by 15 dB. However, it is clearly shown in the Sumby and Pollack study that the size of possible responses (from eight to 256 words) influences the magnitude of the visual effect. There are many other possible factors that could also influence the visual effect, such as the frequency range of masking noise, the type of stimuli, absolute amplitude of auditory signal, and so on.

Experimental procedure could be another factor that influences the results. For example, Sumby and Pollack (1954) had a live speaker recite the test items in front of subjects to provide the audio-visual stimuli. In contrast, in later studies speakers were often video-recorded and video clips were used as visual stimuli (e.g. Summerfield, 1979). Also, there seems to be no standard procedure for obtaining a given S/N ratio. In Sumby and Pollack (1954), the overall signal level was measured in terms of average peak deflection of a VU meter. The S/N ratio was varied by holding the noise level constant and varying the speech level, resulting in greater absolute amplitude for greater S/N ratios. Some later studies also kept the noise level constant, but across studies the constant noise level varied from 60dB (Summerfield, 1979) to 90 dB (Erber, 1969). Other studies held the signal level constant and varied the noise amplitude (Binnie et al. 1974).

Given that there are a number of factors that could influence the results, it is not surprising that various studies have shown different results in terms of the visual contribution to perception. Erber (1969) tested the intelligibility of words with a spondaic stress pattern with and without visual information, and reported that the improvement in the correct response rate with visual information was 28% at -10 dB S/N. Summerfield (1979) used short sentences as stimuli, and the improvement was 20% at -12 dB S/N. In Sumby and Pollack (1954), the improvement at similar S/N ratio varies from over 20% to below 40%, depending on the vocabulary size of stimuli, while it was as much as 49.8 % at -12 dB S/N in Binnie et al. (1974).

The increase in intelligibility can also be shown in terms of speech-reception threshold in noise (SRTN), which is the minimum S/N ratio at which the speech stimuli (e.g., key words) can be identified at a certain accuracy. MacLoed and Summerfield (1987) measured audio-alone and audio-visual SRTNs using short sentences, and the resulting visual benefit ranged from +3 dB to +22 dB among the test sentences, yielding the mean of +11 dB. Later, they developed a more efficient method to obtain SRTN using an adaptive, or "up-down", procedure (MacLoed and Summerfield, 1990). The results revealed a mean visual benefit of +6.4 dB, ranged from +2.7 dB to +9.5 dB across the sentences. Grant and Seitz (2000) investigated the interaction between visual cues and time-aligned auditory cues by testing signal detectability under audio-only (A), time-matched audio-visual (AVm), and time-unmatched audio-visual (AVm) conditions. They used short sentences as stimuli that were carefully balanced in terms of phonetic segments, number of syllables, grammatical structure, and intonation. The amount of threshold reduction (AVm-A) was 1.56 dB on average, yet it varied significantly for the 3 target sentences they tested. These within-experiment variances are intriguing in particular because all the obvious factors that could affect intelligibility (masking noise, the type of stimuli, and absolute amplitude

of auditory signal, etc.) – except for segment-specific properties – were controlled across the sentences.

This inconsistency among the audio-visual perception studies suggests the possibility of additional factors which influence audio-visual intelligibility.

1.2 Confusion Matrices and Segmental Differences

Other studies have investigated auditory-only and visual-only intelligibility. Unlike the audio-visual literature, those studies often present their results in the form of confusion matrices. A confusion matrix shows a row count of actual versus correct responses on a class-by-class basis. In speech perception, it plots each stimulus phoneme against subjects' responses to that phoneme (the actual response), and is interpreted as showing the perceptual confusability of segments, as well as the similarities and differences among them. The most well-known auditory confusion matrices are the ones by Miller and Nicely (1955). They tested auditory perceptual confusion among 16 English consonants in noise, as a function of S/N ratios and frequency ranges of masking noise. The data show clear differences in perceptual confusion between the consonants: for example, the intelligibility of /ʃ/ is usually higher than other segments, while the intelligibility of /θ/ and /ð/ are quite low; these are often misperceived as /f/ and /v/ respectively. The data were further analyzed in terms of five articulatory features: voicing, nasality, affrication, duration, and place of articulation. The results revealed that the perception of any of these features is relatively independent of the perception of the others, and that the place of articulation was the hardest to hear correctly. Miller and Nicely predicted that the addition of visual cues should be likely to eliminate place confusions, although they did not test that prediction. Similar confusion results are reported by many others including Wang and Bilger (1973) and Luce (1986).

Visual confusion matrices, on the other hand, have been reported less often than auditory confusion matrices. Walden et al. (1977) investigated the effect of lip-reading training by conducting visual intelligibility tests of 20 English consonants before and after training sessions. Their visual confusion matrices indicate differences in confusability between the consonants, as well as some response biases: for example, if two consonants are mutually confusable and the distinction between the two is voicing, the voiceless response is favored (except for /p/ vs. /b/). If the distinction is nasality, the oral consonant response is favored. They further performed a cluster analysis to organize the data into structures that reflect the relative visual similarities among the consonants. As a result, nine groups of phonemes that are visually indistinguishable at some degree of accuracy (= visemes) were derived: {p,b,m} {w} {r} {f,v} {θ,ð} {ʃ,ʒ} {s,z} {t,d,n,k,g,j} {l}. They reported that the visemes {s,z} and {r} were relatively visually-undefined (= low accuracy) in the pre-training test, and yet showed the greatest training effect.

Jiang (2003) provided the most extensive visual confusion matrices currently available. He quantified optical (facial) signals of 69 CV syllables in American English (23 initial consonants and 3 vowels), and examined their relationship with acoustic signals as well as with visual speech perception. His visual confusion matrices show clear differences in visual confusion between the consonants. For example, very high intelligibility can be observed for /w/, /h/ and /f/, while /r/, /n/ and /g/ appear to be very confusable. The following six visemes were derived from the matrices: {w} {p,b,m} {r,f,v} {θ,ð} {l,n,k,g,h,j} {t,d,s,z,ʃ,ʒ,tʃ,dʒ}. These visemes, as well as some response biases found in this study, are in overall agreement with those of Walden et al. (1977). His results also indicate that places of articulation are visually distinguishable, supporting the plausibility of the prediction by Miller and Nicely (1955). The

voicing contrast and manner of articulation[1], on the other hand, are shown to be visually indistinguishable.

Another visual confusion matrix from Binnie et al. (1974) shows the confusions among five visemes (bilabials, labiodentals, interdentals, rounded labials, and linguals) derived from 16 consonants tested. They reported that the subjects' ability to visually recognize five places of articulation was nearly perfect (98.2%), while nasality and voicing contrasts are reported to be most difficult to visually recognize. These findings are in overall agreement with those of Jiang (2003), although his results show lower accuracy for place identification. On the other hand, the results from their audio-only condition show that nasality and voicing contrasts are most resistant to masking noise[2], while place of articulation is most severely affected. Again, the plausibility of the prediction by Miller and Nicely (1955) that the addition of visual cues would eliminate place confusions is supported here. Similar results are obtained by other visual-only studies (e.g. Woodward and Barber, 1960).

As discussed above, many auditory and visual confusion studies seem to agree in their findings in general, and all indicate significant differences in intelligibility across segments. This is true even when non-CV stimuli are used: Rosen and Corcoran (1982) tested lip-reading intelligibility of spoken sentences, and reported that sentences that are easy to lip-read are made up of words that start with visibly distinctive consonants, such as "f", "ð", and "b" rather than "t", "s", and "k". For example, more than 90% of native observers with normal hearing correctly lip-read the sentence "The boy's running away", while none managed to lip-read the sentence "The tea towel's by the sink". It is unquestionable that some segments have more robust visual cues than others, just like some have more robust acoustic cues than others, and thus are more

[1] The three labial visemes ({w} {p,b,m} {r,f,v}) differ in manner as well as place, however.
[2]This can also be seen in their audio-visual (A-V) condition.

intelligible than others. Also, the literature on both auditory and visual confusion matrices seems to agree with Miller and Nicely (1955) that (1) voicing and nasality are readily perceived auditorily, while place of articulation is the hardest to hear correctly, and (2) visual cues should eliminate the place confusions that do occur auditorily.

1.3 Aims & Hypotheses

The nature of audio-visual speech perception and bimodal integration has been studied extensively for psychological and cognitive-scientific interest, as well as for clinical interest. However, the existing literature seems to suffer from a weakness in its linguistic treatment of audio-visual speech perception. As mentioned earlier, most studies in audio-visual speech perception focus on the overall increase between audio-only and audio-visual speech intelligibility, and the results are inconsistent in terms of the magnitude of intelligibility gain. On the other hand, segmental differences have been found and studied in both audio-only and visual-only perception. Nevertheless, little is known about the combination of the two modalities: namely, segmental differences in the visual contribution to speech intelligibility. An investigation of segmental differences in audio-visual speech perception seems beneficial for finding a possible source of the previously mentioned inconsistency in the literature as well as for a better understanding of the system.

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. /f/, /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. In particular, if the segment has very poor visual cues, would seeing it still increase visual intelligibility?

The purpose 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 Sumby and Pollack (1954) and other early studies (e.g. Erber, 1969) using up-to-date technology to see if similar results will still be obtained.

Based on these purposes, the following hypotheses are formed:

1. **Consonants differ in their visual contribution to audio-visual speech intelligibility.**
2. **Visually distinctive consonants show greater visual contribution to speech intelligibility than visually confusable consonants.**
3. **The presence of visual information always increases speech intelligibility regardless of its visual salience.**
4. **The visual contribution to audio-visual speech perception increases as the signal-to-noise ratio decreases.**

An experiment involving both audio-visual and auditory-only speech perception was carried out to test these hypotheses. The details of the experiment are described in the following section.

2 Method

2.1 Subjects

Sixteen native speakers of American English with normal hearing and normal or corrected vision served as subjects for this experiment. They were recruited from the UCLA undergraduate population, and included 11 females and 5 males. The age of the subjects varied from 18 to 25 years. The subjects received $10 for their participation.

2.2 Stimuli

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 (as in Miller & Nicely 1955, and Jiang 2003) in order to reduce segmental frequency effects. Nonsense syllables are prone to introduce both segmental and lexical bias, given that many nonsense syllables (CV) are, or are very close to, real words in English. The aforementioned 15 consonants were arranged into six triplet groups: p/t/k, b/d/g, f/θ/s, s/ʃ/tʃ, v/ð/b, r/w/v. Note that /b/, /s/ and /v/ were included in two triplets in the experiment for the sake of forming suitable triplets. The triplets were chosen such that each set contains auditorily highly confusable consonants in noise, based on the confusion matrices from Miller & Nicely (1955) and Wang and Bilger (1973)[3]. 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

[3] For the sake of forming triplets, a few sets included less confusable consonants (i.e., f/θ/s), however.

place of articulation within a triplet[4], 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). The degree of visual confusability was determined from the visual confusion matrices and 3-D multidimensional scaling (MDS) in Jiang (2001 and 2003). Six minimal sets (e.g., *pick, tic, kick*) were then chosen for each triplet group, making up 36 minimal sets (= 108 words) in total. Table 1 shows, as an example, all the sets for the p/t/k triplet group. A complete listing of these stimuli and their frequency counts appears in the Appendix.

Test consonants	Option 1	Option 2	Option 3
p/t/k	**pick**(3418)	tic(269)	kick(988)
	pot(657)	tot(33)	cot(406)
	pin(568)	**tin**(767)	kin(60)
	puff(239)	**tough**(751)	cuff(152)
	perk(47)	Turk(127)	**kirk**(195)
	pill(507)	till(1399)	**kill**(3835)

Table 1: Example of Perception Stimuli
HIs (the words with the highest frequency counts) are shown in bold, and LOWs are shown in shade. The boldface and shaded words were recorded by the speaker and used as stimuli. The remaining words were not recorded, but were provided as response choices to the subjects.

In order to control possible bias due to lexical frequency effects, the relative frequencies in the stimulus set were balanced in the following manner. Within each minimal set, the word with the highest frequency within each minimal set was labeled as "HI" (e.g. 'pick' in row #1 above), and a word with a lower[5] frequency was then labeled as "LOW" (i.e., 'tic' in the row #1 above). The number of HIs and LOWs were balanced for each consonant. The same number of HIs and LOWs (36 HIs & 36 LOWs) were used as actual auditory or auditory-visual stimuli in the

[4] With an exception of one triplet (v/ð/b) which partially contrasts manners.
[5] The word with the lowest frequency count was chosen in general. However, equal distribution of each consonant was the priority and thus the second lowest word was also used occasionally.

experiment, while all 108 words were presented on the screen (in the form of triplets) as the forced-choice response options. By comparing the intelligibilities between the groups HI and LOW, we hoped to find out how much of an effect lexical frequency of the choices could have had.

The stimuli were recorded in a sound booth in the UCLA Phonetics Laboratory. The speaker producing the stimuli was a trained phonetician (a female who did not wear glasses). She was seated in front of a plain dark blue background, approximately 1.5 m away from a video camera (Sony DSR-PD150 DVCAM), which was placed on the table in the sound booth and fixed in location. The experimenter held up 72 word cards, and the speaker was asked to read each word aloud, three times, looking straight into the 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 middle tokens were usually chosen from the three repetitions unless there were problems such as blinking or coughing. After editing, the movie clips were saved as QuickTime files, and the audio tracks were extracted from the movie files into audio files (SoundEdit) at a sampling rate of 22000 Hz. Audio files were extracted from already-edited video files, 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 (white noise), band-pass filtered at frequency between 200-6500 Hz using Kay Elemetrics' MultiSpeech. This noise spectrum was chosen from among those used by Miller & Nicely (1955) because it seemed to induce confusions related to place of articulation in their

study, and it also covers the frequency of the segments of interest in this study, except for /s/. 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 S/N range was chosen because, in preliminary tests, it appeared to produce a performance range from near chance to near perfect. The same range was also used in Wang and Bilger (1973).

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 (Sennheiser HD280). The visual stimuli were presented on a 20-inch computer screen (resolution: 1152 x 870, 75Hz). The size of video clips was 9.5" x 6.5".

2.3 Procedure

The stimuli were presented using Psyscope 1.2.5 (Cohen, et al., 1993). 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. After a brief explanation of the procedure by an experimenter, subjects went through a practice session consisting of five trials prior to the main session. 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 that made up a minimal triplet (as in Table 1) on a screen, and to press the button which corresponded to the word they think 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 (Figure 1), 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.

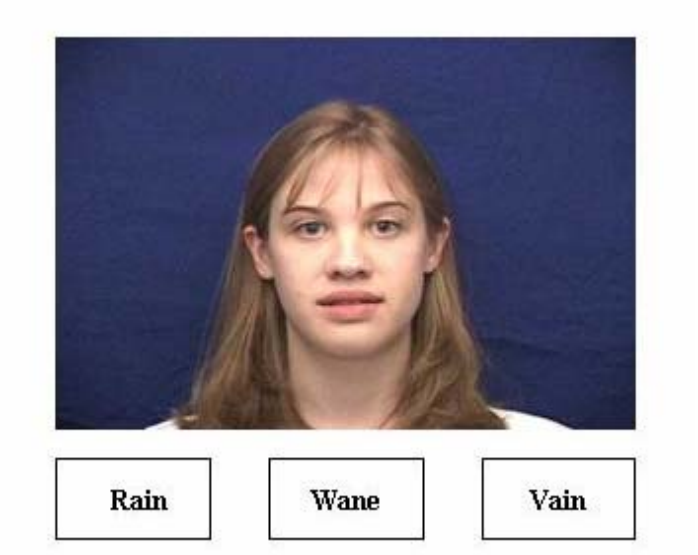

Figure 1 : An example stimulus presentation of the AV block. Subjects were asked to press the button corresponding to what they heard.

The auditory stimuli were arranged such that the correct response was equally distributed across the three buttons. In order to avoid any possible order effects, the presentation order of the two blocks was varied across subjects. 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), 5) lexical frequency, 6) order of A/AV blocks presentation, and 7) word list. The effect of each of these variables on the correct response rate was determined by cross-section time-series logistic regression (Hardin and Hilbe, 2003). Although repeated measures ANOVA is a much-used statistical analysis method in speech-perception studies, this method is not ideal for handling missing data[6], nor is it appropriate for studies with both a large number of independent variables and a small number of subjects. Table 2 summarizes the logistic regression results for each independent variable. The asterisks in the second column mark the independent variables on which the subjects' responses depend in a statistically significant way. In the following, I will present the results for each independent variable in detail.

Since the main question of interest is the additional information provided by visual presentation, the analyses concern only items which were degraded by auditory noise; items which were not degraded by noise cannot contribute to answering the question. In fact, it was found that /s/, as well as /v/ when paired with /r/ and /w/, were identified perfectly both with and without visual information; for this reason, the data for these sounds were excluded from the analysis. In fact, they were not expected to be acoustically confusable within a triplet; they were included in the experiment for the sake of forming triplets.

[6] Due to technical problems during the sessions, a few subjects did not complete the AV block.

Test		Wald chi2	DF	Log likelihood	Prob > chi2
segment class (overall)	*	598.24	14	-2278.7657	<0.001
segment class (A)	*	262.88	14	-1541.539	<0.001
segment class (AV) (not included; see below)		-	14	-319.05215	-
segment class (overall) (with 5 errors added)	*	573.83	14	-2400.1197	<0.001
segment class (A) (with 5 errors added)	*	262.88	14	-1541.539	<0.001
segment class (AV) (with 5 errors added)	*	397.07	14	-560.74805	<0.001
S/N (overall)	*	190.07	1	-2527.782	<0.001
S/N (A)	*	194.96	1	-1592.5069	<0.001
S/N (AV)	*	8.57	1	-735.71645	=0.0034
vision	*	316.45	1	-2445.1647	<0.001
lexical frequency (overall)	*	10.78	1	-2626.9135	=0.0010
lexical frequency (A)		0.35	1	-1639.7285	=0.5523
lexical frequency (AV)	*	36.42	1	-680.20574	<0.001
easy/hard (overall)	*	100.19	1	-2583.9253	<0.001
easy/hard (A)		0.48	1	-1699.4313	=0.4888
easy/hard (AV)	*	233.26	1	-573.55218	<0.001

Table 2: Results of cross-sectional logistic regression (by Stata).

3.1 Signal-to-Noise Ratio (S/N)

Figure 2 shows the speech intelligibility under A and AV conditions as a function of (auditory) S/N ratio. The S/N ratio is plotted along the horizontal axis, while the fraction of correct responses is plotted along the vertical axis. With three response choices, chance performance is .33. 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 previous studies (e.g. Sumby and Pollack, 1954).

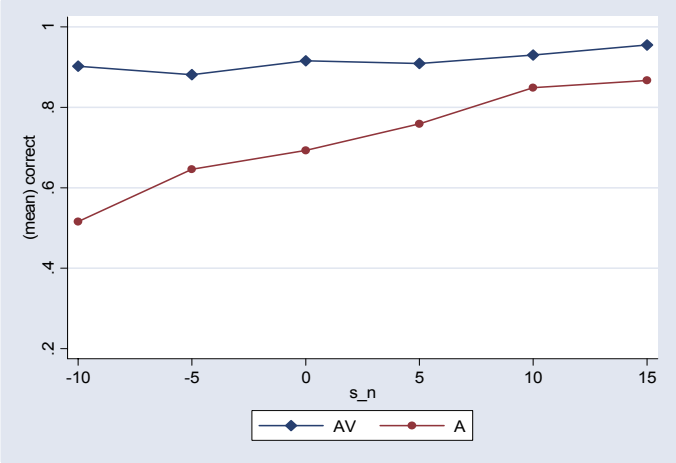

Figure 2: Speech intelligibility under AV (audio-visual) and A (auditory only) conditions as a function of the S/N ratio.

3.2 Presence of Visual Information

The effect of the presence of visual information on speech intelligibility was found to be statistically significant (Wald $\chi^2(1) = 316.45$, $p < 0.001$, see 'vision' in Table 2). The presence of visual information increases the overall intelligibility of speech perception (see Figure 2). This result agrees with previous studies (e.g. Sumby and Pollack, 1954).

3.3 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 $\chi^2(14) = 397.07$, $p < 0.001$). Due to the near perfect performance, it was not possible to obtain a Wald chi square for the AV condition using the raw scores. Five errors were therefore manually added to the score for each segment in order to test significance. This modification

made the data more uniform and less variable, and thus the significant result obtained here is considered to be genuine.

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 (Miller and Nicely 1954, and others). The locations of the AV curves do not vary as much, however, and appear to exhibit a ceiling effect in many cases. Given that most segments tested in this study have contrastive places of articulation from their counterparts in the triplets, and that they are supposed to be easy to lip-read, this ceiling effect in the AV condition was partially expected. However, the same ceiling effect was observed for those segments that contrast manner of articulation (i.e., /tʃ/, /w/) in the triplets. The result for /tʃ/ was particularly unexpected because according to the data from Jiang (2003), the correct response rates for /tʃ/ and its counterpart /ʃ/ were equally low, while that for /w/ was much higher than its counterpart /r/.

Figure 4 shows the visual gain (proportion correct AV– proportion correct A) across the 15 segments. The segment class is plotted along the horizontal axis, while the difference between the AV and A curves from Figure 2 is plotted along the vertical axis. Each data point represents an average across all S/N ratios. Note that there are two segments that show a mean difference below zero, namely, /r/ and /ʃ/. 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, and thus a relatively small visual contribution was expected. However, this negative contribution was not expected at all since no previous study has shown this type of visual effect. This result will be further discussed below.

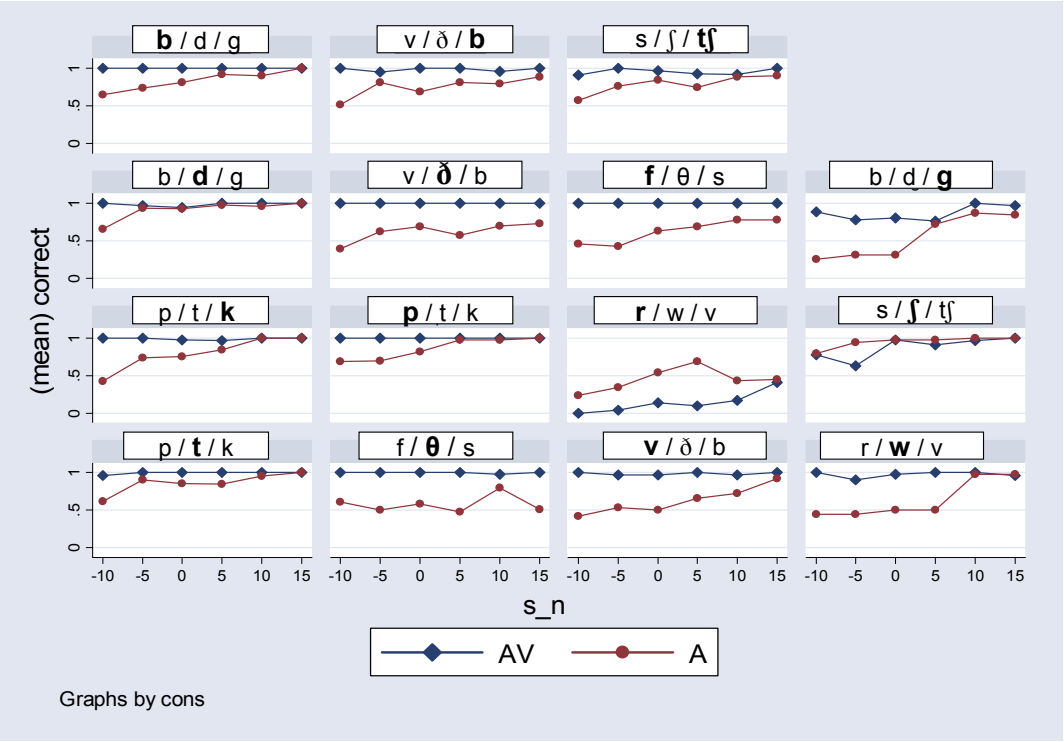

Figure 3: Speech intelligibility under AV (audio-visual) and A (auditory only) conditions as a function of S/N ratio. Bold face consonants were presented as stimuli.

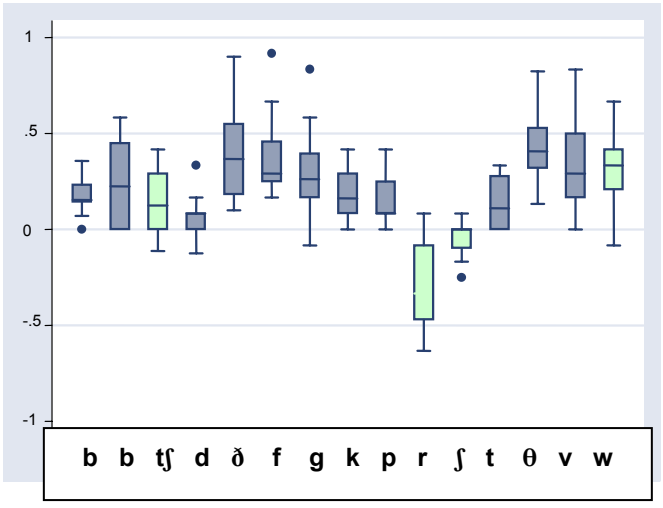

Figure 4: Visual gain across segments with SD. The bars in dark gray represent the Easy category, and the bars in light gray represent the Hard category.

The results in this study also indicate that there is a significant qualitative difference between auditory only and audio-visual speech perception. Figures 5 and 6 show the rate of intelligibility including mistakes as a function of segments under A and AV conditions, respectively. Note that in Figure 5, both possible mistakes are made for most segments. That is, when subjects are wrong, they choose both available wrong responses in a triplet. However, as can be seen from Figure 6, there are almost no mistakes made in the AV condition, and in the few cases where mistakes were made, only one of the available alternatives was favored. This is because the subjects made almost no substitutions of sounds that involve visually salient segments such as labials or labio-dentals. This result will be further discussed below.

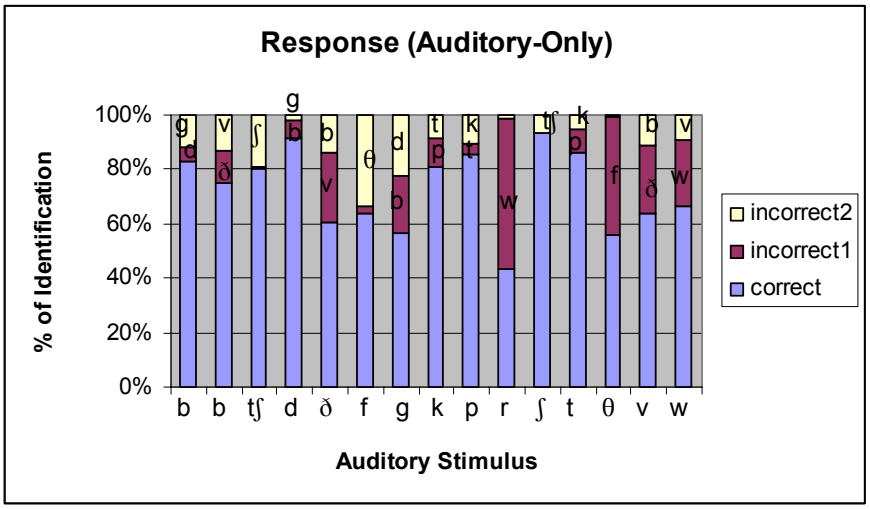

Figure 5: Ratio of correct and incorrect responses as a function of segment under A condition.

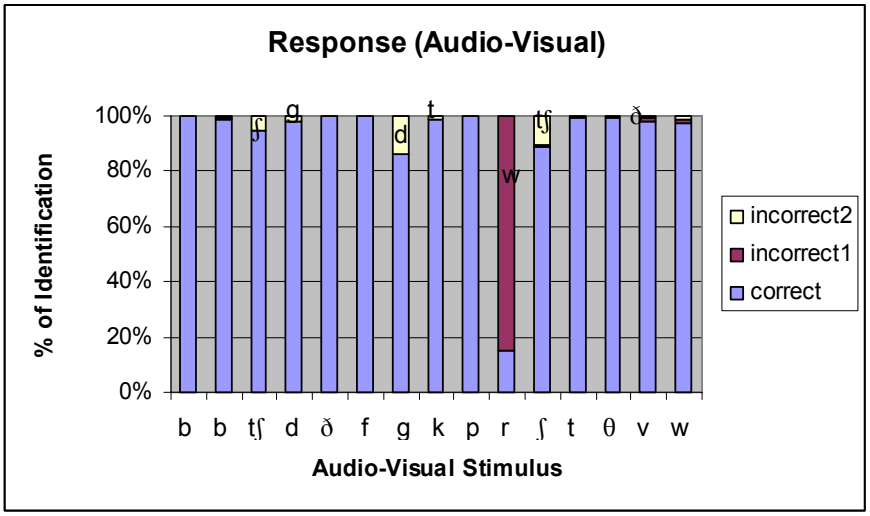

Figure 6: Ratio of correct and incorrect responses as a function of segment under AV condition.

3.4 Easy/Hard

As discussed in the method section, the 15 consonants were classified into two groups (Easy & Hard) according to their visual confusability: Easy {p/t/k, b/d/g, f/θ, v/ð/b}, Hard {ʃ/tʃ/s, r/w/v}. The segments in the Easy group are expected to be easy to distinguish visually, and contrast with their counterparts in place of articulation (e.g., /p,t,k/), while the segments in the Hard group are expected to be difficult to distinguish visually, and contrast in manner of articulation (e.g.,/w,r/). In Figure 4, the dark gray bars represent the Easy group, and the light gray bars represent the Hard group. 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. Figure 7 shows the intelligibility rate as a function of Easy/Hard category, averaged over all the consonants. There is no significant difference between the categories under the A condition, while there is more than a

25% increase of intelligibility for the Easy group under AV condition. Note that among the 4 consonants in the Hard group, the intelligibility for /tʃ/ and /w/ increased (see Figure 4), while it decreased for /ʃ/ and /r/.

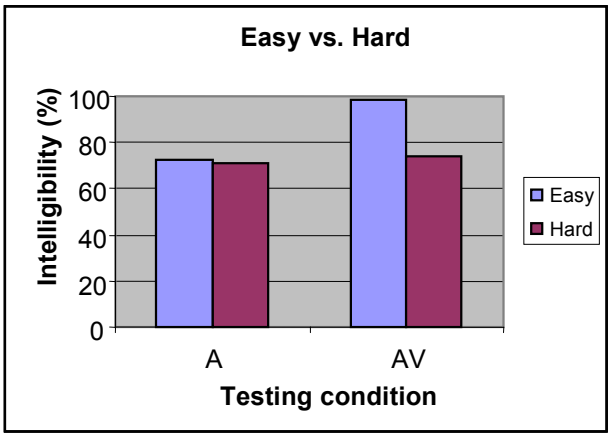

Figure 7: Speech intelligibility by two conditions (A & AV) and Easy/Hard consonants.

3.5 Lexical frequency

In terms of correct response, no significant effect of lexical frequency was found under the A condition, while a small effect was found under the AV condition in the opposite direction from our prediction (A: Wald $\chi^2(1) = 0.35$, $p = 0.5523$, AV: Wald $\chi^2(1) = 36.42$, $p < 0.001$). That is, the intelligibility rates for HI group words (which have the highest lexical frequency count in their minimal triplet) were slightly lower than for LOW group words (which have a relatively lower lexical frequency count) in the AV condition. In addition, contrary to results from previous studies, slightly longer RTs for group HI (high frequency words) were found for both A and AV conditions, although the difference in RT was not significant according to t-tests (assuming unequal variance, two tailed, $a = 0.05$).

Although this unexpected result is very interesting, the primary focus of this study is to look at the visual effect of segment class and Easy/Hard categories on intelligibility, and lexical

frequency was controlled in order to eliminate possible biases on intelligibility. We believe the bias we found here was successfully controlled by mixing the same number of high and low (relative) frequency words, and thus this result will not be discussed further.

3.6 Other

The order of A vs. AV presentation and two kinds of word lists were controlled to rule out any possible biases. There was no significant effect found for these factors (presentation order and word list), and these results will not be discussed further.

3.7 Interim Summary

The purpose of this study was to investigate segmental differences in the visual contribution to speech perception. Identification accuracy of 15 English consonants was measured in both auditory-visual (AV) and auditory-only (A) conditions with varied S/N ratios, and the difference between the two conditions was compared across the consonants to test the following hypotheses (repeated for convenience):

1. Consonants differ in their visual contribution to audio-visual speech intelligibility.
2. Visually distinctive consonants show greater visual contribution to speech intelligibility than visually confusable consonants.
3. The presence of visual information always increases speech intelligibility regardless of its visual salience.
4. The visual contribution to audio-visual speech perception increases as the signal-to-noise ratio decreases.

The results revealed a significant difference between the consonants in their visual contribution to speech intelligibility, supporting hypotheses 1 and 2. Further, the results also revealed that although the visual contribution is mostly positive, it can be negative for a few segments, refuting hypothesis 3. A greater visual contribution was found for lower S/N ratio settings as in Sumby and Pollack (1954), supporting hypothesis 4.

4 Discussion

4.1 Segmental difference

Our results show that the contribution of visual cues to auditory-visual perception differs significantly across segments. The members of the Easy group are the segments which contrast with their counterparts within a triplet in place of articulation, and are therefore expected to have salient visual cues. Indeed, 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/, /ð/) improved more than the ones with relatively salient acoustic cues (such as /d/). Also as expected, those segments which contrast in manner of articulation with their counterparts (= poor visual cues) (such as /r/, /ʃ/, /tʃ/) showed a very small (or even negative) visual contribution to speech intelligibility (see Figure 2 & 3). 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. The segmental difference we found in this

study might explain the large variation of threshold reduction (= visual contribution) in previous audio-visual perception studies discussed in the introduction.

This segmental difference might also explain other variations found on the McGurk effect literature. Sekiyama and Tohkura (1991 and 1993) 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, Massaro *et al.* (1993) examined the effect of language and culture (Japanese, Spanish, and English) on speech perception in face-to-face communication. They reported that there is no difference in the nature of processing across language groups. However, their data showed a relatively weaker effect of visual cues among Japanese speakers when compared to the other two language groups. If there is a cross-linguistic difference in audio-visual perception, where does it come from? Is it the cultural differences as the authors suggested?

The results from this study suggest another possible answer: 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, labio-dentals, interdentals) 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 for the speakers of 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.

As mentioned earlier, the results in this study also indicate that there is a significant qualitative difference between auditory only and audio-visual speech perception. Figures 5 and 6 show the rate of subjects' responses including mistakes as a function of segments under A and AV conditions, respectively. As can be seen in Figure 6, when the additional visual cues were available, the subjects made almost no substitutions of sounds that involve visually salient segments (i.e. labials and labio-dentals). In general, labial consonants have weaker acoustic cues due to lip closure compared to other places: for example, among the three places of articulation for English stops, labial is usually the least salient acoustically. This can be seen in spectrograms, acoustic confusion matrices, and also in the McGurk literature in which auditory /b/ or /v/ are most susceptible to vision-induced illusion (Massaro, 1998). However, with additional visual signals, the intelligibility of labials increases dramatically, and thus in face-to-face communication, the confusability/similarity of labials should be very different from that in auditory-only communication.

In the phonological literature, 'similarity' is often defined in terms of features. Frisch et al. (1997) defined similarity Sim (x,y) as:

$$\frac{\#\text{ of natural classes that include both x \& y}}{\#\text{ of natural classes that include x, y, or both.}}$$

This model is successful in accounting for some similarity-related phonological phenomena such as the Obligatory Contour Principle (OCP) effect. However, given that natural classes are based on articulatory or acoustical features, the output of this model is likely to describe the similarity for audio-only communication. Note that although some articulatory features do describe visual features such as lip rounding, they do not distinguish visible vs. invisible features, and thus the outcome of the model is destined to be unimodal. Therefore, if there is a difference in

'similarity' between face-to-face communication and audio-only communication as our results suggest, this model has no means for describing the difference.

4.2 Negative Effect

Perhaps the most striking finding of this study is that the presence of visual cues can actually decrease speech intelligibility (see 3.3.). We found one such case, /r/. According to the auditory confusion matrix in Luce (1986), 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 2). On the other hand, the visual confusion matrix by Jiang (2003) shows that /r/ was often mistaken as /w/ and /f/ (26%, 26%, and 35%, respectively) while /w/ was rarely perceived incorrectly (89% correct response, which is 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). Jiang also measured the facial movements of his four speakers, and used these measurements as quantified optical signals. According to his data, 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 (except Subject 12), as well as the subjects in Jiang's study, often respond with /w/ when /r/ was presented as the stimulus (see Figure 8)?

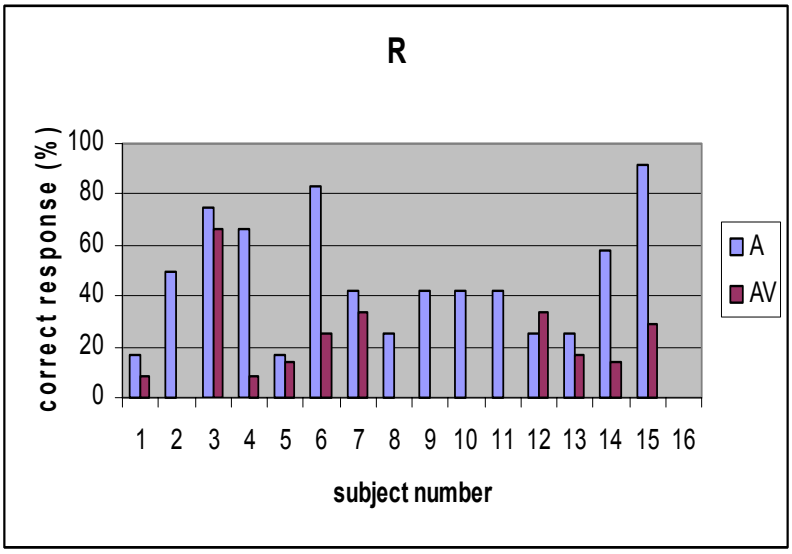

Figure 8: Correct response rate (all S/N collapsed) for /r/ across 16 subjects[7].

One possibility is that this effect in this study is due to the speaker: her speech actually exhibits very distinctive lip-rounding in general. Therefore it is likely that her /r/ involves more rounding than the average speaker's /r/. 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? Could it be regional (the speaker is from the Midwest) or random variation? According to the data from Jiang (2003), 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

[7] Note that the subject #16 had many fewer trials due to some technical problems during her session.

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 factor could be that the speaker in this study knew she was being recorded for a speech perception experiment: could it be that she exaggerated her rounding to enhance the phonetic properties of /r/? If so, then she did it not knowing that it would confuse the subjects. What does it say about the speaker's knowledge of 'careful speech', and listener-oriented speech? Given that Jiang's speakers also knew that they were being recorded, it is unlikely that the negative effect in this study is due to a hyperarticulation.

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, Jiang's physical measurements of speech production 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 Woodward and Barber (1960) and Walden *et al.* (1977)[8]. 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

[8] In their post-training data only. Their pre-training data show a {r,w,j} cluster.

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 Walden (1977) 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.

4.3 Asymmetrical Confusion

As discussed earlier, we found one case (/r/) in which the presence of visual cues actually decreased speech intelligibility. It was often mistaken as its counterpart in the triplet: /w/. The pair (/r/- /w/) appears to be mutually confusable according to the 3-D multidimensional scaling (MDS) plots from Jiang (2001 and 2003), and thus classified as a member of the Hard group. However, our data shows that /w/ was rarely mistaken as /r/, and its intelligibility increased significantly under the AV condition. This asymmetrical confusion (w>r) is in agreement with earlier visual-only studies such as Walden et al (1977) and Jiang (2003). In fact, the asymmetrical confusions found in the two studies are very similar; for example, there is a strong bias favoring voiceless over voiced consonants.

Many cases of similar asymmetrical confusion were also found in the A condition: /p/ was more often mistaken as /k/ than /k/ as /p/, and /θ/ was more often mistaken as /f/ than /f/ as /θ/. /b/ was more often mistaken as /g/ than as /d/, while /g/ was mistaken as /b/ and /d/ equally. What does this asymmetry mean when we talk about similarity between segments, especially

similarity derived from confusion matrices? Counting shared features (including as in the model by Frisch et al.) does not account for any asymmetry.

In search for an appropriate model to separate response bias and symmetric similarity in a confusion matrix, Goldstein (1977) examined various kinds of confusion asymmetries. He predicted that such biases can be shown to correlate either with lexical frequency or phonological naturalness, or both, although how to determine phonological naturalness is still an open question. However, lexical frequency was controlled in the current study. Although segmental frequency was not controlled, it is unlikely that it accounts for the asymmetry between /r/ and /w/. Ohala (1997) discussed the nature of asymmetrical confusion in terms of entropy – a state of greater randomness or noise. The natural course of events is for entropy to increase, so a change from high entropy to low entropy requires more energy (thus it is less likely) than the opposite change. He argues that each sound has a different level of entropy, and that this difference is the source of asymmetrical confusion. Estimating phonological naturalness or entropy of speech sounds is by no means an easy task, given that they should look very different for articulatory and acoustic-auditory domains, let alone the visual domain. However, further examination of asymmetrical confusions both in unimodal and bimodal speech perception might help us understand what it really means for a speech sound to be phonologically natural or to have high entropy.

4.4 Audio-visual integration and the Fuzzy Logical Model of Perception (FLMP)

In his well-known chapter in *Hearing by Eye* (Campbell and Dodd,1987), Summerfield stated the nature of audio-visual perceptual integration as “Information in the two modalities is integrated before phonetic or lexical categorization takes place; the two streams are analogue at

their conflux; categorization follows integration; in this respect, audio-visual speech perception is not a special mode," (p.16). There are some experimental results which support his pre-phonetic integration view. Green and Kuhl (1989) showed that a change in the perceived place of articulation resulting from the McGurk effect influenced the processing of VOT, indicating that auditory and visual information are combined by the time a phonetic decision is made (ruling out the possibility of post-phonetic integration). Results from recent fMRI studies (King and Calvert, 2001) also suggest that cross-modal interactions may be mediated at a relatively early level of processing.

Massaro (1998) examined several speech perception models by comparing the fit between responses predicted by the models and subjects' responses in McGurk type experiments. Subjects' responses were predicted more accurately by his Fuzzy Logical Model of Perception (FLMP), which perceives continuous values of auditory and visual features up to the point of integration, than by a model which made categorical decisions about the phoneme presented in each modality. According to FLMP, well-learned patterns are recognized in accordance with a general algorithm, regardless of modality or the particular nature of the patterns. The key assumptions to this model are: (1) each source of information is evaluated to give the degree to which that source specifies the relevant alternatives, (2) the sources of information are evaluated independently of one another, (3) the sources are integrated to provide an overall degree of support for each alternative, and (4) perceptual identification follows the relative degree of support among the alternatives. For bimodal trials, the predicted probability of a response P(/d/) is equal to:

$$\mathrm{P}(/\mathrm{d}/) = \frac{a_i v_j}{a_i v_j + (1 - a_i)(1 - v_j)}$$

where

a_i = the degree to which the auditory stimulus A_i supports the alternative */d/*
v_j = the degree to which the visual stimulus V_j supports the alternative */d/*

Massaro never applied FLMP to non-McGurk experiments where auditory and visual information are congruent. However, if the model represents the nature of bimodal speech perception more reliably than any other models, it should also be able to predict the results for congruent AV (or real-life) speech perception, for example, the results obtained in this study. In particular, the data for /r/ and /w/ can be fed into the equation to see if the model will predict a probability which is close to our result. In Massaro's study, the stimuli were synthesized varying one factor per modality (e.g. mouth opening for visual, and F2-F3 contour for auditory) and thus the evaluation of stimuli was done mechanically according to the factor's setting. However, there were no production measurements taken in the current study, and thus the correct response rate from our auditory-only trials (A) and Jiang's visual confusion matrices (V) are used to evaluate each segment.

In this study, the correct response rate for /r/ under the A condition was 50% at 0 S/N. According to Jiang (2003), the visual intelligibility rate for /r/ was 26%. Given that our subjects had to choose from three options (r,w,v), this rate should be normalized relative to the total correct response rate for the three segments. However, as mentioned earlier, there is a strong response bias favoring voiceless consonants in the visual confusion matrices, and thus there were more responses of /f/ than /v/. Given that /f/ and /v/ share the same place of articulation, the responses for /f/ were counted for /v/ in this normalization process, yielding 28% visual response rate for /r/. Using these independently evaluated numbers (a = 0.5, v = 0.28), the probability of audio-visual /r/ (P/r/) can be obtained as:

$$P/r/ = (0.5)(0.28) /((0.5)(0.28) + (1- 0.5)(1 - 0.28)) = 0.14 / 0.5 = 0.28.$$

The correct response rate for /w/ under the A condition was also 50% at 0 S/N. The visual response rate for /w/ after the normalization is 93 %.

$$P/w/ = (0.5)(0.93) / ((0.5)(0.93) + (1- 0.5)(1 - 0.93)) = 0.465 / 0.5 = 0.93$$

The actual correct response rates for /r/ and /w/ in the AV condition were 15% (lower than the predicted value) and 97.5% (very close to the predicted value), respectively. Given that the visual data used for the equation was from another study, it is plausible that if we had visual-only responses for our speaker's tokens, the predicted value might have been even closer to the actual value. This shows that FLMP successfully predicted the negative and positive contribution of additional visual cues, even when the data from perception experiments were used as input.

Investigating the nature of audio-visual integration is not the aim of this study, and our results do not provide any conclusive support for either pre-phonetic or post-phonetic integration views. However, simulations by the model which assumes pre-phonetic integration (FLMP) successfully predicted the results 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.

4.5 The Ceiling Effect

Overall intelligibility was much higher than expected in several cases, resulting in a ceiling effect in the AV condition. There seem to be at least two reasons for this result. First, the very small vocabulary size in this experiment (forced-choice from three options) seems to have contributed to the unexpectedly high performance in the A condition. Sumby & Pollack (1954) reported a strong effect of vocabulary size on speech intelligibility. In this study, the S/N ratio

range was chosen based on preliminary tests which involved open-choice responses. Lower S/N ratios could have been selected, and thus at least some of the ceiling effects would have been avoided. Secondly, among the 15 consonants tested in this study, 11 consonants (the Easy group) contrast with their counterparts in place of articulation (e.g., /p,t,k/). As discussed in the introduction, visual intelligibility of place features is known to be very high (e.g. Binnie et al.). Thus, given the nature of the experimental stimuli, high performance was expected for the AV condition (in particular, the Easy group) regardless of S/N ratio.

However, the cross-segmental difference found in this study should be valid for at least some segments. For example, the difference we found in the Hard group (particularly /r/ and /w/) must be fairly reliable given that the performance was not as high as for the Easy group. Also in the Easy group, comparing /t/ and /θ/, it is unlikely that /θ/ will become more auditorily intelligible than /t/ for lower S/N ratios. Regardless of S/N ratio settings, all the confusion matrices (except the matrices with low-pass filtered noise) reported in Miller and Nicely (1955) show higher confusability for /θ/ than /t/. On the other hand, the visual perception confusion matrices by Jiang (2003) show much higher visual intelligibility of /θ/ when it is compared to /t/. Therefore, even if a very low S/N ratio (such as -20 dB) had been used and there were no ceiling effects, the visual contribution for /θ/ would still have been expected to be greater than that for /t/.

On the other hand, there are cases where the ceiling effect seemed to be responsible for our result. For example, among the six stop consonants tested in this study (/p,t,k,b,d,g/), /g/ showed the greatest visual contribution. According to the data from Miller & Nicely (1955), in the same band-pass noise setting as this study (200-6500 cps.), /g/ show the lowest intelligibility among the 6 stops tested. Our data confirms this difference between /g/ and other stops in the

auditory-only setting. However, according to Jiang's visual perception confusion matrix, /g/ is expected to be the most confusable (=less intelligible) among the 6 stops. In particular, the difference in intelligibility between /b/ and /g/ is shown to be very large (correct response: 41% vs. 10%). Our data appears to have failed (due to the ceiling effect in the AV condition) to capture this difference. Therefore, the result of /g/ showing the greatest visual contribution among stops may not hold if the intelligibility was measured using lower S/N ratios or open-choice responses.

4.6 Comparison with Sumby and Pollack (1954)

One of the purposes of this study was to replicate the classic study by Sumby and Pollack (1954) (SP hereafter) using more up-to date methods, and to see if the same results can be obtained. The following are the changes made in the current study: 1)Video clips were used as visual stimuli instead of an actual speaker in front of subjects, so that every subject would see and hear exactly the same tokens across conditions. 2) Real words were used as in SP, although the lexical frequency was controlled in order to avoid possible frequency effects. This was possible by limiting the response to 3 options, and consequently, the chance level was increased to 33%. 3) Subjects' response was open-choice and handwritten in SP, while it was 3-option forced-choice and recorded through a button-box in this study. 4) S/N ratio was defined in a clearer manner: in SP, the signal level was monitored on a VU meter at a constant level by a test supervisor, and S/N ratio was varied in real time by holding the noise level constant and varying the speech level, resulting in absolute dB level variations across the S/N ratios. In the current study, the signal level was kept constant by normalizing all the audio files, and the various levels

of noise were added by a program specially designed for this purpose. All the audio files were then equated for the peak RMS amplitude at 80 dB.

The main findings in SP were (1) the visual contribution to oral speech intelligibility increases as the signal-to-noise ratio is decreased, while (2) its contribution relative to its possible contribution is independent of S/N ratio. Our result generally agrees with their finding (1), although it does not provide strong support for (2). Given that some features are more resistant in noise (e.g. voicing and nasality) than others (e.g. place), auditory intelligibility is dependent on S/N ratios in a *non-linear way*. On the other hand, visual intelligibility is independent of S/N ratio. Taken together, it is unlikely that the relative visual contribution to oral speech intelligibility is independent of S/N ratio.

5 Conclusion

Previous literature in audio-visual speech perception focuses on the overall increase between audio-only and audio-visual speech intelligibility, and little is known about the visual contribution for individual segments. 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 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. Although the segmental differences found in this study are reliable, the magnitude of the visual contribution found for some segments might not be accurate due to ceiling effects in the experiment as well as speaker-specific effect. Nonetheless, those 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.

References

Binnie, C.A., Montgomery, A.A., and Jackson, P.L. (1974). Auditory and visual contribution to the perception of consonants. Journal of Speech and Hearing Research, 17, 619-630.

Erber, N. (1969). Interaction of Audition and Vision in the Recognition of Oral Speech Stimuli. Journal of Speech and Hearing Research, 12, 423-425.

Erber, N. (1975). Audio-Visual Perception of Speech. Journal of Speech and Hearing Disorders, 40, 481-492.

Frisch, S., Broe, M., & Pierrehumbert, J. (1997). Similarity and phonotactics in Arabic. ROA-223, Rutgers Optimality Archive.

Goldstein, L. (1977). Bias and Asymmetry in Speech Perception. UCLA Working Papers in Phonetics, 39, 62-87.

Grant, K.W., and Seitz, P.F. (2000). The use of visible speech cues for improving auditory detection of spoken sentences. Journal of the Acoustical Society of America, 108, 1197-1208.

Green, K. P., & Kuhl, P. K. (1989). The role of visual information in the processing of place and manner features in speech perception. Perception & Psychophysics, 45, 34-42.

Hardin, W. and Hilbe, M. (2002). Generalized Estimating Equations. Chapman & Hall/CRC.

Jiang J., Alwan, A., Auer, E., and Bernstein, L. (2001). Predicting visual consonant perception from physical measures. Proc. Eurospeech'01, Aalborg, Denmark, 179-182.

Jiang J. (2003). Relating Optical Speech to Speech Acoustic and Visual Speech Perception. Ph.D. dissertation, Dept. of Electrical Engineering, UCLA.

King, A. J. and Calvert, G. A. (2001). Multisensory integration: Perceptual grouping by eye and ear. Current Biology, 11, 322-325.

Luce, P. A. (1986). Neighborhoods of words in the mental lexicon. Research on Speech Perception Progress Report, No. 6. Bloomington: Indiana University, Psychology Department, Speech Research Laboratory.

MacLeod, A. and Summerfield, Q. (1987). Quantifying the contribution of vision to speech perception in noise. British Journal of Audiology, 21, 131-141

MacLeod, A. and Summerfield, Q. (1990). A procedure for measuring auditory and audio-visual speech-reception thresholds for sentences in noise: rationale, evaluation, and recommendations for use. British Journal of Audiology, 24, 29-43

Massaro, D. W. (1998). Perceiving Talking Faces. MIT Press. Cambridge, MA.

Massaro, D. W., Tsuzaki, M., Cohen, M., Gesi, A., & Heridia, R. (1993). Bimodal speech perception: An examination across languages. Journal of Phonetics, 21, 445-478.

Massaro, D. W. and Cohen, M. M. (2000). Tests of auditory-visual integration efficiency within the framework of the fuzzy logical model of perception. J. Acoust. Soc. Am., 108, 784-789.

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

Miller, G. and Nicely, P. (1955). An Analysis of Perceptual Among Some English Consonants. Journal of the Acoustical Society of America, 27, 338-352.

Ohala, J.J. (1997). Comparison of speech sounds: distance vs. cost metrics. Speech Production and Language. Mouton de Gruyter. New York, NY.

O'Neill, J.J. (1954). Contribution of the visual components of oral symbols to speech comprehension. Journal of Speech and Hearing Disorders, 19, 429-439.

Rosen, S.M. and Corcoran, T. (1982). A video-recorded test of lipreading for British English. British Journal of Audiology, 16, 245-254.

Sekiyama, K., & Tohkura, Y. (1991). McGurk effect in non-English listeners: Few visual effects for Japanese subjects hearing Japanese syllables of high auditory intelligibility. Journal of the Acoustical Society of America, 90, 1797–1805.

Sekiyama, K., & Tohkura, Y. (1993). Inter-language differences in the influence of visual cues in speech perception. Journal of Phonetics, 21, 427–444.

Sumby, W. and Pollack, I. (1954). Visual Contribution to Speech Intelligibility in Noise. The Journal of the Acoustical Society of America, 26, 212-215.

Summerfield, Q. (1979). Use of Visual Information for Phonetic Perception. Phonetica, 36, 314-331.

Summerfield, Q. (1987). Some Preliminaries to a Comprehensive Account of Audio-visual Speech Perception, Hearing By Eye: The Psychology of Lip-Reading (ed. B. Dodd and R. Campbell), 3-51. Lawrence Erlbaum Associates, London.

Wang, M. and Bilger, R. (1973). Consonant confusion in noise: a study of perceptual features. The Journal of the Acoustical Society of America, 54, 1248-1266.

Walden, B.E., Prosek, R.A., Montgomery, A.A., Scherr, C.K., and Jones, C.J. (1977). Effect of training on the visual recognition of consonants. Journal of Speech and Hearing Research, 20,130-145.

Woodward, M. F. and Barber, C. G. (1960). Phoneme perception in lipreading. Journal of Speech and Hearing Research, 3, 212-222.

Appendix: Word List

p/t/k

1.	**pick**(3418)	tic(269)	ick(988)
2.	**pot**(657)	tot(33)	cot(406)
3.	pin(568)	**tin**(767)	kin(60)
4.	puff(239)	**tough**(751)	cuff(152)
5.	perk(47)	Turk(127)	**kirk**(195)
6.	pill(507)	till(1399)	**kill**(3835)

b/d/g

7.	**big**(7201)	dig(762)	gig(14)
8.	**birth**(1128)	dearth(16)	girth(27)
9.	bait(829)	**date**(1576)	gate(1236)
10.	bore(708)	**door**(6924)	gore(30)
11.	bowl(820)	dole(115)	**goal**(939)
12.	bun(117)	done(146)	**gun**(1771)

f/θ/s

13.	**four**(5873)	Thor(?)	ore(297)
14.	**feign**(68)	thane(2)	sane(149)
15.	fin(139)	**thin**(1556)	sin(699)
16.	fought(?)[9]	**thought**(3434)	sought(?)
17.	feign(68)	thane(2)	**sane**(149)
18.	fin(139)	thin(1556)	**sin**(699)

[9] fight(3216)

s/ʃ/tʃ

19.	**seat**(2280)	sheet(1037)	cheat(275)
20.	**suck**(672)	shuck(12)	chuck(88)
21.	sip(376)	**ship**(1378)	chip(374)
22.	sop(8)	**shop**(2427)	chop(451)
23.	seep(85)	sheep(718)	**cheap**(1262)
24.	Sue's(?)	shoes(1437)	**choose**(3211)

v/ð/b

25.	**vine**(111)	thine(30)	bine(2)
26.	**vow**(202)	thou(229)	bow(376)
27.	van(1034)	**than**(29721)	ban(556)
28.	vat(110)	**that**(217376)	bat(371)
29.	vow(202)	thou(229)	**bow**(376)
30.	v (?)	thee(?) **bee**(?)	

r/w/v

31.	**rain**(1559)	wane(97)	vain(242)
32.	**rail**(472)	wail(217)[10]	veil(194)
33.	reel(325)	**wheel**(958)	veal(88)
34.	Rhine(81)	**wine**(1421)	vine(142)
35.	rain(1559)	wane(97)	**vein**(269)
36.	rail(472)	wail[11] (217)	**veil**(194)

[10] wale(613)
[11] wale(613)