International Journal of Linguistics, Literature and Translation
ISSN: 2617-0299 (Online); ISSN: 2708-0099 (Print)
DOI: 10.32996/ijllt
Journal Homepage: www.al-kindipublisher.com/index.php/ijllt

IJLLT
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

Active Models in Speech Perception: A Critical Review in Phonetics

Wael Almurashi
Department of Languages and Translation, Faculty of Arts and Humanities, Taibah University, Medina, Saudi Arabia
Corresponding Author: Wael Almurashi, **E-mail**: wmurashi@taibahu.edu.sa

| ABSTRACT
This critical review examines four prominent active models of speech perception: quantal theory, motor theory, hyper-hypo-articulation theory, and exemplar theory. Each of these models offers distinctive insights into how listeners perceive, interpret, and categorize speech sounds. Quantal theory explores the non-linear relationship between articulation and acoustic output, emphasizing stable and unstable regions in speech production. Motor theory posits that speech perception is grounded in articulatory gestures, suggesting a direct link between production and perception. The hyper-hypo-articulation theory focuses on the speaker's adaptive articulation, which ranges from hyper- to hypo-articulation depending on the listener's needs and the communicative context. Exemplar theory, rooted in cognitive psychology, highlights the role of memory and experience, with individual speech sounds compared to stored exemplars in memory to aid perception. Despite their contributions, none of these models offers a comprehensive solution to all speech perception challenges. Each theory is suited to specific contexts, speech features, or tasks, making it crucial for researchers to understand their limitations. Furthermore, the review underscores the importance of selecting the appropriate model for particular research or practical applications, given the variability in their functionality. Finally, it calls for future research to delve deeper into passive models of speech perception and suggests that comparative studies between active and passive models could yield valuable insights.

| KEYWORDS

Speech perception; Active models; Phonetics, Sounds.

| ARTICLE INFORMATION

ACCEPTED: 01 October 2024 **PUBLISHED:** 16 October 2024 **DOI:** 10.32996/ijllt.2024.7.10.19

1. Introduction

Speech perception, which encompasses the processes of hearing, interpreting, and comprehending every sound a speaker produces, is integral to understanding spoken language. It involves organizing auditory features in a way that reflects the speech patterns of a specific language. Speech perception includes both phonological and phonetic elements, as well as the syntactic and semantic aspects of a spoken message. For effective speech perception, a model is required that integrates these components to create a coherent message. Over the past four decades, categorical perception has gained significant attention, prompting the development of various theories that attempt to explain how people perceive sounds. Several models have emerged to help understand the processes underlying speech perception, with some focusing solely on perception or production, while others combine both. The earliest theories date back to the mid-1900s, and models have continued to evolve since then. Much remains unknown about how hearers evolved into listeners capable of decoding meaning from speech signals. The existing literature proposes two listening models: active and passive (Crystal, 1997). However, both models present certain challenges. Passive models emphasize the sensory aspect of speech perception, suggesting that listeners recognize sounds through internal filtering. In contrast, active listening models, which are the focus of this paper, argue that listeners rely on their pre-existing language knowledge to decode novel speech strings. In other words, they use their understanding of sound production to recognize words (O'Grady, 2012). Numerous variations of active models exist, but the aim of this paper is to critically review key active speech

perception theories—specifically motor theory, quantal theory, hyper-hypo articulation theory, and exemplar theory—and to provide a clear overview of the issues associated with each.

2. Active Models in Speech Perception

2.1 Motor Theory

2.1.1 Core Principles

Liberman and colleagues proposed the formation of a specialized speech mechanism, arguing that categorical perception is exclusive to speech sounds and does not apply to non-speech sounds (Liberman, 1996). The core idea behind this view, which is central to the motor model of speech perception, is that speech production strongly influences perception (Lindblom, 1996; Gerrits, 2001). Liberman (1996) suggested that listeners recognize sounds by simulating the speaker' articulatory movements. The Motor theory also sheds light on the intriguing 'McGurk effect', which highlights the multimodal nature of speech perception. For example, in a video where a speaker repeatedly produces the syllable [gagaga], but the original soundtrack is replaced with a recording of [dadada], viewers perceive a mismatch. While watching the video, listeners may perceive the sounds as [gagaga], but when they close their eyes and listen carefully, they will actually hear [dadada] (O'Grady, 2012).

2.1.2 Limitations

Despite its contributions, the motor theory presents several issues. First, it does not demonstrate the ability of individuals with speech disorders to comprehend speech effectively. Additionally, the motor theory struggles to account for certain subtle distinctions in speech sounds that are difficult to demonstrate, such as the voicing contrast between /do-to/, which highlights a significant production gap (Liberman et al., 1961). The distinction between /d/ and /t/ involves fine differences in vocal cord vibration that may not be fully captured by observable gestures, making it difficult for motor theory to explain how listeners reliably perceive this contrast. Similarly, the motor theory cannot adequately explain the weak perception of the affricates (e.g., /tʃ/) and fricatives (e.g., /ʃ/), where a clear articulatory gap exists (Howell & Rosen, 1983). Another issue is the motor theory's reliance on phonemic gestures as invariant units based on articulatory gestures; however, gestures are not invariant and vary across speakers. This raises the question: How can a speaker's gestures be properly defined when each individual has a different vocal tract and sound production method? Moreover, speech perception is not unique to humans, as animals are also capable of distinguishing speech sounds. For instance, chinchillas can discriminate human speech sounds, suggesting that the production system is not essential for sound recognition (Kuhl & Miller, 1975). This finding supports the argument that the production system is not necessary for basic speech sound recognition and raises the question: if non-human animals can recognize these sounds without a production system, why do humans rely on it for speech perception? Evidence from developmental studies provides further insight. Infants can distinguish sounds they cannot yet produce. For example, a study by Eimas et al. (1971) showed that even at a very young age, infants are capable of discriminating between phonetic contrasts like /ba/ and /pa/, well before they develop the ability to articulate these sounds. This suggests that speech perception is independent of the motor ability to produce sounds and, therefore, the production system may not be necessary for recognition. Thus, there is no clear developmental explanation for why the production system would be necessary for sound recognition, as even non-verbal animals and pre-verbal infants demonstrate the ability to perceive speech sounds without using the production system (Stasenko et al., 2013). Lastly, the motor theory does not clarify how phonetic knowledge is linked to linguistic knowledge, or how allophonic variation and connected speech are processed.

2.2 Quantal Theory

2.2.1 Core Principles

The second influential active listening approach is quantal theory (Stevens, 1989), which is based on the existence of nonlinearities in the mapping between articulatory configurations and acoustic outputs. Perturbations in articulatory parameters in certain regions of the articulatory 'space' result in minimal changes to the acoustic output, while similar perturbations in other regions cause significant acoustic changes. Quantal theory posits that the selection of preferred sound categories occurs in stable regions, which are separated by unstable regions. In summary, the key assumptions of this model are as follows: Speech is quantal in nature. There are abrupt shifts in the relationship between continuous articulatory variation and discontinuous acoustic parameter values, as well as between continuous acoustic variation and discontinuous auditory response values. Speech involves alternating between plateau-like regions (stable articulatory or acoustic regions) and transition-like regions (where rapid changes in acoustic or auditory limits occur). Invariance is distributed across different constraints, and quantal jumps can be observed among articulatory, acoustic, and auditory parameters. These assumptions lead to the production of a set of features that listeners use to identify the unique characteristics of a language.

2.2.2 Limitations

This theory faces several challenges. The first issue is the difficulty in defining the amount of precision and variation required to account for no perceived change. For example, at 800 Hz, a 50 Hz difference is noticeable, but at 2000 Hz, the same difference is not. Furthermore, while a 50 Hz change is easy to perceive at lower frequencies, larger jumps are necessary to distinguish between categories at higher frequencies, such as moving from 800 Hz to 2000 Hz in the second formant. In this case, a jump of 100 Hz or

150 Hz may be required to perceive differences between categories. Additionally, the relationship between articulatory parameters and their acoustic consequences is complex because multiple articulators can produce similar acoustic results. For instance, labiality, retroflexion, and pharyngealization, which occur in the same quantal regions, all result in a lower F2. This lowering of F2 can draw attention to the second formant, but it can be difficult to determine the specific articulatory cause. If F1 is rising, the phenomenon is identified as pharyngealization. If F1 is falling, it is considered labialization, and if F3 is falling, it indicates retroflexion. Consequently, determining the cause based on a single acoustic output is challenging. Therefore, a combination of formants is often used to define more categories.

In some studies on Arabic pharyngealization, the focus has been primarily on the lowered F2 as the main cue. However, the question arises: was the cue truly pharyngealization, or could it have been labialization or retroflexion? This highlights the need for multiple formants to avoid simple errors that could lead to significant inaccuracies and disrupt results. Another issue is that quantal theory considers 'rare' sounds, such as retroflex and pharyngeal consonants, but overlooks more common sounds like alveolar consonants. Alveolars are particularly complex due to the many variations caused by different parts of the tongue, which can dramatically alter the acoustic output. Lastly, the theory suggests that vowels such as /i/, /a/, and /u/ occur in stable regions, reinforcing the idea that they occupy similar positions across languages. However, this is misleading, as these vowels are sufficiently distinct. For instance, in a vowel system with 12 vowels like those in English, the vowels occupy specific positions that do not shift, and speakers produce them consistently in those positions. In Arabic, which has fewer vowels, including /e/ and /a/, the vowel /e/ can be produced as /a/ in place of diphthongs. For example, the words /zait/ and /zeit/ both mean 'oil' in Arabic, even if the vowel quality changes. While such vowel substitutions do not affect meaning in Arabic, in languages with a larger vowel inventory, changing vowels would alter the meaning of the word.

2.3 Hyper-Hypo-Articulation Model

2.3.1 Core Principles

The third model is the hyper-hypo-articulation (H&H) model. Lindblom (1990) observed that in this model, word pronunciation occurs along a spectrum, ranging from hyper to hypo articulation. In hyper articulation, words are pronounced more clearly than usual, linked to acoustic-phonetic features that reflect the speaker's increased effort, such as longer durations and expanded vowel spaces. On the other hand, hypo articulation involves less clear pronunciation, characterized by reduced vowel spaces, shorter durations, and sometimes dropped phonemes. Recent research has focused on when and how speakers use hypo- and hyper-articulation. Lindblom's H&H theory posits that speakers only need to articulate clearly enough for listeners to distinguish the intended words, as long as the listeners have access to context or other independent information to aid comprehension. In summary, the foundation of this model is primarily on articulation rather than perception. Moreover, the model offers a general framework for speech perception, aiming to explain the variability in the signal while downplaying the role of the sensory model.

2.3.2 Limitations

Sensory models emphasize that listeners extract important phonetic details from the acoustic signal. In contrast, direct realism suggests that sounds are perceived as articulatory gestures rather than being perceived solely as acoustic signals, which are considered less meaningful in this framework. In direct realism, phonological units (like phonemes) are considered less central because the goal of speech perception is to perceive gestures and meaning, not to focus on individual phonemes. Sensory models, however, view phonetic details and phonemes as important for identifying speech sounds.

This critique aligns with aspects of phonological theory that suggest an overemphasis on phonetic details can detract from the listener's primary goal of understanding the message. The H&H model argues that focusing on meaning, rather than phonetic minutiae, is more important for speech perception. The theory emphasizes that understanding the meaning of the message is more important than analyzing fine-grained acoustic details, such as how combinations of F1, F2, and F3 form specific vowels or phonemes. In the H&H model, acoustic and articulatory tools, along with phonological theory and phonemes, function as tools with the primary purpose of recognizing meaning. Only a few studies have focused on actual meaning. However, this theory is truly active because it involves real communication between speakers and listeners, who adapt to each other. For example, if a speaker is habitually hypo-articulating and notices that listeners are struggling to understand due to background noise, the speaker will tend to hyper-articulate to ensure comprehension. This demonstrates that humans are capable of adapting their communicative functions. The H&H model provides evidence that humans are active in both speech production and perception.

2.4 Exemplar Model

2.4.1 Core Principles

The final theory discussed in this paper is the exemplar model, which originated in psychology as a model of perception and categorization before being applied to speech (Johnson, 1997; Pierrehumbert, 2001). Exemplar theory is based on the connection between past experiences with words and memory. The model aims to explain how listeners recall acoustic episodes, which are experiences involving spoken words. Research has shown that listeners remember details of specific acoustic events if those events

are familiar (Goldinger, 1996). Listeners are more likely to recognize words they have heard before if the same speaker repeats them at the same speaking rate, indicating the familiarity of the acoustic episode. According to exemplar theory, each word spoken leaves a distinct imprint in the listener's memory, which later helps in recalling and recognizing words. When listeners encounter a new word, its imprint is compared to previous words to detect similarities (Goldinger, 1998). Exemplar theory emphasizes that as listeners hear and learn new words, their memory becomes more stable, leading to lexical improvement. In summary, in the exemplar model, a large cloud of recalled symbols from a category, organized within a cognitive map, represents each category. Memories of very similar situations are grouped closely together, while dissimilar memories are spaced farther apart. Speech production involves reproducing the acoustic signals of previously heard tokens, reflecting the relevant contextual information. An example of this process can be seen in Egyptian speakers, who typically produce /z/ instead of /ð/ in everyday speech but can acquire the /ð/ sound in standard Arabic, often substituting /z/ due to frequent exposure in daily communication. Rather than extracting features in speech perception, each sound is compared to stored exemplars for that category, and it is classified with the exemplar collection that most closely matches it. Listeners develop sensitivity to surface forms and retain fine phonetic details, enabling them to categorize sounds more effectively. New experiences can shift category boundaries (plasticity) by creating memory traces, and with increased experience, category stability improves. The exemplar-based model has proven useful in bridging the gap between sociolinguistics, phonetics, and speech perception (Gahl & Yu, 2006). For instance, Hay, Jannedy, and Mendoza-Denton (1999) found that the phonetic implementation of certain features can influence both lexical frequency and the ethnicity of the addressee. Similarly, Clopper and Pisoni (2004) showed that dialect perception aligns well with an exemplar-based account. Hay, Nolan, and Drager (2006) further examined the implications of the exemplar-based model for speech perception, demonstrating that listeners' vowel-identification judgments depend on their assumptions about the speaker's dialect origin. Explaining these findings is challenging without assuming that individual words have associated phonetic distributions linked to social and indexical information.

2.4.2 Limitations

While the exemplar model has proven useful in linking sociolinguistics, phonetics, and speech perception, it faces some challenges. One significant issue is the difficulty in defining word similarities, which can vary widely between individuals and are influenced by contextual, phonetic, and social factors (Johnson et al., 1999). This variability complicates the consistency of categorization, as different listeners might store different exemplars for the same sounds based on their unique experiences and the situations in which they hear them. Moreover, the model assumes that every experience leaves an imprint, but it does not fully account for how irrelevant or erroneous exemplars are filtered out or updated in memory. Exemplar theory lacks a formalized process for prioritizing or ignoring certain exemplars, which can lead to inefficiencies in how memories are stored and retrieved. Another challenge is the variability in acoustic episodes. The same speech event may not result in identical perceptual experiences for different listeners, and even the same listener may perceive it differently on different occasions. This variability complicates how well the model can explain stable and consistent speech perception across different contexts and speakers. The theory doesn't fully address how listeners manage to extract reliable patterns despite hearing speech in varying conditions. Additionally, exemplar theory struggles to explain how listeners generalize from specific stored examples to abstract phonetic categories, which may be necessary for learning new languages or dialects with unfamiliar sound patterns. While the model accounts for the storage of detailed phonetic information, it lacks a clear mechanism for the formation of higher-level phonological rules or grammatical abstractions based on these exemplars. Lastly, although exemplar theory effectively incorporates social and indexical information, such as dialect and speaker identity, it struggles to explain how this information is integrated with broader linguistic knowledge, such as grammatical structures or phonological rules. This makes it difficult to apply exemplar theory to more abstract levels of language processing, beyond the surface phonetic details.

3. Conclusion

In conclusion, this critical review examined several active theories of speech perception, including quantal theory, motor theory, hyper-hypo-articulation theory, and exemplar theory. While these are not the only active models available, they are commonly referenced, and their perceptual properties are relatively well understood. As discussed, various models can be applied to perceive speech, each with its own approach and function. The choice of a model depends on the specific speech feature or the purpose for selecting it. Each theory is unique and operates according to its design, but none of them is without limitations, as no perfect model for speech perception exists. If one model fails to solve a perception issue, there is often a strong possibility that another model can provide the solution. Each model discussed in this paper has a specific function and performs best when used for that purpose. The development of speech perception theories is aimed at assisting with the detection and interpretation of speech, especially in challenging situations where distinguishing utterances becomes difficult. It is important to recognize the variations among these models, as selecting the right one can make solving a speech perception problem much easier compared to using an unsuitable theory. This paper focused on only a few active models of speech perception. Future critical reviews could delve deeper into one of these models or explore other active models that were not covered. Additionally, since this paper focused solely on active models, it would be valuable for future studies to examine passive models and compare them with active ones.

Funding: This research received no external funding.
Conflicts of Interest: The authors declare no conflict of interest.
Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers.

References

[1] Clopper, C. G., & Pisoni, D. B. (2004). Some acoustic cues for the perceptual categorization of American English regional dialects. *Journal of Phonetics*, *32*(1), 111–140.
[2] Crystal, D. (1997). *The Cambridge Encyclopaedia of Language*. 2nd edition. Cambridge: Cambridge University Press.
[3] Eimas, P. D., Siqueland, E. R., Jusczyk, P., & Vigorito, J. (1971). Speech perception in infants. *Science, 171*(3968), 303–306.
[4] Gahl, S., & Yu, A. C. (2006). Introduction to the special issue on exemplar-based models in linguistics. *The Linguistic Review*, *23*(3), 213–216.
[5] Gerrits, P. A. M. (2001). *The categorisation of speech sounds by adults and children*. Utrecht, The Netherlands: LOT.
[6] Goldinger, S. D. (1996). Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* 22, 1166–1183.
[7] Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, *105*(2), 251–279.
[8] Hay, J., Jannedy, S., & Mendoza-Denton, N. (1999). Oprah and /ay/: Lexical frequency, referee design and style. In *Proceedings of the 14th International Congress of Phonetic Sciences* (pp. 1389-1392). Berkeley, CA: University of California.
[9] Hay, J., Nolan, A., & Drager, K. (2006). From fush to feesh: Exemplar priming in speech perception. *The Linguistic Review*, *23*(3), 351–379.
[10] Howell, P., & Rosen, S. (1983). Natural auditory sensitivities as universal determiners of phonemic contrasts. *Linguistics, 21*(1), 205–235.
[11] Kuhl, P. K., & Miller, J. D. (1975). Speech perception by the chinchilla: Voiced-voiceless distinction in alveolar plosive consonants. *Science*, *190*(4209), 69–72.
[12] Johnson, K. (1997). Speech perception without speaker normalization: An exemplar model. In K. Johnson & J. Mullennix (Eds.), *Talker variability in speech processing* (pp. 145–165). San Diego: Academic Press.
[13] Johnson, K., Strand, E. A., & D'Imperio, M. (1999). Auditory–visual integration of talker gender in vowel perception. *Journal of Phonetics*, *27*(4), 359–384.
[14] Liberman, A. M. (1996). *Speech: A special code*. MIT Press.
[15] Liberman, A.M., Harris, K.S., Kinney, J.A., & Lane, H. (1961). The discrimination of relative onset time of the components of certain speech and nonspeech patterns. *Journal of Experimental Psychology, 61*(5), 379–388.
[16] Lindblom, B. (1990). Explaining phonetic variation: A sketch of the H&H theory. In *Speech production and speech modelling* (pp. 403–439). Springer Netherlands.
[17] Lindblom, B. (1996). Role of articulation in speech perception: Clues from production. *The Journal of the Acoustical Society of America, 99*(3), 1683–1692.
[18] O'Grady, G. (2012). *Key concepts in phonetics and phonology*. Palgrave Macmillan.
[19] Pierrehumbert, J. (2001). Exemplar dynamics: Word frequency, lenition, and contrast. In J. Bybee & P. Hopper (Eds.), *Frequency effects and the emergence of lexical structure* (pp. 137–157). Amsterdam: Benjamins.
[20] Stevens, K. N. (1989). On the quantal nature of speech. *Journal of Phonetics*, *17*(1), 3–45.
[21] Stasenko, A., Garcea, F. E., & Mahon, B. Z. (2013). What happens to the motor theory of perception when the motor system is damaged?. *Language and Cognition*, *5*(2-3), 225–238.