Journal of Language
Volume 3, Number 2, pp: 95-102, November 2021
e-ISSN: 2685-8878 | p-ISSN: 2655-9080

PERCEPTION OF SPEECH USING AUDIO VISUAL AND REPLICA FOR STUDENTS OF SULTAN AGENG TIRTAYASA UNIVERSITY

Ediwarman Ediwarman[1], Syafrizal Syafrizal[2], John Pahamzah[2]
Indonesian Department, University of Sultan Ageng Tirtayasa[1]
English Department, University of Sultan Ageng Tirtayasa[2]
e-mail: syafrizal@untirta.ac.id

Received: 2021-04-20 Accepted: 2021-11-03

Abstract

This paper exmined the perception of speech using audio visual and replica for students of Sultan Ageng Tirtayasa Univesity. This research was aimed at discussing face-to-face conversation or speech felt by the ears and eyes. The prerequisites for audio-visual perception of speech by using ambiguous perceptual sine wave replicas of natural speech as auditory stimuli are studied in details. When the subjects were unaware that auditory stimuli were speech, they only showed a negligible integration of auditory and visual stimuli. The same subjects learn to feel the same auditory stimuli as speech; they integrate auditory and visual stimuli in the same way as natural speech. These research result suggests a special mode of perception of multisensory speech.

Keywords: Perception, Audio Visual, Replica, ambiguous perceptual, natural speech, multisensory speech

1. Introduction

Speech perception is whether speech is perceived as all other sounds or whether a special mechanism is responsible for encoding an acoustic signal into phonetic segments. "Speech mode" refers to a speech module that is structurally and functionally packaged selectively on articulatory movements, or a perceptual mode that focuses on cues. Phonetics in speech signals.

Interesting demonstration of speech mode uses a sine wave speech replica that varies over time. A stimulus wave consists of a sine wave positioned at the center of the three or four lowest formant frequencies, namely the vocal tract resonance of natural speech. The resulting sine wave replica lacks all of the other cues typical for natural speech such as regular pulsations of the vocal cords, aperiodicities, and broadband formant structures. The naïve subject perceives wave stimulation primarily as a non-speech whistle, bleep or "computer sound". When other groups of subjects are instructed about the speech-like nature of the wave stimuli, they can easily assign the linguistic content to the same stimuli.

In face-to-face conversations, speech is received by the ears and eyes. Watching congruent articulatory movements improves the perception of degraded acoustic speech stimuli by presenting them in noise or by reducing them to replica sine waves. In some cases, observing a speaker's articulatory movements incompatible with acoustic speech

can alter auditory perception, even when the acoustic signal is clear. For example, when subjects see an articulating face / ga / and are simultaneously presented with acoustic / ba /, they usually hear / da /. This provides an example of multisensory integration in which the subject combines visual articulatory information with acoustic information unexpectedly at a high level of complexity. An example of non-speech is the audio-visual integration of the "plucks" of cello playing. This suggests that not only speech, but also ecologically valid combinations of auditory and visual stimuli can integrate in a complex manner. Although audio-visual speech perception has been suggested to provide evidence for specific modes of speech perception, to date there is no conclusive empirical evidence to suggest that this type of integration would be specific to speech.

2. Literature Review

This study examines whether the subject's expectations about the nature of auditory stimuli have an effect on audio-visual integration. Sine wave replicas of non-speech language are presented to the subject either alone or dubbed into a visual display of a congruent or inappropriate articulate face. In Experiment 1, in non-speech mode, subjects were trained to classify wave stimuli into two arbitrary categories and were not informed of their speech-like nature. In speech mode, the same subjects are trained to feel the same wave stimuli as speech. We studied whether subjects integrated acoustic and visual signals in the same way in these two modes of perception. Our hypothesis is that if audio-visual speech perception is specific, then integration will only occur when the subject perceives the wave stimuli as speech. In comparison, natural speech stimuli are also used. Subjects were asked to always report how they heard auditory and audio-visual only stimuli. Audio-visual integration is defined here as the amount of visual influence on auditory perception (Calvert, 2001) although we are aware that this definition may not apply if the integration mechanism is very high. Non-linear (Massaro, 2000) mention that performance is measured by calculating the percentage of correctly identified auditory portions of stimuli (hereafter "correct identification"). For unsuitable audio-visual stimuli, a low percentage of correct identification will indicate strong integration because integration will lead to illusory perception. Experiment 2 is designed to ensure that learning effects cannot explain the results of Experiment.

3. Research Method

The method used with the experiment, Experiment 1 sample of ten students of Sultan Ageng Tirtayasa University. All normal hearing is reported and vision is normal or corrected to normal. Neither of the subjects had previous experience with wave stimulation. Two subjects were excluded from the subject pool because they reported seeing wave stimuli as speech before being instructed about their speech-like traits.

1.1.2. Stimulus

Four auditory stimuli (natural / omso / and replica sine wave) and digital video clips of articulating / omso / and / onso / male faces were used. These stimuli were chosen because, for natural speech, inappropriate audio-visual combinations of / m / and / n / have been shown to produce such a powerful effect that the visual component modifies the perception of auditory speech. In addition, based on an informal pilot study, the inclusion of fricative / s / enhances the peculiarity of sine wave speech stimuli. The natural sound token produced by one of the authors was recorded in a silencer using condenser microphones and digital video cameras. The audio channel was

transferred to a microcomputer (digitized at 22,050 Hz, 16 bit resolution) and a sine wave replica of / omso / and / onso / was created by Praat software with a script provided by Chris Darwin (http://www.biols.susx .ac.uk / home / Chris_Darwin / Praatscripts / WAVE). The script creates a three-note stimulus by positioning a time-varying sine wave at the center frequency of the three lowest formants of the natural speech token.

Four audio-visual stimuli are created for natural speech and WAVE conditions by dubbing auditory stimuli to articulate faces using the video editing software FAST Studio Purple by replacing original acoustic speech with natural audio tracks or waves: two unedited / omso / congruent and / onso / stimuli in which the face and auditory signals are the same, and two incompatible stimuli, where auditory / onso / are dubbed visual / omso / and auditory / omso / are dubbed visual / onso /. In addition, for visual control tasks only, two visual stimuli of the articulating / omso / and / onso / voiceless faces were created.

1.1.3. Procedure

An experiment consists of six tasks which are always performed in the following order:

1. Training in non-speech mode. Subjects are taught to categorize two sine wave speech tokens into two non-speech categories without any knowledge of speech traits of sound. Subjects are told that they will hear two different auditory stimuli (it may sound strange). They were asked to press the button labeled "1" if they heard stimulus number one (sine wave replica / omso /), and "2" if they heard stimulus number two (sine wave replica / onso /). Both voices are played back several times and at each presentation the correct response code is demonstrated. When the subject feels that they have learned the correspondence, classification performance is tested by presenting both stimuli 10 times randomly. All learning subjects classify stimuli accurately.
2. Waves in non-speech mode. Self-presented or audio-visual waves with congruent or inappropriate visual articulation state that each stimulus was repeated 20 times. The subject's task is to focus on the mouth movements of the faces displayed on the computer screen and listen to what is playing on the loudspeaker. Subjects were never told that the mouth movements were actually articulatory movements, but were only told that they would be looking at a face with a moving mouth. They were instructed to indicate by pressing a button whether they heard a stimulus "1" or "2". After the test, the subjects were asked about the nature of the stimulation of the waves to find out if they spontaneously sensed the presence of phonetic elements in the stimuli of the waves. Two subjects reported hearing speech sounds / omso /, / onso / or / oiso /, and they were excluded from the subject pool.
3. Natural speech. The same tests as in the second task were performed but now natural auditory stimulation / onso / and / omso /. Subjects were asked to indicate using the keyboard whether the consonant they heard was / n /, / m / or something else.
4. Training in speech mode. Similar training sessions as in the first phase in non-speech mode were given but now the subjects (eight still under the impression that the wave stimuli are non-speech sounds) are taught to categorize the wave stimuli as / omso / and / onso /. Learning is tested by presenting both stimuli 10 times randomly. All subjects are studied to categorize the stimulus waves as / omso / and / onso /. They

were also asked how they heard the stimuli, and all reported that they now considered them to be speech sounds.

5. Wave is in talk mode. The same test as in the second task was given but the subjects responded as in the third task.
6. Only visual. Only articulate faces are presented with instructions to try to speak reading what the faces are saying. The number of alternative responses is not limited. As in assignments 3 and 5, / omso /, / onso / or "something else" is given as an example of a response.

4. Result and Discussion

Responses (percentage of correctly identified auditory portions of stimuli) were subjected to a two-way repeated measure of variance (ANOVA) analysis with two factors in the subject, Three-level conditions (waves in non-speech vs natural speech vs speech mode) and Type of stimulus with three levels (auditory only vs. congruent audio-visual vs. unsuitable audio-visual). The result is shown in Figure 1.

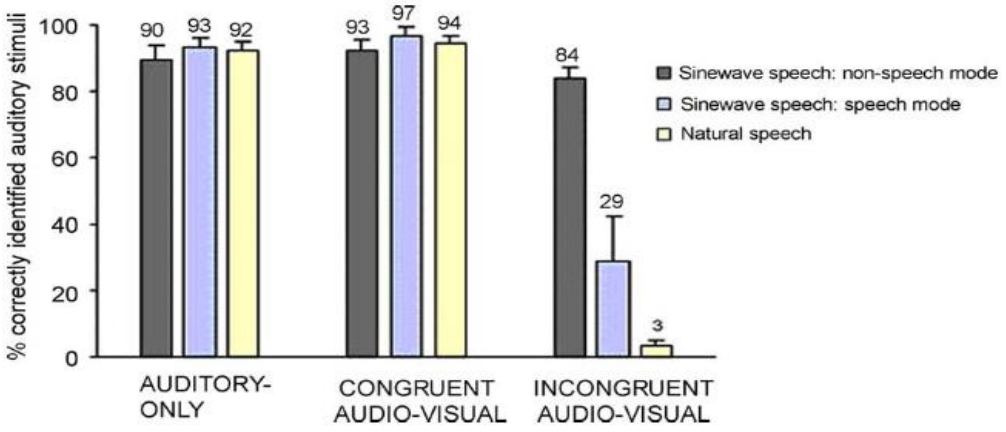


Figure1. Experiment 1

Picture1. Experiment 1: Percentage of correctly identified auditory stimuli (standard error C of the mean) for auditory-only stimuli, congruent audio-visual stimuli (visual / onso / C auditory / onso / and visual / auditory / omso / C auditory / omso /), and inappropriate audio-visual stimuli (visual / onso / C auditory / omso / and visual / omso / C auditory / onso /). The gray and light blue bars indicate wave identification in non-speech and speech mode respectively, and light yellow bar identification for natural speech. A low percentage of correct auditory identification with inappropriate audio-visual stimuli indicates strong audio-visual integration. revealed the main effects of both Conditions (F (2.14) Z12,922, P <0.001), due to higher correct identification scores for wave stimulation in non-speech mode, and Type of Stimulus (F (2.14) Z148, 959, P <0.001), due to lower identification scores for inappropriate stimuli, and the significant interaction of these factors (F (4.28) Z27.958, P <0.001). The significant interaction effect was followed up by performing one-way ANOVA separately for the condition and type of stimulus factors.

The results of the analysis showed no significant difference between conditions in the presentation of auditory-only and congruent stimuli (both F 1) but a significant main effect in incompatible stimuli (F (2.14) Z26,504, P> 0.001). The post hoc t-test shows that this effect is due to the fact that the identification performance with the

incompatible wave stimulation in the non-speech mode (84%) is significantly better than the wave in the speech mode (29%, t (7) Z4, 271, PZ0.004) and natural speech (3%, t (7) Z24.177, P 0.001). The identification scores for wave stimulation in speech and natural speech modes did not differ significantly from each other (t (7) Z1, 769, PZ0.120, n.s.). Separate comparisons of conditions across stimulus types reveal the main effect across all conditions (waves in non-speech mode: F (2.14) Z8, 739, PZ0.003; waves in speech mode: F (2.14) Z26, 285, P <0.001; natural speech: F (2.14) Z522.901, P> 0.001). In all conditions, the patterns were similar: identification of mismatches, but not of congruent stimuli, distinct from auditory-only baseline stimuli (all P <0.001 except for wave stimuli in non-speech mode, PZ0.012).

Thus, the results show that strong audio-visual integration effects occur only when auditory stimuli are perceived as speech. The integration effect was also observed in the non-speech mode, but its magnitude was minimal (decreased from 90 to 84%) compared to wave stimulation in speech mode (decreased from 93 to 29%) and natural stimulation (decreased from 92 to 3%).

Table Wave Mode

F	Nilai P
F (2,14)	P>0,001
F (2,14)	P>0,001
F (2,14)	P>0,001

4.1 Experiment

In Experiment 1, different tasks were always performed in the same order, so that the non-speech mode always preceded the speech mode for wave stimulation. The reason for this is that once the subject "enters speech mode" it is impossible to hear the wave stimuli as non-speech. However, this procedure may have created a learning effect with the subject becoming more accustomed to wave stimulation. Then at least part of the observed large integration effect with inappropriate stimuli could be due to this learning effect. To control for this, we presented new subjects with wave stimulation in speech mode as the first block, and reasoned that if the subjects showed comparable performance without prior extended exposure to WAVE stimuli, then a large integration effect could not be caused by learning. For comparison purposes, we also present natural speech stimuli.

4.2 Method

4.2.1. Subject

Thirteen students of Sultan Ageng Tirtayasa who did not participate in Experiment 1 were studied. All had normal hearing and vision normal or were corrected to normal. Neither of the subjects had previous experience with wave stimulation.

4.2.2. Stimulus

The same stimulus material was used as in Experiment 1.

4.2.3. Procedure

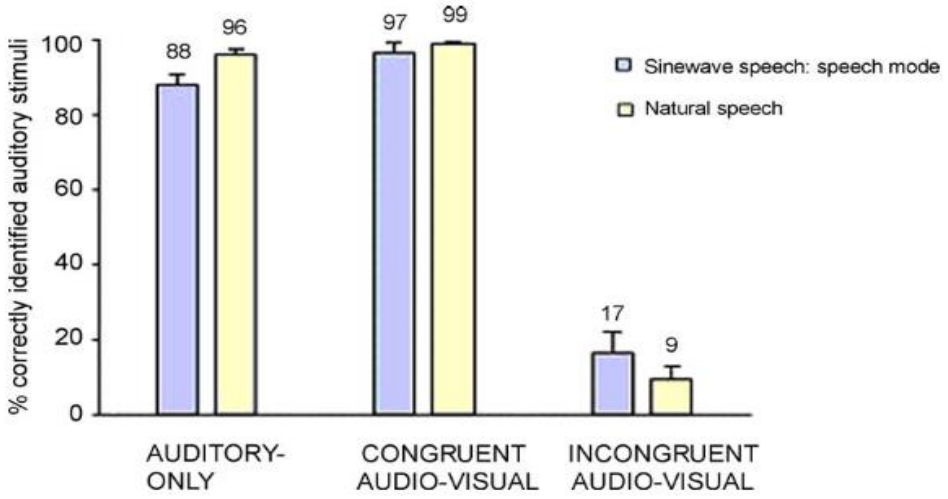
Experiments consist of four tasks with the same instructions as in Experiment 1. However, the order of the tasks is different from Experiment 1.The tasks are always performed in the following order:

1. Training in speech mode.
2. Wave is in talk mode.
3. Natural speech.
4. Only visual.

4.3 Result

Fig. 2 shows the results of Experiment 2 which replicate the findings of Experiment 1 that the waves in speech and natural speech modes provide the same and low number of auditory responses to inappropriate audio-visual stimuli, indicating strong and similar audiovisual integration.

Direct comparison of identification performance with wave stimulation in speech mode and with natural stimulation between Experiment 1 and Experiment 2 was carried out by performing three-way ANOVA with Experiment with 2 levels (first vs second) as a factor between subjects, and conditions with two levels (waves in mode natural speech vs. speech) and three levels of stimulus type (auditory only vs. congruent vs. out of sync) as a factor in the subject. The results showed the main effect of stimuli.



Type (F (2, 34) Z428, 273, P <0.001), due to lower identification score of inappropriate stimuli, and interaction between Stimulus Condition and Type (F (2, 34) Z8, 492, PZ 0.001) in the same way as in Experiment 1.Most importantly, there are no main effects of Condition (F (1, 19) Z2, 773, PZ0.112, ns) or Experiment (F! 1), and no interactions involving Experimental factors statistically significant. This pattern of results indicates that SWS stimulation in speech mode (and natural stimuli) was identified in the same way in Experiment 1 and Experiment 2.Thus, the large integration effect observed in Experiment 1 was not based on the learning effect due to the order in which the stimulus conditions were presented.

Our results show that acoustic and visual sound is strongly integrated only when the receiver interprets the acoustic stimuli as speech. If SWS stimuli are always processed in the same way, the effect of visual speech should be the same in both speech and non-speech modes. These results do not depend on the amount of exercise with listening to wave stimulation as confirmed by the results obtained in Experiment 2. We suggest that when wave stimulation is considered non-speech, acoustic and

5. Conclusion

Visual tokens do not form naturally multisensory objects, and are processed almost independently. When wave stimuli are perceived as speech, acoustic and visual signals are naturally combined to form phonetic perception. We interpret our current findings to be strong evidence for the existence of a specific audio-visual perception mode of speech. We have previously shown that visual speech has a greater influence on the perception of audio-visual speech when the subject pays attention to the speaking face. Here we propose that attention is also involved in the current case, albeit in a very different context. It has been suggested that attention can guide which features of the stimulus bind to the object during the perceptual process (Treisman & Gelade, 1980). Hence, depending on the perceptual mode, a different set of features may become the focus of attention. While in speech mode, attention may have increased the processing and binding of these features in our stimuli which form phonetic objects. When the same stimulus is considered non-speech, attention may have focused on some other feature (such as a certain frequency band containing prominent acoustic energy) that can be used to distinguish the stimulus. The features in the voice or face that are less important for speech perception are not expected to have a major influence on audio-visual speech perception and Hietanen, Manninen, Sams, and Surakka (2001) for the effects of speaker identity and facial configuration on speech perception, and Kamachi, Hill, Lander, and Vatikiotis-Bateson (2003) to show that speaker's identity can be extracted from visions and auditions by matching faces to sentences). Indeed, the difference between the spatial location of acoustic and visual speech has only a slight effect on the strength of the McGurk effect (Jones & Munhall, 2002), and the effect also occurs even when male voices are dubbed into female faces and vice versa. Speech will thus guide attention to the special features of speech in both auditory and visual stimuli, producing integration only when they provide coherent information about phonetic objects. Our account can be seen as an extension of the object-based theory of selective attention in vision to the multisensory domain. When a visual object is attended, the processing of all the features belonging to that object is enhanced, and this increase affects all areas of the brain where the relevant visual features are processed. In this experiment, when the subject perceives wave stimuli as speech, attention is focused on the phonetic object. The processing of phonetic objects in the auditory domain may have enhanced the processing of phonetically relevant visual features, resulting in strong audio-visual integration. It should be noted that we also observed a small integration effect in the non-speech mode, which is minutes in magnitude compared to the talk mode. One possible explanation is that the effect is due to the weak integration of the non-sound features of acoustic and visual stimuli. Features that can be integrated in non-speech mode can include the size of the mouth opening and the loudness of the auditory stimuli. In conclusion, our results support the existence of a special speech processing mode, which also operates in audio-visual speech perception. We suggest that an important component of the mode of speech is the selective and enhanced processing of these features in acoustic and visual stimuli relevant to phonetic perception. Selectivity and enhancement can be achieved through attention mechanisms.

References

Calvert, G. (2001). Cross-modal processing in the human brain: insights from functional neuroimaging studies. Cerebral Cortex, 11, 1110–1123.

Hietanen, J. K., Manninen, P., Sams, M., & Surakka, V. (2001). Does audiovisual speech perception use information about facial configuration? European Journal of Cognitive Psychology, 13, 395–407.

Jones, J. A., & Munhall, K. G. (2003). The effects of separating auditory and visual sources on audiovisual integration of speech. Canadian Acoustics, 25(4), 13–19.

Kamachi, M., Hill, H., Lander, K., & Vatikiotis-Bateson, E. (2003). 'Putting the face to the voice': matching identity across modality. Current Biology, 13, 1709–1714.

MacDonald, J., & McGurk, H. (2011). Visual influences on speech perception processes. Perception and Psychophysics, 24(3), 253–257.

Massaro, D. W. (1998). Perceiving talking faces. Cambridge, MA: MIT Press.

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

Treisman, Anne and Gelade, Garry. (1980). A feature-integration theory of attention. Cognitive Psychology, Vol. 12, No. 1, pp. 97-136.