https://doi.org/10.14400/JDC.2017.15.12.509

A review of speech perception: The first step for convergence on speech engineering

Young-lim Lee
Dept. of Psychology, Dankook University

말소리지각에 대한 종설: 음성공학과의 융복합을 위한 첫 단계

이영림
단국대학교 심리학과

Abstract People observe a lot of events in our environment and we do not have any difficulty to perceive events including speech perception. Like perception of biological motion, two main theorists have debated on speech perception. The purpose of this review article is to briefly describe speech perception and compare these two theories of speech perception. Motor theorists claim that speech perception is special to human because we both produce and perceive articulatory events that are processed by innate neuromotor commands. However, direct perception theorists claim that speech perception is not different from nonspeech perception because we only need to detect information directly like all other kinds of event. It is important to grasp the fundamental idea of how human perceive articulatory events for the convergence on speech engineering. Thus, this basic review of speech perception is expected to be able to used for AI, voice recognition technology, speech recognition system, etc.

Key Words : Convergence, Event perception, Speech perception, Direct perception theory, Motor theory

요 약 사람들은 항상 사건들과 접하고 말소리 지각과 같은 사건을 지각하는데 별 어려움이 없다. 생물학적 운동의 지각과 마찬가지로, 말소리 지각에 대한 두 이론이 논쟁해 왔다. 이 논문의 목적은 말소리 지각에 대해 설명하고 말소리 지각에 대한 운동이론과 직접지각 이론을 비교하는 것이다. 운동이론학자들은 인간은 운동신경의 명령에 의해 말소리를 지각하고 생성해 내기 때문에 인간은 말소리 지각에 있어서 특별한 감각을 가지고 있다고 주장해 왔다. 하지만, 직접지각 이론학자들은 말소리 지각은 여느 다른 소리를 지각하는 것과 다르지 않다고 제안했다. 왜냐하면, 말소리를 지각하는 것은 다른 모든 사건을 지각하는 것과 마찬가지로 필요한 정보를 직접 탐지하면 되기 때문이다. 음성공학과의 융합에 있어서 이러한 인간의 기본적인 말소리 지각 능력을 먼저 이해하는 것이 중요하다. 따라서 이러한 말소리 지각에 대한 기본적인 이해는 인공지능, 음성 인식 기술, 음성 인식 시스템 등에 사용될 수 있을 것으로 기대된다.

주제어 : 융복합, 사건지각, 말소리지각, 직접지각이론, 운동이론

* This work was supported by the research grant of Dankook University in 2017.
Received 30 October 2017, Revised 30 November 2017
Accepted 20 December 2017, Published 28 December 2017
Corresponding Author: Young-lim Lee
(Dept. of Psychology, Dankook University)
Email: younglee13@dankook.ac.kr

ISSN: 1738-1916

1. Introduction

People perceive not only numerous objects but also lots of events in our environment, such as a person walking, a person talking, a ball to catch, water falling, etc. Perceptual events basically involve the processing of temporally extended, dynamic information. One of the distinctive events humans perceive is speech perception. We do not have any difficulty to recognize the nature of an event but the questions are how people recognize events and what information enables observers to recognize events. There have been debates on speech and biological motion perception compared to other events because two main theories have different views on perception of these events. Motor theorists argue that human observers have special sensitivity when perceiving articulatory or biological motion events because we both produce and perceive those events. Direct perception theorists however, argue that perceiving events, such as speech or biological motion is not special Lee[1] reviewed event perception theory by comparing motor theory and direct perception but this review only dealt with biological motion. Regarding speech perception, these two theorists have similar ideas. It is important to understand how we perceive speech because speech perception is closely related to speech science, specifically speech engineering and learning. For instance, older adults have a difficulty with speech perception because of lower speech perception performance with noise as well as the auditory problem or cognitive variables such as deficiency of working memory capacity. People with a cochlear implant also found to have difficulties when they learn second language. Yim, Kim, and Rhee[2] suggested that patients with cochlear implant would make listening errors with different patterns when processing the second language compared to second language learners with normal hearing. Thus, understanding of how we perceive and produce speech is necessary to broaden speech engineering and to find a way of how people with auditory needs perform speech perception better.

In this paper, I briefly describe speech perception and review two main theories of speech perception.

2. Introduction to speech perception

The principle idea of both motor theory and direct perception theory is that the objects of speech perception are not acoustic or auditory, rather articulatory[3]. Individual phonetic sounds (i.e., vowels and consonants) are referred as sets of coordinated gestures of various vocal configurations such as lips, tongue body and tip, larynx, soft palate, and jaw[4,5]. Before reviewing speech perception studies, the nature of speech sound waves and the basic concepts in the acoustics of speech production need to be introduced[6]. Sound waves are one example of wave motions produced by—and consisting of—the vibration of certain quantities. Sound waves are produced by the vibration of air particles. Vibration is described as the to and fro motion of the mass spring and consists of amplitude, frequency, and period. If the mass is displaced from its rest position (e.g., stretching or compressing mass spring) and released, it moves back and forth through the rest position (i.e., vibration or oscillation). When the displacement (i.e., the distance of the mass from its rest position) is maximum, it is called the amplitude of the vibration. If there is no friction (i.e., no energy losses), the amplitude will be the same on both sides of its rest position over time. When the mass moves from one side of the rest position (i.e., displacement point) to opposite side of the rest position and back to the rest position, it is called one cycle of oscillation. We refer the number of cycles per second to the frequency of the oscillation. The period of vibration is referred as the time taken for one cycle of oscillation. Speech sound waves are generated by the vocal organs such as lips, tongue, nose, jaw and throat. An air-filled tube whose resonances play a major role in speech perception is formed by these organs. The specialized movements of the organs vary the shape of the vocal tract, which is the part of the tube that lies above the larynx. These movements produce the different sounds of speech by adjusting the vocal tract. The vocal tract configuration sets the values of its resonant frequencies and as the tract configuration is changed, the amplitudes will peak at different frequencies. We refer resonances of the vocal tract to as formants, and their frequencies to the formant frequencies. A set of characteristic formant frequencies is determined by each configuration of the vocal tract. Formant frequencies will not be uniform, such that some of them will be higher and others lower. The formant is named depending on the height in frequency; the first formant is called for the lowest formant frequency, the second formant is the next highest frequency, and so on.

Like other events, speech perception is dynamic. Remez, Rubin, Pisoni, and Carrell[7] found that a linguistic message in speech can be perceived because time-varying properties of artificial acoustic cues provide sufficient information, even though there was no acoustic elements for phonetic segments in their stimuli. Their finding is similar as Johansson's[8] point-light demonstrations in which we recognized several lights as a motion pattern when they changed over time[9].

As I mentioned earlier, theorists of two major theories, motor and direct perception theory agree that speech perception is articulatory (i.e., dynamic), rather than acoustic or auditory. They, however, had different views on question of whether speech perception is special compared to nonspeech perception and thus, I compare the notions of two major theories.

3. Comparison of two major theories

3.1 Motor theory

Motor theory suggests that speech perception is special to humans because we can perceive and produce speech sounds. The perceived speech sounds are constantly articulated and compared with the auditory sequence of the articulation[6]. Liberman and Mattingly[10] suggested that articulatory commands play an important role in speech perception. First, the speech perception is the phonetic gestures the speaker intends. The human ability to perceive speech sounds are mediated by neuromotor commands that call for articulatory movements through certain linguistically significant gestures. Phonetic segments are composed of one or more articulatory gestures such as lip rounding, tongue backing, jaw raising, etc. That is, perception of speech sounds is perception of a specific pattern in intended phonetic gestures. Second, since same mechanism is used both for speech perception and speech production, they must be internal and innate. To perceive speech sounds is linked to phonetic gestures depending on vocal-tract shapes, articulatory movements. Thus, speech perception is special compared to nonspeech perception because the link between perception and production innately specified occurs only in speech. Eimas, Siqueland, Jusczyk, and Vigorito[11] found that there was no difference on performance of 1- and 4-month-old infants compared with adults to distinguish acoustic features between the voiced stop consonants and voiceless (i.e., /b/ vs. /p/). The acoustic cue between /b/ and /p/ is voice onset time (VOT) defined as the time between the release burst and the onset of voicing; /b/ has a short VOT, whereas /p/ has a long VOT. To investigate whether infants are able to distinguish these two sounds, the first speech sound was repeatedly presented to infants. Next, acoustic sounds within the phonemic categories on the basis of VOT were presented as a second speech sound. When the first and the second sound were from the same category, infants habituation was not recovered. When the two sounds were from different phonemic categories, however, infants showed greater recovery from habituation. This finding of prelinguistic infants ability has supported that the link between perception and production is not associated with learning, rather is innately specified.

The fundamental argument of motor theory is that humans are more sensitive to perceive speech sounds relative to nonspeech sounds. Researchers have investigated whether we perceive speech sounds differently from nonspeech sounds[12,13,14].

Liberman et al.[13] investigated the discriminability of speech sounds relative to nonspeech sounds. They used the pattern playback to convert the spectrogram patterns to the sound stimuli /do/ and /to/. The pattern of sound /do/ is distinctive from the pattern of sound /to/ only in the relative time of onset of the first formant relative to other two formants (the second and third formants). Nonspeech stimuli were simply made by turning the speech spectrograms upside down. The patterns of nonspeech stimuli show the same acoustic differences (i.e., the relative time of onset of the formants) as the patterns of speech stimuli, even though nonspeech stimuli are not perceived as speech. Thus, observers can discriminate nonspeech stimuli if they use only the relative onset time. three stimuli were presented and observers were asked to decide whether the third stimulus was same as other two stimuli. The results showed that observers performed much poorly to discriminate nonspeech sounds compared to speech sounds. Thus, Liberman et al.[13] concluded that speech and nonspeech sounds are dissimilar.

Although Liberman et al.[13] found that perceiving speech sounds is different from perceiving nonspeech sounds, other researches have shown somewhat different results. Pisoni et al.[14] and Diehl and Walsh[12] investigated whether stops and glides (e.g., /b/ and /w/) are distinguishable. They used speech stimuli, /ba/ and /wa/, and compared these stimuli with corresponding nonspeech stimuli. From the auditory principle of durational contrast, the perception of length of an adjacent segment is affected by the duration of acoustic segments[15]. For instance, a longer vowel will produce shorter formant transitions and thus, the stimulus would be identified more as stop sounds. In both studies, the results showed that frequency transition duration was an effective cue to distinguish not only speech sounds, stops and glides, but nonspeech sounds, abrupt and gradual onsets. That is, distinction of speech stimuli was not different from that of nonspeech stimuli on the basis of transition duration. Diehl and Walsh[12], however, found that speech and nonspeech sounds are different with respect to amplitude rise time, even though they are similar with respect to transition duration. When transition duration was fixed and amplitude rise time was varied, effect of variation in rise time was small to discriminate speech sounds (stops vs. glides) as well as nonspeech sounds (abrupt vs. gradual onset). A stimulus length, however, had an effect on distinction between speech sounds. That is, a longer vowel shifted the boundary of stop/glide distinction toward being longer and thus, glide sounds were identified more as stop sounds.

These mixed findings in comparisons of speech with nonspeech sounds are not consistent with the fundamental notion of motor theory in which speech perception is special. Motor theorists suggest two possibilities to explain divergent findings. One possibility is that different processes constrain the ability to perceive differences among speech sounds, that is, some processes applied to speech sounds are special while others are not[16]. It is also possible that some nonspeech sounds are so speechlike as to be perceived as speech while others are not[12,17]. However, these possibilities, cannot be sufficient to support the notion that speech perception is special. Therefore, Direct perception theorists claim that speech perception is not special relative to nonspeech perception.

3.2 Direct perception theory

In contrast to motor theory's claim that speech sounds are perceived by special innate mechanisms to produce speech, direct perception theory denies the special link between perception and production[18]. Since different vocal-tract configuration could be used for the same acoustic signals and different acoustic signals could be formed by the same vocal-tract configurations, motor theorists argue that the nervous system internally computes speech signals to be perceived[10]. That is, the speech motor system is used in perception to help extract articulatory movements which produce the acoustic speech signal together. On the other hand, as Gibson[19] proposed, direct perception theorists claim that perceiving the acoustic signals is directly picking up the structure in sine waves. Thus, speech events are not distinguished from nonspeech events. Motor theorists suggest that some perceptual processes applied to speech acoustic signals, while other different processes applied to nonspeech signals because observers respond to speech signals differently to nonspeech signals[13]. On the other hand, direct perception theorists infer that the different perceptual processes do not produce different responses to speech and nonspeech signals. Instead, responses to acoustic signals are occurred by what the signals are perceived as[20]. That is to say, acoustic signals are perceived depending on information directly picked up in the environment regardless of what signals are, speech or nonspeech.

Fowler[20] investigated whether speech perception is special or not by using nonspeech signals similar to the /ba/ and /wa/ stimuli Miller and Liberman used[15]. As I mentioned earlier, Pisoni et al.[14] and Diehl and Walsh[12] investigated whether observers discriminate speech stimuli (stops vs. glide speech) and nonspeech stimuli (abrupt vs. gradual onset). Pisoni et al.[14] found that responses to speech stimuli were similar to nonspeech stimuli by using durational information in both speech and nonspeech stimuli. Diehl and Walsh [12] replicated the Pisoni et al.'s study comparing responses to speech and nonspeech stimuli because they concerned that the nonspeech stimuli used in Pisoni et al.'s study might be processed as speech due to since wave segments. Diehl and Walsh generated nonspeech stimuli using a single sine wave segment, rather than sine wave segments. They nevertheless found that speech sounds are similar to nonspeech sounds on the basis of transition duration. However, they also found that speech sounds are dissimilar to nonspeech sounds based on amplitude rise time.

Fowler[20] used nonspeech stimuli similar to the /ba/ and /wa/ stimuli instead of the synthesized nonspeech stimuli used in previous studies. She produced a steel ball rolling down a set of steel tracks for nonspeech events. Sounds were recorded as a ball rolling down from the downward slopes onto the flat or upward sloping tracks. Each event consisted of two phases stored separately. The phase 1 sound was recorded during the downward slopes and the phase 2 sound was recorded during either the flat or upward sloping track. Two phases of the event were constructed by connecting different phase 2 sounds to the five phase 1 sounds. The phase 1 sound was produced by one of five downward slopes at 50, 40, 30, 20, and 10 degrees relative to the horizontal and the phase 2 sound was produced by either 10 or 50 degree tracks. In both phases, a steeper slope is associated with shorter duration. Durations in phase 2 are also related to those in phase 1. Duration in phase 2 is longer as the slope of phase 1 is shallower (i.e., 10 degrees) in the event with the flat track. Since a long-duration phase 2 indicates a long-duration phase 1, durations in phase 2 are positively related to those in phase 1. On the other hand, duration in phase 2 is longer as the slope of phase 1 is steeper (i.e., 50 degrees) in the event with the upward sloping track. In this event, a long-duration phase 2 indicates a short-duration phase 1 and thus, durations in phase 2 are negatively related to those in phase 1. The phase 1 of each event (i.e., downward sloping part of each track) was covered with sandpaper to make the sound in phase 2 distinctively as the steel ball rolled onto the flat (or upward slope) from the downward sloping part of the track. Observers were asked to judge whether the downward sloping part of each track (the phase 1 part) is steep or shallow.

Fowler[20] predicted that judgments of phase 1 slopes would be affected by the durations of phase 2 if acoustic signals were used as information to be perceived directly, rather than through perceptual processes. The results were consistent with the prediction. In the flat condition, observers judged the slopes of phase 1 part steeper (i.e., short duration) when followed by the short (50 degree) duration phase 2 than by the long (10 degree) duration phase 2. In the upsloping condition, on the other hand, the slopes of phase 1 part were judged steeper when followed by the long (10 degree) duration phase 2 than by the short (50 degree) duration phase 2. There was difference in perception of phase 1 between two conditions in phase 2. In the flat condition in phase 2, durations of phase 1 were perceived shorter (i.e., perceived as steeper slopes) when the duration of phase 2 was short (50 degree) because the relation between phase 1 and 2 is positive. In the upsloping condition, on the other hand, since phase 1 is negatively related to phase 2, durations of phase 1 were perceived longer (i.e., perceived as shallower slopes) when the duration of phase 2 was short. Thus, even nonspeech stimuli were perceived differently depending on the durations of phase 2 used as information for the slope of a track causing the phase 1 sound structure. These results are similar to the finding of Miller and Liberman[15] in which speech perception is affected by the durational contrast.

Direct perception theorists, therefore, claim that there is no difference between speech and nonspeech perception because we pick up information in the structure of the acoustic signals and use it to perceive nonspeech sounds as well as speech sounds.

4. Conclusion

Researchers in speech perception have tried to map properties of the acoustic signal with linguistic elements such as phonemes and distinctive features. Both motor theorists and direct perception theorists have claimed that perceiving speech sounds is articulatory event rather than acoustic or auditory events. Speech perception cannot be separated from speech production. Yoon[21] found that native English speakers who received perceptual training of Korean vowels improved in both perception and production. The main difference between two theories, however, is two-folds. First, motor theorists claim that speech perception is invariant because speech perception is processed by neuromotor commands, while direct perception theorists claim that speech perception is processed by simply detecting the relevant information. Second, motor theorists claim that perceiving speech sounds special in which it is different from perceiving nonspeech sounds, whereas direct perception theorists claim that perceiving speech events is not distinguished from perceiving nonspeech events.

As smart phones have been used widely, researchers investigated to improve speech recognition performance, voice recognition technology or system [22,23,24]. Other than technologies itself, situational satisfaction with voice recognition technology of smart phones and humanistic measure about cultural changes of voice recognition technology were also investigated[25,26]. Although researches on artificial intelligence have grown numerically these days, the field of research in Korea is narrow focusing on local and technical aspects[27]. It is necessary to develop many aspects to build human-like AI. Especially, speech production and as well as voice recognition or speech perception are inevitable to interaction between AI and humans. Thus, we need to understand basic mechanisms how we perceive speech sounds and it is expected that this review could be used to extend the researches related to AI and speech engineering in future.

In sum, the arguments of two theories for speech perception is similar to those for perception of biological motion These two events are distinctive from other events because humans can both produce and perceive these events. The motor theory claims that speech perception is mediated and reconstructed by innate neuromotor commands, whereas we perceive speech directly by detecting the appropriate information from the direct perception theory.

ACKNOWLEDGMENTS

This work was supported by the research grant of Dankook University in 2017.

REFERENCES

[1] Y. Lee, “A review of event perception: The first step for convergence on robotics”, Journal of Digital Convergence, Vol. 13, No. 4, pp. 357-368, 2015.

[2] A. Yim, D. Kim, and S. Rhee, "Korean ESL learners' perception of English segments: a cochlear implant simulation study", Phonetics and Speech Sciences, Vol. 6, No. 3, pp. 91-99, 2014.

[3] R. L. Diehl, and K. R. Kluender, "On the objects of speech perception", Ecological Psychology, Vol. 1, pp. 121-144, 1989.

[4] C. A. Fowler, "An event approach to the study of speech perception from a direct-realist perspective", Journal of Phonetics, Vol. 14, pp. 3-28, 1986.

[5] J. A. S. Helpso, B. Tuller, E. Vatikiotis-Bateson, and C. A. Fowler, "Functionally specific articulatory cooperation following jaw perturbations during speech: Evidence for coordinative structures", J. of Experimental Psychology: Human Perception and Performance, Vol. 10, pp. 812-832, 1984.

[6] P. B. Denes, and E. N. Pinson, "The speech chain: The physics and biology of spoken language", New York: W. H. Freeman and Company, 1996.

[7] R. E. Remez, P. E. Rubin, D. B. Pisoni, and T. D. Carrell, "Speech perception without traditional speech cues", Science, Vol. 212, pp. 947-950, 1981.

[8] G. Johansson, "Visual perception of biological motion and a model for its analysis", Perception & Psychophysics, Vol. 14, pp. 201-211, 1973.

[9] C. A. Fowler, and B. Rakerd, "Work group on speech and sign language", In W. H. Warren & R. E. Shaw (Eds.), Persistence and Change, Hillsdale, NJ: Erlbaum, pp.283-298, 1985.

[10] A. M. Liberman, and I. G. Mattingly, "The motor theory of speech perception revised", Cognition, Vol. 21, pp. 1-36, 1985.

[11] P. Eiman, E. R. Siqueland, P. Jusczyk, and J. Vigorito, "Speech perception in infants", Science, Vol. 171, pp. 125-138, 1985.

[12] R. L. Diehl, and M. A. Walsh, "An auditory basis for the stimulus-length effect in the perception of stops and glides", J. of Acoustical Society of America, Vol. 85, pp. 2154-2164, 1989.

[13] A. M. Liberman, K. S. Harris, J. Kinney, and H. Lane, "The discrimination of relative onset-time of the components of certain speech and nonspeech patterns", J. of Experimental Psychology, Vol. 61, pp. 379-388.

[14] D. B. Pisoni, T. D. Carrell, and S. J. Gans, "Perception of the duration of rapid spectrum changes in speech and nonspeech signals", Perception & Psychophysics, Vol. 34, pp. 314-322, 1983.

[15] J. L. Miller, and A. M. Liberman, "Some effects of later-occurring information on the perception of stop consonant and semivowel", Perception & Psychophysics, Vol. 25, pp. 457-465, 1979.

[16] P. Eimas, "The equivalence of cues in the perception of speech by infants", Infant Behavior and Development, Vol. 8, pp. 125-138, 1985.

[17] C. T. Best, M. Studdert-Kennedy, S. Manuel, and J. Rubin-Spitz, "Discovering phonetic coherence in acoustic patterns", Perception & Psychophysics, Vol. 45, pp. 237-250, 1989.

[18] C. A. Fowler, and B. Galantucci, ″The relation of speech perceptio nand speech production″, In D. B. Pisoni & R. E. Remez (Eds.), The Handbook of Speech Perception, Oxford, UK: Blackwell, pp. 633-652, 2005.

[19] J. J. Gibson, ″A theory of direct visual perception″ In J. Royce & W. Rozeboom (Eds.), The Psychology of Knowing, New York and London: Gordon and Breach, pp. 215-227, 1972.

[20] C. A. Fowler, ″Sound-producing sources as objects of perception: Rate normalization and nonspeech perception″, J. of Acoustical Society of America, Vol. 88, pp. 1236-1249, 1990.

[21] E. Yoon, ″The effects of perceptual training on speech production: Focusing on Korean vowels″, Studies in Foreign Language Education, Vol. 22, No. 2, pp. 1-27, 2013.

[22] J. Hwang, ″Voice recognition performance improvement using the convergence of Bayesian method and selective speech feature extraction″, J. of the Korea Convergence Society, Vol. 7, No. 6, pp. 7-11, 2016.

[23] J. Lee, J. Lee, and J. Lee, ″Speech recognition of Korean phonemes ′ㅅ‘,’ㅈ‘,’ㅊ‘ based on sign distribution volatility″, J. of KIISE: Computing Practices and Letters, Vol. 19, No. 7, pp. 377-382, 2013.

[24] S. Nam, E. Jean, and I. Park, ″A real-time embedded speech recognition system″, The Institute of Electronics Engineers of Korea-Computer and Information, Vol. 40, No. 1, pp. 74-81, 2003.

[25] Y. Lee, and S. Kim, ″Study on the situational satisfaction survey of smart phone based on voice recognition technology″, J. of Digital Convergence, Vol. 15, No. 8, pp. 351-357, 2017.

[26] H. Yuk, and B. Cho, ″A study on the humanistic measure about cultural changes of voice recognition technology″, J. of Digital Convergence, Vol. 13, No. 8, pp. 21-31, 2015.

[27] M. Chung, S. Park, B. Chae, and J. Lee, ″Analyses of major research trends in artificial intelligence through analysis of thesis data″, J. of Digital Convergence, Vol. 15, No. 5, pp. 225-233, 2017.

이 영 림(Lee, Youn Lim)

- 2000년 2월 : 성신여자대학교, 영어영문학과(학사)
- 2003년 8월 : Western Kentucky University, Experimental Psychology (석사)
- 2009년 9월 : Indiana University Bloomington, Psychological & Brain Sciences / Cognitive Sciences (박사)
- 2016년 3월 ~ 현재 : 단국대학교 심리학과 강의전담 조교수
- 관심분야 : 3D 시각지각, 직접지각 접근법
- E-Mail : younglee13@dankook.ac.kr