CGS603: Experiment Synopsis
McGurk Effect across Sexes

Gargi Singh 160259
Aviraj Mishra 150170
Mugdha Arora 150427
Mohit Duseja 150419
Gurpreet Singh 150259

1. Hypothesis

The McGurk effect has been consistently used as an argument for the involvement of visual stimuli in speech perception. However, the level of influence of this effect varies by factors such as gender, linguistic and cultural background, age, bilingual/monolingual etc. Most reports showing this difference have been inconsistent. We aim to show this difference between different sexes by using a strict inclusion criteria to avoid other factors introducing biases in our reportings.

In order to study the difference in the influence of the McGurk effect we formulate our hypothesis as given below.

Null Hypothesis There is no difference in the occurrence of McGurk effect across genders.

$$\pi_F - \pi_M = 0 \tag{1}$$

where π_F is defined as the proportion of **women** exhibiting McGurk effect and π_M is defined as the proportion of **men** exhibiting McGurk effect.

Alternate Hypothesis There is a difference in the occurrence of McGurk effect across genders with women exhibiting the effect more than men. Following the same convention as in 1.1, this translates to:

$$\pi_F \neq \pi_M \tag{2}$$

2. Literature Review

2.1. The McGurk Effect

The McGurk effect is a perceptual phenomenon that demonstrates an interaction between hearing and vision in speech perception. The illusion occurs when the auditory component of one sound is paired with the visual component of another sound, leading to the perception of a third sound.

The effect was first described in 1976 in a paper by Harry McGurk and John MacDonald, titled "Hearing Lips and Seeing Voices" (1) in Nature (23 Dec 1976). Speech perception was regarded as a purely auditory process before this paper, which confirms an observation that, on being shown a video clip, in which repeated utterances of the syllable [ba] had been dubbed on to lip movements for [ga], normal adults reported hearing [da]. With the reverse dubbing process, a majority reported hearing [bagba] or [gaba]. However, without visual input, the subjects reported the syllables accurately.

To further confirm and generalise the original observation, new materials were prepared. Videos and audios of four utterances: [ba-ba], [ka-ka], [pa-pa], [ga-ga] were mixed. A correct response was defined as an accurate repetition of the auditory component of each recording. Under the auditory-only condition accuracy was high, with averages of 91%, 97% and 99% for pre-school, school age and adult subjects respectively, while under the auditory-visual condition, where subjects heard the original soundtrack, errors were substantial. For pre-school subjects average error rate was 59%, for school children 52% and for adults 92%.

This paper showed that the contemporary, auditory-based theories of speech perception were inadequate to explain these new observations and there was a certain need to incorporate a role for vision i.e. perceived lip movements in the perception of speech.

2.2. Common Problems with analyzing the McGurk Effect

(2) point out some of the common problems faced across studies experimenting with the McGurk effect. These are highlighted in brief in this section. We aim to design our experiment taking into consideration such issues and suggestions mentioned by the authors in order to obtain significant and representative results.

2.2.1. Validating the Visual Influence

The original definition of the McGurk effect overlooked many deviations or instances in which the visual information overrides the auditory component in speech processing. For example, reporting 'ga' in response to $A_{ba} + V_{ga}$ and similar instances are overlooked as part of the McGurk effect.

More researchers have lately opted for a more liberal and lesser theory-bound scoring method. Later researchers defined the effect to be an instance of an incongruent visual speech resulting in an incorrect perception of the auditory information.

Secondly, it is difficult to validate the visual influence as it is non-trivial to attribute an incorrect perception directly to incongruent visual information as there might be other factors affecting the listener's speech perceptiveness. In order to avoid this, as suggested by (2), we ought to add baseline unimodal conditions to ensure that illusory reports are explained by genuine integration mechanisms.

Error-correction (2) suggest a corrective method to minimize such biases in the collected data by subtracting the total number of fused responses by the number of auditory misidentifications for each subject *i.e.*

$$\text{Illusory percepts} = \text{Fused percepts - Auditory Misidentifications}$$

We plan on using this correction while collecting data to ensure low biases while studying the McGurk effect.

2.2.2. Quality of the Auditory and Visual Information

Noise factors within the auditory signals such as release bursts, aspiration and voiced format transitions negatively affect the speech perception of the listeners. Similar factors which enhance the effects of incongruent visual information can induce bias within our data. Such issues have been prominent in many earlier studies, however it is not of much significance to us because of the advancements in technology since the time the original paper on the McGurk effect was published.

2.2.3. Clarity of the Task Instructions and Structure

Different task instructions can lead to subtle differences in the way subjects approach the task. An instruction that says "report what you heard" might potentially increase the perceptual weight of the auditory cue and thus lead to less illusory percepts. We aim to ensure presenting an objective instruction set saying "report what the talker said" to avoid such biases in collecting data.

The structure of the task can also pose a problem as there could be priming, selective adaptation, and/or visual recalibration that could potentially affect the categorization decision process. In order to avoid such problems, we need to choose an appropriate gap between two tests. There is a tradeoff as choosing a short gap can cause the aforementioned problems whereas a larger gap would hinder the data collection as the subjects would not be able to submit a lot

of responses.

2.2.4. Response Structure

There are two categorizations to the response structures – open set, in which subjects can respond with any syllable, and closed sets, in which subjects have to choose from specific response options. Many earlier experiments were conducted with closed set responses however, as (2) point out, elicit substantially more illusory responses. This variability was observed across studies and is believed to be because of the subject's innate tendency to equalize the number of responses between choices.

Moreover, closed set responses might not be able to capture the actual response a the perceptual utterance experienced by the listener. We plan on using open set responses to ensure no bias is induced within the perception of the subjects because of irrelevant factors.

2.2.5. Variability in Inner-Subject Perception

Many studies have reported varying scales of the McGurk effect based on natural factors such as linguistic and cultural background, gender, age, etc. Subjects belonging to one set of natural factors might be more susceptible to the McGurk effect.

For these reasons, we choose a strict inclusion criteria to avoid large variability within such factors. This inclusion criteria is discussed in the next section.

2.3. Variation across Sexes

A few studies have indicated that women are more influenced by visual information than are men when presented with incongruent auditory and visual McGurk stimuli. However, reports of sex difference in language processing are inconsistent and are thought to vary by task type and difficulty.

A number of fMRI studies have shown sex differences in the pattern of neurological activation in the brain for tasks that require phonological processing. It is believed that the differences emerge as more difficult tasks are encountered, with women performing better at them.

In their paper titled "A Sex Difference in Visual Influence on Heard Speech" (4), Fowler, Irwin and Whalen investigate this difference with brief (100msec) and full (temporally equivalent to the auditory) incongruent consonant-vowel stimuli. They report that females experience visually influenced percepts in 8.8% more instances than males for brief visual stimuli (considered the difficult task). There were no such significant effects for full video stimuli (considered the easier task). However, there were differences amongst what were considered easier and more difficult voices when it came to the brief visual stimuli.

3. Experiment Design

The experiment will be an attempt to replicate the work by Fowler et al. (4) to examine the influence of visual information on heard speech in male and female perceivers in setting of native Hindi speakers instead of native English speakers.

3.1. Sample

- **Sample Size:** Around 40 males and 40 females
- **Inclusion criteria:**
    1. Subjects should belong to the age group of 18-30 years
    2. Subjects should be native speakers of Hindi
    3. Subjects should be multilingual
    4. Subjects should have proper or corrected vision
    5. Subjects should have proper or corrected auditory sense

3.2. Stimuli

The subjects will be shown a series of speech stimuli (50 trials), each of which will be of four kinds:

1. has all visual frames of utterance of syllable
2. has three static visual frames of utterance of syllable with clearest indication; other frames are clear
3. has three dynamic visual frames of utterance of syllable with consonant closure and release; other frames are clear
4. has two dynamic visual frames of utterance of syllable made by deletion of middle frame of the three dynamic frames mentioned before

All these stimuli will be made with visual /ba/, /va/ and /ga/, and audio /ba/. The series will be a random combination of these.

3.3. Variables

- **Independent Variables:** Stimuli kind and gender

- **Dependent Variables:** Response for each trial
- **Control Variables:** Number of trials for each subject for both experiments and time for each trial

3.4. Division between Subjects

Since the trials are short we would not face the common issues usually faced while conducting within-subject experiments. Moreover, such a division would allow us to collect more data as compared to a between-subject division. Therefore, we opt a **within-subject division** and each subject will be shown trials of each stimulus.

3.5. Procedure

Two experiments will be performed on each subject, one with full stimuli (stimulus kind 1) only and the other with brief stimuli(stimulus kind 2, 3, 4) only. In both, same methodology will be followed.

We will design a webapp which would allow the subjects to mark their answers for corresponding trials among given options of '/ba/', '/da/', '/tha/' and 'others' (in order to ensure open-set responses, as mentioned earlier). They will be shown a series of audio and visual samples with a warning of "ready" before each stimuli to warn the person about the onset.

The task instruction for each stimulus will be "report what the talker said" and the subject will be asked to choose his response corresponding to the syllable perceived by clicking on one of the options displayed on the screen. The messages will be displayed for 1 and 5 seconds respectively. The participants will be instructed to pay attention to the screen for each trial and mark whatever sound they believe the talker uttered.

3.6. Data

The data will be collected into a database (MySQL or MongoDB) with each response submitted by the participants mentioning the sound they perceive for each trial. This would allow for easy data processing and analysis which is discussed in detail in the next section.

4. Data Analysis

The first step shall be to obtain the proportion values (π_F & π_M) for each of the experiment settings stated in 3.2. A subject will be considered a **non-perceiver** of the effect if he/she is incorrect less than 15% times in a trial of 50.

The following data analysis plan shall be used separately for each setting.

Significance value for determining whether to accept or reject H_0 is taken as **0.05**.

2 - tailed t-test is used to determine if the observed difference in proportions differ significantly from the hypothesized difference in proportions.

4.1. Computing P Value

We shall calculate the sample standard error as:

$$\sigma_{d_{sd}} = \sqrt{\frac{\pi_F(1-\pi_F)}{n_F} + \frac{\pi_M(1-\pi_M)}{n_M}} \tag{3}$$

where n_M and n_F are the number of observations for the corresponding genders.

Now let $\pi_d = \pi_F - \pi_M$, then our test statistic: t* will be calculated as:

$$t^* = \frac{(\pi_F - \pi_M) - 0}{\sigma_{d_{sd}}} = \frac{\pi_d}{\sigma_{d_{sd}}} \tag{4}$$

Now finally, P Value, the probability of observing a sample statistic as extreme as t^*, can be obtained from t-Table as the probability associated with t^* taking the degree of freedom as $(n_F + n_M - 2)$.

4.2. Interpreting the Results

Finally, we shall compare the P value of our test with the significance level (0.05).

If $P < 0.05$, we will reject our Null Hypothesis H_0 and accept that our Alternate hypothesis H_a is statistically significant.

Otherwise, we have failed to observe any difference in proportions and we shall accept the Null Hypothesis H_0.

5. Model to expain the McGurk Effect

The work by Magnotti et al. (5) describes McGurk effect by a noisy encoding of disparity. It determines the probability that the measured disparity for a given stimulus is below a participants threshold by:

$$p(x < T_j | D_i) = \int_{-\infty}^{T_j} N(x; D_i, \sigma_j) dx \tag{5}$$

where N is the Normal (Gaussian) distribution with mean D_i and standard deviation σ_j for stimuli i and participant j. T_j is the measure of each participant's prior probability for fusing the auditory and visual components of the syllable. If the measured disparity is smaller than T_j for a participant j, the audio and visual are fused and not otherwise. The threshold T_j can be modelled to accommodate attribute of gender if alternate hypothesis is accepted to predict difference in McGurk effect across genders.

6. Possible Confounds

- The sample in the original experiment consisted of native English speakers unlike in the proposed experiment design. The occurrence of McGurk effect might differ because of different linguistic nativity.
- The sample in the proposed experiment design will consist of multilingual people but the original experiment does not account for multi-lingual capabilities and corresponding effect on results.
- Some people included in the sample might find it relatively difficult to access short visual or audio signals.
- The number of trials have been reduced to 50 per person as opposed to 70 and 90 respectively for the two experiments in the referred paper. This might lessen deliberate answers but also might interfere with assessment of perceptual capabilities.
- Age might not just be a number when assessing perceptual capabilities, the age of the participants might affect results.

References

[1] Harry McGurk, John MacDonald: Hearing Lips and Seeing Voices [Nature]

[2] Agnès Alsius, Martin Paré and Kevin G. Munhall: Forty Years After Hearing Lips and Seeing Voices: the McGurk Effect Revisited https://doi.org/10.1163/22134808-00002565

[3] John MacDonald: Hearing Lips and Seeing Voices: the Origins and Development of the McGurk Effect and Reections on Audio-Visual Speech Perception Over the Last 40 Years https://doi.org/10.1163/22134808-00002548

[4] Irwin, J.R., Whalen, D.H. Fowler, C.A. Perception Psychophysics (2006) 68: 582 https://doi.org/10.3758/BF03208760

[5] John F. Magnotti, Michael S. Beauchamp, The Noisy Encoding of Disparity Model of the McGurk Effect (2015) https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4370809/