

Segmental differences in the visual contribution to speech intelligibility

Kuniko Nielsen

Department of Linguistics

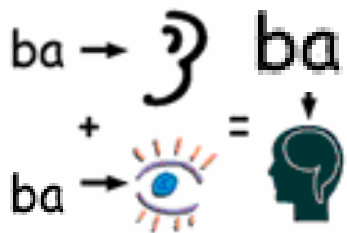
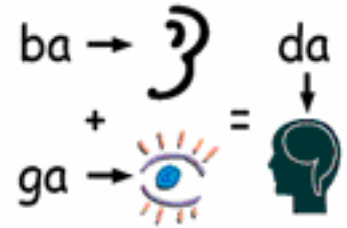
UCLA

kuniko@humnet.ucla.edu

<http://www.linguistics.ucla.edu/people/grads/kuniko/>

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**: (a.k.a. **McGurk Effect**)



- It **improves overall intelligibility** when the information from the two modalities are **congruent** (O'Neill 1954, Sumbly & Pollack 1954, Erber 1969).
- However, previous audio-visual studies focus on **overall intelligibility increase**, and not much is known about **individual segments**.

Background: Audio-Visual Speech Perception

- Sumbly and Pollack (1954)
 - Results: (1) the visual contribution to overall speech intelligibility increases as the signal-to-noise ratio decreases, (2) the proportion of actual contribution to its possible contribution is consistent across S/N. Best known as :

‘visual cues = +15 dB’

- However, there are many factors that influence the visual effect. In fact, various studies have shown different results (e.g., Erber 1969, Summerfield 1979, Binnie et al. 1974). There is variation even within an experiment (Rosenblum, et al. 1996). **What could be the source of this variation?**
- No previous audio-visual studies have controlled ‘segments’ in terms of their visual contribution. If there is a segmental difference, it could be another factor that affects audio-visual speech intelligibility...

Background: Segmental differences

- On the other hand, **segmental differences** have been studied extensively in both auditory and visual speech perception.
- Auditory Confusion Matrix
 - Miller and Nicely (1955)
 - differences in auditory confusion between consonants.
 - voicing and nasality are easy to perceive auditorily, while **place of articulation is the hardest**.
 - Prediction: additional **visual cues eliminate place confusions**.
- Visual Confusion Matrix
 - Jiang (2003)
 - differences in visual confusion between consonants.
 - **places of articulation are visually distinguishable**.
- Results from other studies are in agreement (e.g. Wang and Bilger 1973, Luce 1986, Binnie et al. 1974, Walden et al. 1977, etc.).

Background Summary & Motivation

- From the literature on audio-visual speech perception, we know that visual information improves **overall** intelligibility.
- From the literature on both auditory and visual (uni-modal) speech perception, we now know that :
 - Both Auditory perceptual intelligibility and Visual perceptual intelligibility differ across consonants.
 - Voicing and nasality are relatively easy to hear, while place of articulation is the hardest to hear, and
 - Places of articulation are easy to see, while voicing and nasality (or manner) are difficult to see.
- **Nevertheless, little is known about the combination of the two : segmental differences in the audio-visual contribution to speech intelligibility.**

Aims and Hypotheses

- Given that listeners seem to use the visual signal regardless of auditory signal quality, 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 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,
 3. whether the contribution of visual cues is always to increase speech intelligibility, and
 4. whether the visual contribution to audio-visual speech perception increases as the signal-to-noise ratio decreases (as in Sumbly and Pollack, 1954).

- Speech intelligibility with and without supplementary visual observation of the speaker's facial movements (**Auditory-only vs. Audio-Visual**) was tested across S/N ratios and segments.

Method

- Subjects: Sixteen native speakers of American English with normal hearing and vision: 11 females and 5 males (18 to 25 yrs).
- Stimuli: 108 English words that met the following criteria:
 - (1) monosyllabic (CVC)
 - (2) the initial consonant was one of the 15 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.
- The relative frequencies in the stimulus set were balanced.
- 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.
- Each triplet contains **auditorily confusable** consonants in noise, and contrasts either **place** or **manner of articulation**.
 - Example: **pot tot cot**

- The six triplet groups were classified into two groups according to the degree of **visual confusability**:
 - **Easy** group: [p-t-k], [b-d-g], [f-θ-s], [v-ð-b]
 - **Hard** group: [r-w-v], [s-ʃ-tʃ]

/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 darker blue, and LOWs are shown in lighter blue.

- The stimuli were digitally video-recorded by a phonetician.
- The video clips were edited for each word → the audio tracks were extracted from the video clips → the audio files were equated for the peak RMS amplitude at 80 dB.
- Six levels of noise (S/N ratios: -10, -5, 0, 5, 10, 15 dB) were mixed with the audio files.
 - Noise: flat shape, band-pass filtered at 200-6500 Hz.
- The audio files were added to the video clips.
- The audio files were presented as stimuli in an Audio-only block (A), and the video clips were presented as stimuli in an Audio-Visual block (AV). (Exactly the same audio for A and AV blocks).

Procedure

- The stimuli were presented using Psyscope 1.2.5.
- The audio channel was presented at 85 dB SPL.
- In the auditory-only (A) block, subjects were asked to listen to a word while they saw three response options on a screen, and to press the button which corresponded to the word they think they heard. (e.g., Rain Wane Vain)
- In the auditory-visual (AV) block, the same subjects were asked to watch and listen to the video clips of the speaker on the screen, and to press the button which corresponds to the word they thought they heard.
- The subjects were told to guess when unsure.

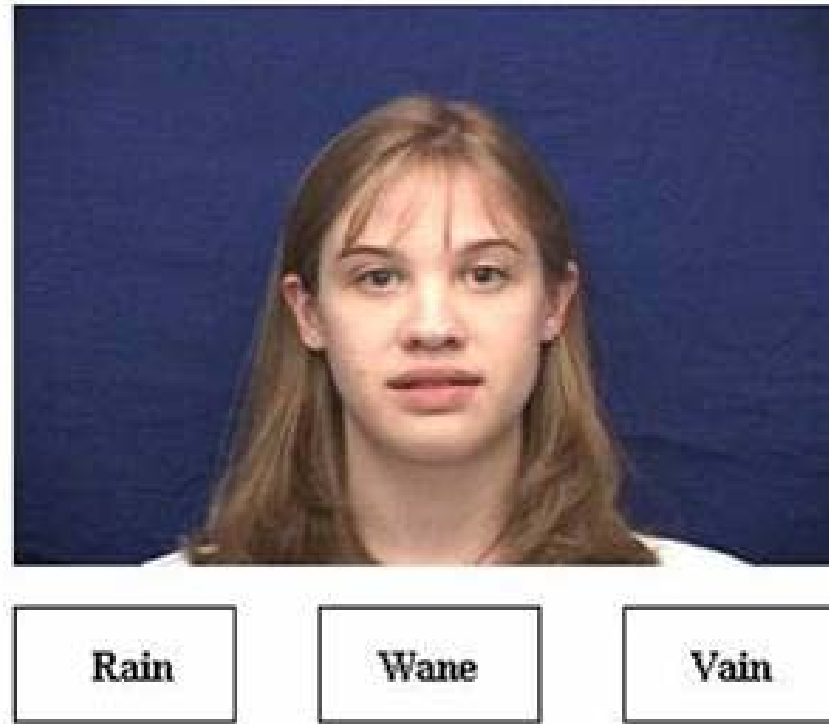


Figure 1 : An example stimulus presentation of the AV block. Subjects were asked to press the button corresponding to what they heard.

- 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.
- The order of blocks and the word lists within blocks were counterbalanced across subjects.

Results

The main effects on speech intelligibility:

- **Segment:** **significant** in both A and AV conditions.

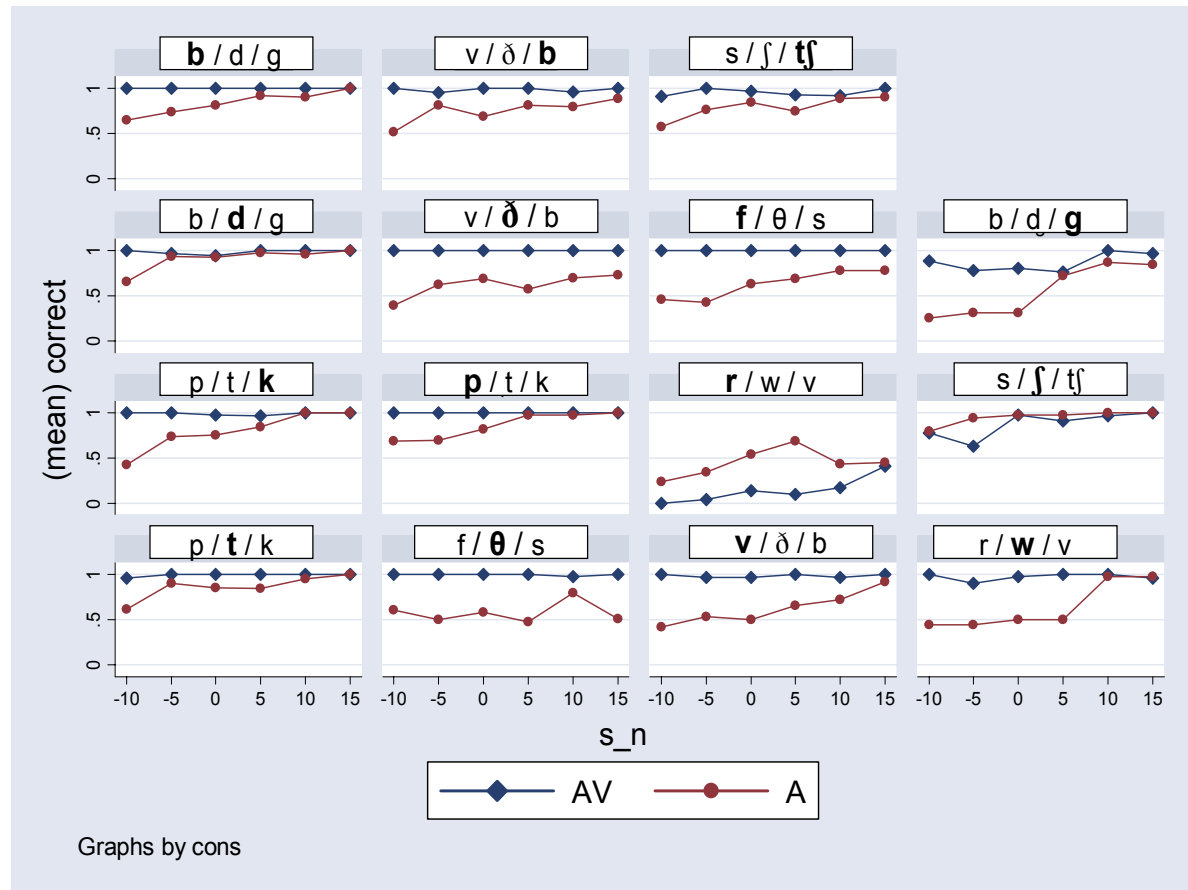


Figure 2: Speech intelligibility under AV and A conditions as a function of S/N ratio. Boldface consonants were presented as stimuli.

- **Easy/Hard** (visibility of cues) : **significant** in (AV), but **not significant** in (A)
 - This result was predicted given that this Easy/Hard classification was made solely based on visual confusability.

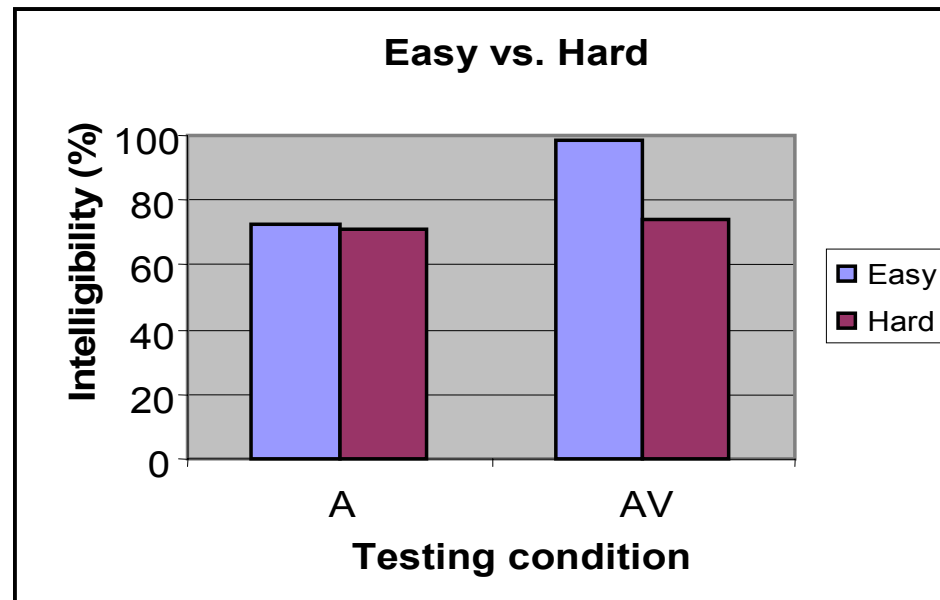


Figure 3:
Speech intelligibility by two conditions (A & AV) and Easy/Hard consonants, averaged over all the consonants.

- **S/N ratio**: **significant** in both conditions
- **A/AV** (Presence of visual information): **significant**
 - These two results confirm Sumbly and Pollack (1954).
- **HI/LOW** (Lexical Frequency): **not significant** (A), while **significant** (AV) in the opposite direction from our prediction.
 - The difference in RT was not significant according to t-tests (assuming unequal variance, two tailed, $\alpha = 0.05$).
- There was no significant effect found for the order of A vs. AV presentation and two kinds of word lists.

Results Summary

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

Discussion

(1) Segmental difference

- Our results show that the contribution of visual cues to auditory-visual perception differs significantly across segments.

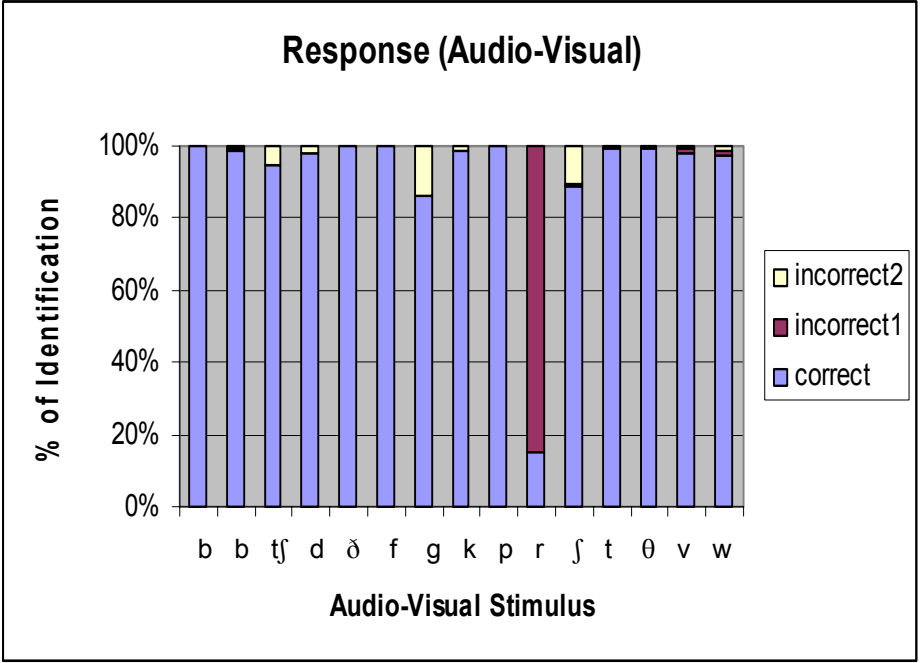
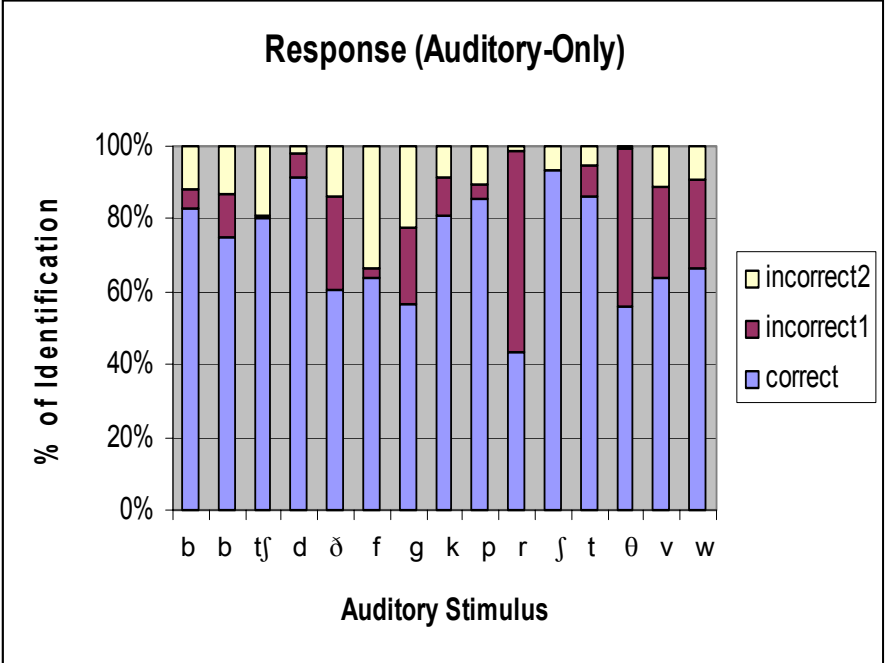
- The Easy group [p-t-k], [b-d-g], [f-θ-s], [v-ð-b] showed larger visual effects than the Hard group [r-w-v], [s-ʃ-tʃ]

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.

- This segmental difference might explain the variation of visual contribution in previous studies discussed in the introduction.¹⁷

- It might also explain other **cross-linguistic variation** found in the McGurk effect literature:
 - Sekiyama and Tohkura (1991 & 1993) as well as Massaro *et al.* (1993) reported a weaker McGurk effect among Japanese speakers than English speakers.
 - Every language has a different phoneme inventory, and some languages might have more visually distinctive segments.
 - 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 (such as Japanese) has fewer segments with strong visual contribution (e.g., /θ/) than a language (such as English) which has many of those segments, the **weight** of visual information in processing speech may be much smaller for the former (even if the nature of processing is the same for the two languages).

- The results in this study also indicate that there is a significant **qualitative difference** between auditory only and audio-visual speech perception.
 - In AV, the subjects made almost no substitutions of sounds that involve visually salient segments (i.e. labials and labio-dentals).



(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/. (The same effect was also found for /ʃ/, but it wasn't statistically significant).
- Luce (1986): the intelligibility of /r/ and /w/ are similarly low at -5 dB S/N and were both often misheard as /j/. Our results confirm this acoustic feature of the two consonants (see Figure 2).
- Jiang (2003): /r/ was often mistaken as /w/ and /f/ (26%, 26%, and 35%, respectively)
- On the other hand, /w/ was rarely perceived incorrectly (89% correct response: the highest among 23 consonants)

- Taken together, /r/ seems to have weak auditory cues and poorly utilized visual cues, while /w/ has weak auditory cues yet very strong visual cues (even more than other labials).
- Hypothesis: perceivers have little 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.
- 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 Walden (1977) which showed that /r/ was relatively **undefined** in the pre-lip-reading-training testing, and yet it demonstrated the largest training effect.

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

- Visual and auditory information are integrated “**before phonetic or lexical categorization takes place; the two streams are analogue at their conflux;**” (Summerfield, 1987)
- Some experimental results supporting **pre-phonetic** integration:
 - Green and Kuhl (1989) showed that a change in the perceived place of articulation resulting from the McGurk effect influenced the processing of VOT (ruling out the possibility of post-phonetic integration).
 - King and Calvert (2001) fMRI study also suggests 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.
 - For bimodal trials, the predicted probability of a response $P(/d/)$ is equal to:

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

- If the model represents the nature of bimodal speech perception, it should also be able to predict the results for congruent AV (or real-life) speech perception (i.e., the results in this study).
- In particular, we can feed our data for /r/ and /w/ into the equation to see if the model will predict probabilities which are close to our result (e.g. **our negative effect**).
 - Our data for (A) @ 0 dB S/N were used as ***a***. Since we do not have visual only data, the results from Jiang (2003) were used as ***v***:
- /r/ (actual result = **15%**)
 - $a = 0.5, v = 0.28$ ■ $P_{/r/} = \mathbf{0.28}$.
- /w/ (actual result = **97.5%**)
 - $a = 0.5, v = 0.93$ ■ $P_{a/w/} = \mathbf{0.93}$

- 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.
- 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.
- Although the current model **weighs** the input from two modalities equally, it may be able to predict the **cross-linguistic variation** reported in previous McGurk studies by assigning different weights to each modality.

(4) Comparison with Sumbly and Pollack

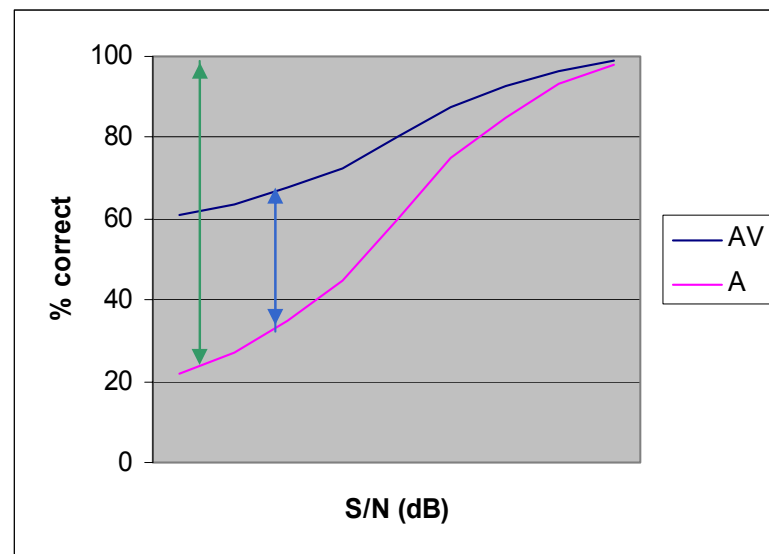
- S&P Main findings : (1) the visual contribution (**A**) increases as the S/N decreases, and (2) its contribution (**A**) relative to its possible contribution (**B**) is consistent across S/N ratio.

$$(R = A/B)$$

- Our result agrees with (1), although it does not provide support for (2).

- Given that some features are more resistant in noise (e.g. voicing) 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.



Sumbly & Pollack (1954)

Conclusion

- This study examined **segmental differences** in their visual contribution to speech intelligibility.
- As expected, there were **significant differences** in the visual contribution for the different consonants, with visual cues greatly improving speech intelligibility for most segments.
- Those segments with more salient visual cues (e.g. /f/, /θ/) displayed greater improvement.
- Surprisingly, the results also suggest that **the presence of visual cues can reduce intelligibility**.
- 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 the McGurk effect.

References

- Binnie, C.A., Montgomery, A.A., and Jackson, P.L. (1974). Auditory and visual contribution to the perception of consonants. *JSHR*, 17, 619-630.
- Erber, N. (1969). Interaction of Audition and Vision in the Recognition of Oral Speech Stimuli. *JSHR*, 12, 423-425.
- 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.
- 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.
- 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.
- 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. *JASA*, 27, 338-352.
- Rosenblum, L., Johnson, J., and Saldaña, H. (1996). Point-light facial displays enhance comprehension of speech in noise. *JSHR* 39(6), 1159-1170.
- Sekiyama, K., & Tohkura, Y. (1991). McGurk effect in non-English listeners: Few visual effects for Japanese subjects hearing Japanese syllables of high auditory intelligibility. *JASA*, 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. *JASA*, 26, 212-215.
- 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. *JASA*, 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. *JSHR*, 20, 130-145.