Journal of Modern Education Review, ISSN 2155-7993, USA
June 2024, Volume 14, No. 5–6, pp. 171–174
Doi: 10.15341/jmer(2155-7993)/5-6.14.2024/003

http://www.academicstar.us

Description and Evaluation of Theory and Research in Speech Understanding

Lazaros Kostanasios
(Directorate of Primary Education of Ioannina, Greece)

Abstract: This work explores the theory and research on speech comprehension, emphasizing the importance of distinguishing speech sounds from noise. It discusses the perceptual processing of speech sounds, including phonemes and their categorical perception, and how variability in speech sounds and segmentation challenges are addressed. It evaluates the impact of linguistic, speaker, and listener characteristics, as well as environmental factors, on speech comprehension. The paper reviews the TRACE model, which integrates bottom-up and top-down processes in word recognition, and the McGurk effect, demonstrating the integration of auditory and visual information. It concludes with the significance of understanding speech for effective communication and the role of poly-sensory integration in speech perception.

Key words: theory, speech comprehension

1. Introduction

Saussure (1979) defines language as a convention that it is based on the ability of speech and which was defined to achieve communication between the social whole. Also, the execution of the faculty of speech in man is speech. Speech is a supra-individual product while speech is an individual realization. According to Tzouriadou (1995), discourse is a process useful for communication, while language is a system. With speech, the person comes into contact with and understands those around him, and they understand him accordingly, while speech is the embodiment of speech.

This paper presents the theory and research on speech understanding. It is talked about the properties of speech sounds, phonemes, categorical perception, variability of speech sounds and the problem of segmentation. Related studies as well as the McGurk effect are also evaluated.

2. Main Subject

It is important to separate what constitutes speech sound and what is simple noise. First, the speech sounds make sense and are easy to understand be recognized. Kellogg (2007) defines speech sounds as “Complex, information-rich auditory signals (that) we are perceptually aware of at an extremely rapid rate”. Their main characteristics are the wavelength and the frequency of the wave. Also, the human brain processes speech sounds differently from other sounds (mainly in the left hemisphere, categorical perception). Another key feature is

Lazaros Kostanasios, Educational Consultant, Directorate of Primary Education of Ioannina; research area: social science (linguistics). E-mail: kostanasio@gmail.com.

phonemes which are “the smallest unit of sound that makes a difference in meaning in a given language” Galloti (2008). Phonemes are very easy to recognize.

Regarding Repp’s (1984) categorical perception, phoneme perception does not faithfully reflect the properties of sounds but adapts to the distinctions between different phonemes of language. Thus, people perceive phonemes in a categorical way. Perception is what categorizes sounds, making similar sounds sound exactly the same and difficult to distinguish. This allows the perceptual system to correctly recognize vowels even when they have been roughly pronounced.

There are many factors that influence speech comprehension Alderson (2000). Linguistic factors such as co-articulation and assimilation where phonemes sound different depending on the sounds that precede and follow them due to the articulatory process. Also, dialects and accents as well as speaking with a foreign accent. Another factor is the characteristics of the speaker (age, gender, alcohol, tiredness, worry, joy, excitement, pronunciation errors, flow problems). In addition, the characteristics of the listener (language knowledge, spelling knowledge), but also the characteristics of the situation (noise from the environment, simultaneous conversations) play a big role.

In terms of the segmentation problem Pinker (1994)& Altmann (1997), speech is a continuous stream of sounds in which there are no clear boundaries between words, with individual phoneme sounds being they fall on top of each other. Thus, to decide where a word begins and ends in the speech flow, the listener uses top-down information that enables rapid analysis of phonemes and segmentation of speech into
a sequence of words. However, there may be ambiguities, such as the listener mishearing what the speaker said.

McClelland & Elman (1986), McClelland (1991) examined the “Trace” model, which emphasizes the role of context in word recognition. It includes bottom-up and top-down processes as all sources of information are used by the listener simultaneously. It is an interactive activation model and contains three levels: features, phonemes and words. It places the different kinds of processing (acoustic signal, phonemes, words) in isolated processing layers, allowing the activated units to communicate with the other layers, having competition between layers, until the “winner” is “recognized” by the model. There is a mental unit for each attribute, when units are activated beyond some threshold, then they can affect other units in two ways: Nodes that are compatible with each other share activation (increased activation of other units) and nodes that are incompatible with each other share inhibitory connections (decreased activation of other units). However, the activation of incoming information depends on whether and to what extent the same information is processed at the various stages.

The positives of the “Trace” model are that it manages to a satisfactory degree the existence of context, and the influence of other levels on recognition. It also partially explains perception ability given the lack of acoustic uniformity, as well as phoneme restoration phenomena in co-pronunciation. Additionally, in the model the existence of co-pronunciation aids recognition, as it does with human recognition and locates word boundaries and addresses the issue of word recognition in noisy environments.

As for negatives, sentences and expressions are not represented and thus the model cannot provide an explanation for the effects of propositional context. Furthermore, visual information (lip reading), general knowledge and orthography are not represented so that the important role of these factors in speech understanding cannot be explained by this model.

McGurk and MacDonald (1976) reported a strong multisensory illusion that occurs with audiovisual speech. They recorded a voice articulating one consonant and transcribed it with a face articulating another consonant. Although the auditory speech signal was well recognized on its own, it sounded like another consonant after

dubbing with mismatched visual speech. The illusion was called the McGurk effect.

It has been replicated many times and has sparked a wealth of research. The reason for the high impact is that it is an impressive display of multi-sensory integration. It shows that auditory and visual information are merged into a unified, integrated perception. It is a very useful research tool, as the strength of the McGurk effect can be considered to reflect the strength of audiovisual integration.

There are many variations of the phenomenon McGurk (MacDonald and McGurk, 1978). The best-known case is when dubbing a voice that says [b] to a person that articulates [g] results in hearing [d]. This is called the fusion effect. Many researchers have defined the McGurk effect exclusively as the fusion effect, because here integration results in the perception of a third consonant, apparently merging information from hearing and vision (Wassenhove, Grant & Poeppel, 2007. Keil, Muller, Ihssen& Weisz, 2012. Setti, Burke, Kenny& Newell, 2013). This definition ignores the fact that other incongruent audiovisual stimuli produce different types of percepts. For example, a reverse combination of these consonants, A[g]V[b], sounds as [bg], i.e., the visual and auditory elements one after the other. There are other pairings, which result in hearing according to the visual component, for example, the acoustic [b] presented with a visual [d] is heard as [d]. The definition of the McGurk effect should be that an auditory utterance sounds like another utterance when presented with a different visual articulation. This definition includes all variants of the illusion and has been used by MacDonald and McGurk (1978) themselves, as well as by several others (Rosenblum & Saldaña, 1996; Brancazio, Miller & Paré, 2003).

During experiments (Reisberg, McLean & Goldfield, 1987), when the task is to report what was heard, the observer reports the conscious auditory perception elicited by the audiovisual stimulus. If there is no multisensory integration or interaction, perception is identical for the audiovisual stimulus and the auditory component presented alone. If audiovisual integration is present, the conscious auditory perception changes. The extent to which visual input influences perception depends on how consistent and reliable information each modality provides. Coherent information is integrated and weighted, for example, according to the reliability of each mode, which is reflected in unperceived distinctiveness.

This perceptual process is the same for audiovisual speech (Sumby & Pollack, 1954), whether it is natural congruent audiovisual speech or artificially incongruent McGurk speech stimuli. The result is conscious auditory perception. Depending on the relative weighting of hearing and vision, the outcome for McGurk stimuli can range from hearing according to the auditory component (when hearing is more reliable than vision) to percepts of fusion and combination (when both ways are informative to some extent) to hearing according to the visual component (when sight is more reliable than hearing). Corresponding audiovisual speech is not treated differently, showing a visual effect, when auditory reliability is reduced.

The McGurk effect is an excellent tool for investigating multisensory integration in speech perception. First, the McGurk effect must be defined as a change in auditory perception due to incongruent visual speech, such that observers hear a speech sound other than that uttered by the voice, and second, the perceptual properties of the auditory and visual stimulus elements will must be considered when interpreting the McGurk effect as a reflection of integration.

3. Conclusions

In conclusion, from the above it is understood how important the understanding of speech is for man to communicate and understand those around him. The speech is special and the speech sounds have great diversity.

Listeners process speech as a series of phonemes. Also, the perception of speech is categorical and there are many factors that influence its understanding reason. The TRACE model explains the interaction of bottom up and some top down processes. Finally, the McGurk effect is a key research tool and is a demonstration of multisensory integration as it shows that auditory and visual information merge into a unified integrated perception. However, further research into speech understanding in the future could uncover important insights.

References

Alderson J. C. (2000). *Assessing Reading*, New York: Cambridge University Press.

Altmann G. T. M. (1997). *The Ascent of Babel*, Oxford: Oxford University Press.

Brancazio L., Miller J. L. and Paré M. A. (2003). "Visual influences on the internal structure of phonetic categories", *Percept. Psychophys*, Vol. 65, pp. 591-601, doi: 10.3758/BF03194585.

Galotti K. M. (2008). *Cognitive Psychology in and Out of the Laboratory* (4th ed.), Belmont, CA: Thomson Wadsworth.

Keil J., Muller N., Ihssen N. and Weisz N. (2012). „On the variability of the McGurk effect: Audiovisual integration depends on prestimulus brain states", *Cereb. Cortex 22*: 221-231, doi: 10.1093/cercor/bhr125.

Kellogg R. T. (2007). *Fundamentals of Cognitive Psychology*, London: Sage.

McClelland J. L. and Elman J. L. (1986). "The TRACE model of speech perception", *Cognitive Psychology 18*: 1-86.

McClelland J. L. (1991). "Stochastic interactive processes and the effect of context on perception", *Cognitive Psychology*.

MacDonald J. and McGurk H. (1978). "Visual influences on speech perception processes", *Percept. Psychophys. 24*: 253-257, doi: 10.3758/BF03206096.

McGurk H. and MacDonald J. (1976). "Hearing lips and seeing voices", *Nature 264*: 746-748, doi: 10.1038/264746a0.

Pinker S. (1994). *The Language Instinct*, London: Penguin. [A readable introduction to research into speech perception].

Reisberg D., Mclean J. and Goldfield A. (1987). "Easy to hear but hard to understand: A lip reading advantage with intact auditory stimuli", in: B. Dodd & R. Campbell (Eds.), *Hearing By Eye: The Psychology of Lip Reading*, London: Lawrence Erlbaum Associates Ltd., pp. 97-113.

Rosenblum L. D. and Saldaña H. M. (1996). "An audiovisual test of kinematic primitives for visual speech perception", *J. Exp. Psychol. Hum. Percept. Perform. 22*: 318-331, doi: 10.1037/0096-1523.22.2.318.

Setti A., Burke K. E., Kenny R. and Newell F. N. (2013). "Susceptibility to a multisensory speech illusion in older persons is driven by perceptual processes", *Front. Psychol. 4:* 575, doi: 10.3389/fpsyg.2013.00575.

Sumby W. H. and Pollack I. (1954). "Visual contribution to speech intelligibility in noise", *J. acoust. soc. amer. 26*.

Saussure F. (1979). *Courses in General Linguistics*, Athena. Editions Papazisi.

Tzouriadou M. (1995). *Thessaloniki*, Prometheus.

Wassenhove V., Grant K. W. and Poeppel D. (2007). "Temporal window of integration in auditory-visual speech perception", *Neuropsychologia 45*: 598-607, doi: 10.1016/j.neuropsychologia.2006.01.001.