universität wien

MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

"Is Perceiving Speech Perceiving Gestures?
The Role of the Motor System in Phoneme Perception"

verfasst von / submitted by
Viktoria Groiß

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of
Master of Arts (MA)

Wien, 2023 / Vienna, 2023

Studienkennzahl lt. Studienblatt/ degree programme code as it appears on the student record sheet:	UA 066 867
Studienrichtung lt. Studienblatt/ degree programme as it appears on the student record sheet:	MA Allgemeine Linguistik: Grammatiktheorie und kognitive Sprachwissenschaft
Betreut von / Supervisor:	Univ.-Prof. Dr. Jutta L. Mueller

Abstract

Language is a cognitive function with input and output being both auditory and motoric in nature. The motor system is a crucial component of speech production, but its role in speech perception has been an ongoing debate. Although a primary role of the articulatory motor system can be excluded, some experiments have shown brain activity of the articulatory motor cortex in response to passive listening and an influence of manipulation of the articulators or the corresponding brain regions on phoneme perception. The study presented here aims to expand the current body of knowledge by using a simple mechanical stimulation (wooden tongue depressor) of the articulators in a behavioral experiment where participants are asked to categorize speech sounds from a continuum (/i:/-/y:/, /ba/-/da/, /ga/-/ka/). Logistic curves were fit to their identification data and midpoint and boundary range were then compared in a repeated measures ANOVA. No significant effects were found on the midpoints or ranges. However, a subsequent analysis of the response times revealed a significant main effect of the mechanical stimulation, compared to the baseline and control conditions. A non-linguistic task would be needed to decide whether this is an effect of divided attention or reflects a motor influence on phoneme perception.

Sprache ist eine kognitive Funktion deren Input und Output sowohl auditiver als auch motorischer Natur ist. Das menschliche Bewegungssystem ist ein essenzieller Bestandteil in der Sprachproduktion, doch seine Rolle in der Sprachwahrnehmung ist eine fortlaufende Debatte. Auch wenn eine primäre Rolle des artikulatorischen Bewegungssystems ausgeschlossen werden kann, zeigen eine Reihe von Studien eine gesteigerte Aktivität im Motor Cortex während dem passiven Hören von Sprache und Abweichungen in der Wahrnehmung von Phonemen bei Manipulation der Artikulatoren oder der entsprechenden Hirnregionen. Die hier vorgestellte Studie soll den bestehenden Wissenstand ergänzen, indem eine simple mechanische Stimulation (Holzspatel) der Artikulatoren in einem Verhaltensexperiment eingesetzt wird, bei dem die ProbandInnen Sprachlaute aus einem Kontinuum (/i:/-/y:/, /ba/-/da/, /ga/-/ka/) kategorisieren müssen. An ihre Identifikationsdaten wurden logistische Kurven angeglichen, deren Mittelpunkte und Grenzweite

anschließend in einer ANOVA mit Messwiederholung verglichen wurden. Weder für die Mittelpunkte noch die Grenzweiten konnte ein signifikanter Effekt gefunden werden. Eine anschließende Analyse der Reaktionszeiten zeigte jedoch einen signifikanten Haupteffekt der mechanischen Stimulation im Vergleich zur Baseline und der Kontrollbedingung. Um zu entscheiden, ob es sich um eine Effekt durch geteilte Aufmerksamkeit oder einen motorischen Einfluss auf die Phonemwahrnehmung handelt, bräuchte es eine nicht-sprachliche Aufgabe.

Contents

Acknowledgments

I want to thank everyone who has supported me in the process of writing this thesis. First and foremost, I want to thank Univ.-Prof. Dr. Jutta Mueller for being the best supervisor possible. Without your guidance and support, this project could not have been realized. You have always found time in your busy schedule to discuss the project and the problems along the way. Thank you, also, for giving me the freedom and encouragement to explore my own ideas.

I want to thank my friends, who have supported me for years in my personal and academic goals. I thank Pia Kappler and Marjana Gruber for being wonderful study buddies throughout the years and Andrea Topalović, Theresa Michels, and Matúš Kresky for playing guinea pig in the development of the experiment.

Without my family's support, I would not have gotten this far. I thank you for encouraging me to follow my passion and supporting my decisions.

The biggest thanks go out to my partner Manuel Schmitzer, who gave me the necessary push when I felt like progress was stalling. Thank you for writing the funniest comments during proofreading. You make me laugh even in the most stressful moments.

Lastly, I want to thank all my participants for their time. Without you, this would not have been possible.

1 Introduction

Speech exists in two states, acoustic patterns of sound pressure waves and the articulatory movements that produce them. While the perception of speech is typically considered an auditory skill, the two states are undoubtedly linked. This link certainly also exists in the human brain.

Studies finding activation of motor areas during passive listening have renewed a debate around the motor influence on speech perception. The most extreme theory is the motor theory of speech perception (Liberman et al., 1967, Liberman and Mattingly, 1985, which proposed that the objects of speech perception are not the acoustic patterns but the motor commands used to generate them. While nobody today would take this extreme view, newer theoretical accounts propose tightly linked neural circuits between action and perception areas leading to interactive processing of speech (Pulvermüller, 2018). While many still argue that the motor system does not contribute to speech perception at all (Lotto et al., 2009), more and more studies accumulate that demonstrate the effects of manipulation of the motor system on phoneme perception. At this point, even some of the very skeptics criticizing proposals of motor involvement tried to account for these findings and integrate them into their models (Hickok et al., 2011).

In section 2, this thesis will start with a brief overview of the original motor theory of speech perception to introduce some important concepts relevant to speech perception and discuss the shortcomings of the theory. Section 3 will give an introduction to the human motor system to better understand the following summary of studies that found activity in the motor cortex during passive listening in section 4. Section 5 explores how perception and production areas may be linked in the brain by looking at mirror neurons and associative learning. Section 6 will then go over the different roles the motor system might play in speech perception and discuss studies where the motor cortex or the articulators were manipulated. After an interim summary in section 7, the experiment conducted as part of this thesis is laid out, with a description of the methods in section

9, a summary of the results in section 10, and a discussion of these results in section 11.

Part I

Theoretical Background

2 The Rise and Fall of the Motor Theory of Speech Perception

Discussing motor involvement in speech perception processes is impossible without at least mentioning the motor theory of speech perception (Liberman et al., 1967, Liberman and Mattingly, 1985). It was the first theory that attributed a major role to the motor system, specifically the part of the brain responsible for articulation, in phoneme perception and, despite a clear consensus that the theory in its original form does not hold up, finds its way into discussion time and time again. So, while it might seem pointless to lay out an outdated theory, it still offers a good starting point for introducing some important concepts and discussing some of the major problems that theories of speech perception need to address. It further offers a good opportunity to explain why the original Motor Theory failed and through that define the limits within which modern theories can include motor system recruitment.

2.1 Research Findings of the Haskins Laboratories

The motor theory of speech perception was developed at a time when interest in the cognitive processes underlying speech perception was still new. In the 1940s, speech perception research was largely focused on improving the intelligibility of speech transmission systems developed during World War II. After the war, research at the Haskins Laboratories was driven by the goal of building a reading machine for blind veterans and therefore attempted to create an acoustic alphabet (for more details see Cooper et al., 1984). Only when this seemingly simple approach failed, because test subjects could not understand the acoustic code at higher speed, did the acoustic details of speech sounds and the cognitive processes underlying speech perception become of interest. It raised the question of how speech differs from these artificial sound codes and why it can transmit information at rates that other codes (e.g. morse code) cannot.

The development of the sound spectrograph and a pattern playback machine allowed the creation and manipulation of speech sounds and opened the door for finding the acoustic patterns that are characteristic for identifying and discriminating phonemes[1]. Usually, participants are given multiple steps on an acoustic continuum with a single acoustic variable being varied continuously. Participants are then asked to categorize these sounds into predefined categories. Through this technique, Liberman et al. (1952) were able to identify a small number of cues for each phoneme of the English language.

Following the account of Cooper et al. (1984), the new advances in speech synthesis and knowledge about acoustic phonetics allowed a new approach to the development of reading machines. Instead of inventing a new acoustic code, each letter of an input text could be translated into speech-like output through speech synthesis. But matching letters (or more precisely, phonemes) with invariant acoustic patterns turned out to be more difficult than expected.

The speech signal is not segmental

One observation was the fact that the speech stream could not be cut into phoneme segments as easily as expected. The phonemes seemed to overlap, which yielded an unintelligible sequence when the segments were reordered (Cooper et al., 1984, 59f). These transitions result from the articulators moving from one position to another to produce the relevant speech sounds, a phenomenon called *coarticulation*. It seemed, however, that these transitions were not just noise but relevant for speech (Liberman et al., 1952, Liberman, 1957).

Transitions carry important cues for both phonemes

The rapid changes in formant frequency[2] between two phonemes were found to be an important cue for identification (Liberman et al., 1958). When listeners were presented with the vowel /a/ starting with different transition patterns, they could identify them as /ba/, /da/, or /ga/. This means that the transitions carry information about the preceding and following phoneme and might therefore be a factor for why speech is much

1 Phonemes are defined as the smallest units that distinguish between two words like *fish* and *dish*. Such contrasts are also called *minimal pairs*. Phonemes are in some ways analogous to letters of the alphabet. But the letters and phonemes of a word may differ greatly, as English is the prime example for this (e.g. *thought* - /θɑ:t/).

2 Formants are certain parts of the frequency spectrum where the energy is at maximum. This is due to the resonances formed in the vocal tract. As the vocal tract changes shape during articulation, the formants differ between phonemes (Johnson et al., 2011).

more efficient than other acoustic codes. These transition patterns, however, would vary not just for each consonant but also for each vowel. Figure 2.1 shows some examples of these transition patterns. For the consonants /d/ and /g/ in combination with five vowels, the first and second formant frequencies are shown. The straight part of the line represents the frequencies typical for the vowel. For every consonant, the line starts at a different frequency, and for every vowel ends at a different frequency. Therefore every combination of consonant and vowel shows a unique pattern which might show a falling formant frequency (as for /du/ and /gu/ as well as /gi/) or a rising formant frequency (as for /di/) (Liberman, 1957, Liberman et al., 1967).

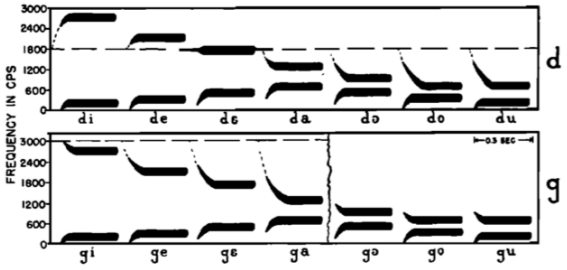


Figure 2.1: Spectrographic patterns for /d/ and /g/ with various following vowels (Liberman, 1957, p. 121)

Acoustically different stimuli can be perceived as the same phoneme

Figure 2.2 shows how often listeners categorized a specific starting frequency for a consonant-vowel-transition as a certain consonant. Depending on the following vowel, a specific starting point would be identified as very different consonants. Also, the same consonant needs very different starting frequencies to be identified correctly depending on the following vowel (Liberman et al., 1952, Liberman et al., 1967).

Acoustically identical stimuli can be perceived as two different sounds based on context

These formant transitions can not only be perceived as belonging to different phonemes based on the context but also in isolation. Without context, these same transitions do not even remotely sound like speech but instead are perceived as chirps. Even more surprisingly, when the transition is played in one ear and a matching context is played in the other ear at the right moment, listeners perceive both the phoneme and the chirp simultaneously (Liberman et al., 1967). This raised the question of how one perceptual process could result in two very different percepts at the same time. They concluded,

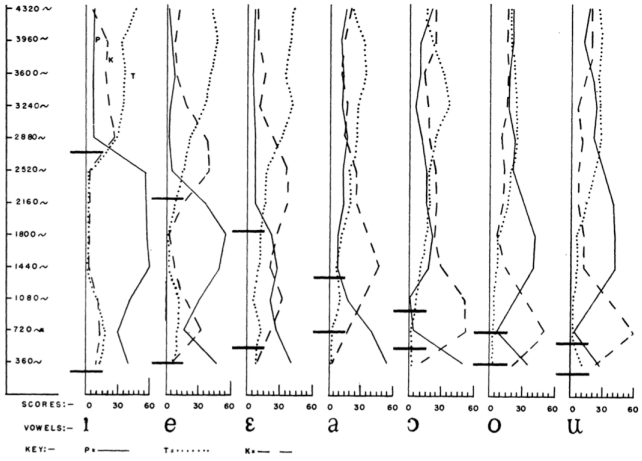


Figure 2.2: The number of identifications as /p/, /t/, or /k/ for each vowel depending on the frequency of the consonant burst (Liberman et al., 1952, p. 504).

that speech perception and general auditory perception were two separate "modes" that are usually in competition with each other (Liberman et al., 1967, Mann and Liberman, 1983.

When the acoustic signal is changed continuously, perception follows articulation rather than acoustic information (*Categorical Perception*)

While the acoustic speech signal operates in a continuous space, our phoneme categories do not. The boundaries for these categories could in principle be anywhere in the acoustic continuum. Yet, the category stereotypes, i.e. the parts of the continuum that are clearly identified as a particular phoneme and rated as a good representation, are located where the acoustic pattern matches the outcome of articulation closely (Liberman et al., 1967).

Cues interact with each other

Cues like formant transitions interact with other cues (e.g. voice onset time). Each of them allows proper identification of the phoneme. A single missing cue can be compensated by the others, so no single cue is truly necessary (Liberman et al., 1967). The effects of manipulating one cue can be counteracted, by manipulating another cue.

On the other hand, slightly offsetting two cues also has the potential to combine and produce a perceptual shift (Fitch et al., 1980).

To summarize, the speech signal is highly varied, and mapping the continuous acoustic signal onto discrete categories is one of the major cognitive processes required for speech perception and in need of explanation. General auditory perception is typically continuous and perceives small changes in the signal. The categorical nature of phoneme perception does not fit with this process. How can speech perception be the same as general auditory perception? How can phonemes possibly be described in purely acoustic terms? And how are listeners capable of perceiving speech and assigning phoneme categories to the hugely variable signal? Thus, a new idea about the processing of speech was put forward.

2.2 Main Claims of the Original and Revised Motor Theory

Because of this massive variance in the acoustic signal and the significant overlap of phonemes due to coarticulation, Liberman et al. (1967) proposed that the invariant object of speech perception and mental representation of phonemes are not acoustic in nature. Every phoneme can be broken down into subphonemic features, where one feature is equivalent to a state of one articulator (e.g. vocal fold vibration, closed velum, rounded lips, etc.), and each articulator can be moved independently. So if every feature can be equated to a motor command sent to an articulator, then these commands should show the invariance needed to code the phonetic segments. For example, while /di/ and /du/ will show very different acoustic patterns, one having a rising formant transition and one having a falling transition and no possibility of perceiving /d/ without these transitions, the features i.e. the motor commands for producing /d/ will always be the same. The acoustic signal then codes the information about the state of the articulators as well their change of state over time. Therefore, in order to perceive the phoneme one needs to extract these commands in a kind of backward production process.

Speech perception should, therefore, be a special form of perception where the acoustic input stream is analyzed for extracting the motor patterns that most likely produced the acoustic patterns and as such be very different from general auditory perception. Through this special mechanism, the perception of speech sounds can be categorical while the perception of other acoustic inputs can be perceived continuously. This is also how Liberman et al. (1967) explain that in the right context, formant transitions sound like speech and otherwise like chirps.

A tight connection between speech production and perception areas is thus assumed. The nature of this connection is described as "overlapping activity of several neural networks - those that supply control signals to the articulators and those that process incoming neural patterns from the ear" (Liberman et al., 1967, p. 454). The most likely cause for this connection is learned association in infancy.

> [...] the motor theory begins with the assumption that coarticulation, and the resulting overlap of phonetic information in the acoustic pattern, is a consequence of the efficient processes by which discrete phonetic gestures are realized in the behavior of more or less independent articulators. The theory suggests, then, that an equally efficient perceptual process might use the resulting acoustic pattern to recover the discrete gestures. Liberman and Mattingly (1985, p. 14)

This idea was revised in Liberman and Mattingly (1985) in two main ways. First, the objects of speech perception are no longer the gestures themselves, but the "intended" gestures, because only the intended gestures are truly invariant. The invariant representations of the speech sounds were, therefore, proposed to be higher up in the processing hierarchy and, following Liberman and Mattingly (1985), should be found in the "more remote structures that control the movements" (p. 23). These remote structures would be the motor system.

Second, speech perception and production are linked, not through learned association, but through a special module that is innate and separate from general auditory perception. This revised theory, therefore, reflects the trends in research by integrating modularity (Fodor, 1983) and the idea of innate language mechanisms made popular by Chomsky (1968). Based on the concept of modularity, this speech perception module would be no different than other specialized modules that take sensory input and calculate a complex percept like sound localization and depth perception. Additionally, this phonetic module is proposed to compete with general auditory perception, leading to the double perception effect, where a speech sound and a chirp are perceived in parallel.

2.3 Critique and Shortcomings

The motor theory was met with critique even in the earliest days. Despite the appeal of the theory to researchers outside the speech perception field, some simple counterexamples show that the motor theory of speech perception, at least in the form stated above, does not hold up.

Perception without production

One of the key predictions of the motor theory is that perception and production are tightly linked and that perception depends on production processes. So, naturally, we should assume that speech perception should not be possible without the ability to produce speech. But data from different populations without production abilities demonstrate the opposite.

Various pathological conditions demonstrate that perception usually works well despite production impairments. Broca's aphasia, for example, is characterized by non-fluent speech, where patients struggle with finding words, producing grammatical structures (telegraphic speech), repeating words, and producing precise articulation (adding or omitting phonetic features). Difficulties in speech perception are, if at all present, usually limited to syntactic structure (Damasio, 1992). MacNeilage et al. (1967) studied a patient with severe impairments in the movement of the oral and laryngeal muscles. Despite the resulting inability to articulate, speech perception was not affected. Patients with aphemia, a condition characterized by articulation difficulties likely due to lesions in the motor cortex, also do not show signs of impaired speech perception (Fox et al., 2001).

Another case contradicting the motor theory, is preverbal infants. Infants typically start to babble at around six months of age and only start producing words at 12 months. Still, preverbal infants show signs of categorical perception for consonants (Eimas et al., 1971) and vowels (Aldridge et al., 2001) and show a general asymmetry between perception and production abilities.

Categorical perception is not exclusive to speech

Categorical perception, one of the main reasons why a separate motor-based perception mechanism for speech was proposed, is also not unique to speech (see Repp, 1984, Kronrod et al., 2016). Categorical perception of acoustic signals has been demonstrated even in the 1960s (e.g. Kopp & Udin, 1969). Cross et al. (1965) used a visual continuum to show that categorical perception is not special to speech perception or general auditory perception. Categorical perception has also been demonstrated for familiar faces (Angeli et al., 2008) and facial expressions (Etcoff & Magee, 1992). In fact, categorical perception can apply to any arbitrary object categories, so that objects that are grouped together are also perceived to be more similar (Goldstone et al., 2001).

Categorical perception is not exclusive to humans

Lastly, categorical perception is also not unique to humans. It has been demonstrated in various animal species, e.g. in chinchillas (Kuhl and Miller, 1975, Kuhl, 1981), monkeys (Kuhl and Padden, 1983, Morse and Snowdon, 1975) and several types of birds, e.g. Japanese quails (Kluender & Lotto, 1994), budgerigars (Dooling et al., 1989, Dent et al., 1997). None of these animals have the physiological capabilities to produce human speech sounds, yet their discrimination mirrors that of humans. These boundaries may exist because of a natural property of the acoustic signal or because of general mechanisms of perception that we share with animals. Either way, if categorical perception had anything to do with articulation, then these observations should not be possible

Two things are clear about phoneme perception: i) categorical perception, though it is an interesting phenomenon that any theory of speech perception must account for, is neither special to the domain of language nor to humans as a species, ii) phoneme production abilities, though they may be linked to perception, cannot be essential to phoneme perception. While it does not mean that production and perception need to be entirely separate, or, that production processes cannot have an influence on perceptual processes, research into motor involvement can only assume a facilitatory role of the motor system.

3 The Human Motor System and Its Role in Speech Production

The discovery of mirror neurons led to renewed interest in the motor theory. The finding that some neurons fire both when an action is observed in another subject or performed by the subject themself, seemed to parallel the perception mechanism proposed in the motor theory. Some of the methods used to study mirror neurons were transferred to speech perception (e.g. Fadiga et al., 2002) and led to new research in the field, which now no longer relied on pathological cases but made use of methods like repetitive transcranial magnetic stimulation. Before going through their findings, one should have a good understanding of how the human motor system is organized. This chapter will, therefore, describe the anatomy and function of the human motor system and how it contributes to speech production.

3.1 The Motor System Hierarchy

The motor system is organized in a hierarchical way, consisting of the spinal cord, the brainstem, the motor cortex, and the association cortex with "side loops" into the cerebellum and the basal ganglia (Nolte, 2009, see also figure 3.1). The spinal cord is responsible for low-level commands that control the exact positioning of the individual muscles and muscle groups as well as rhythmic behaviors like walking. The lower motor neurons connect the brainstem and spinal cord with the muscles of the body. While the motor neurons responsible for the movement of the limbs and body are located in the spinal cord (*corticospinal system*), the motor neurons controlling head and face muscles are situated in the brain stem (*corticobulbar system*).

The motor cortex controls voluntary movements on a larger scale than the brainstem and spinal cord. It manages whole body parts instead of individual muscles and coordinates

more complex movements and movement sequences. It will be described in more detail in section 3.2.

The association cortex, though not part of the motor areas itself, is crucial for the proper response of the motor cortex to the environment. The prefrontal cortex plays a role in action selection based on behavioral context, while the posterior parietal cortex is involved in matching the correct action to an object. The association cortex, the cerebellum, and the basal ganglia all modulate the motor cortex. They do not produce motor output themselves but regulate the motor and premotor cortex.

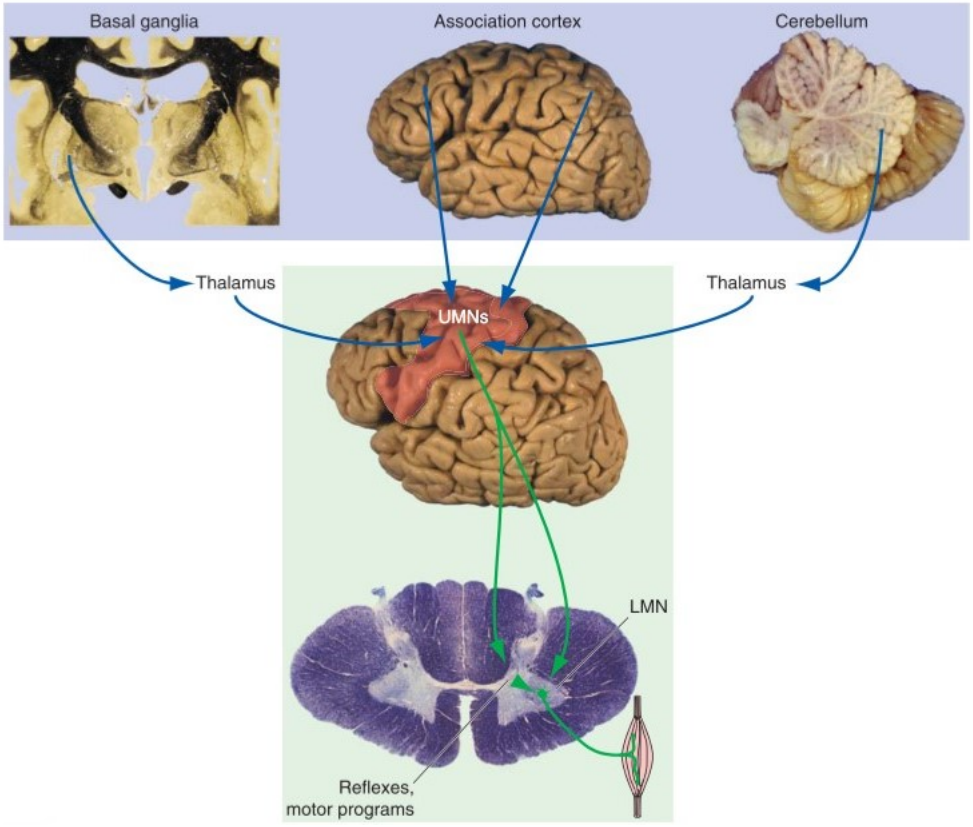


Figure 3.1: Schematic overview over the components of the motor system hierarchy (Nolte, 2009, p. 461)

3.2 The Motor Cortex

The motor cortex consists of the primary motor cortex, premotor cortex, and supplementary motor areas (Nolte, 2009, Geyer et al., 2012).

Primary motor cortex (M1)

The primary motor cortex is located in the precentral gyrus and is somatotopically organized, with every body part being associated with a specific region in the motor cortex. This somatotopic organization can be seen in figure 3.2. The primary motor cortex codes movement on a larger scale by controlling the direction of movement, the extent of movement, and the speed of movement. The motor neurons in the primary motor cortex fire 5-100 ms before the onset of movement and forward the motor command to the lower-level motor neurons that control the contraction of individual muscles (Tremblay et al., 2016). Crucially, the M1 is not the only starting point of corticospinal axons, but they also arise in PM, SMA, and somatosensory cortex (Nolte, 2009).

Premotor cortex (PM)

The premotor cortex is located anterior to the primary motor cortex and is similarly organized. It can be divided into a dorsal and ventral part also referred to as PMd and PMv respectively. The premotor cortex is associated with the coordination of movements involving large muscle groups and the learning of new visuomotor associations (Buccino et al., 2001, Binkofski and Buccino, 2006). The premotor cortex signals the preparation and suppression of movements that are internally initiated (Weeks et al., 2001, Wiese et al., 2004) and seems to be sensitive to the behavioral context of movements (Iacoboni et al., 2005).

Supplementary motor area (SMA)

The supplementary motor area lies anterior to the portion of the primary motor cortex that controls the feet and can be divided into SMA-proper and pre-SMA. It is associated with sequences of movements, e.g. complex finger movements (Roland et al., 1980), and learning of such sequences as well as executive control functions such as adjusting or aborting actions (Nachev et al., 2008).

Somatosensory cortex (S1)

The somatosensory cortex lies posterior to the primary motor cortex in the postcentral gyrus. The primary somatosensory area is equally somatotopically organized as the primary motor area and is responsible for tactile processing and proprioception (Hill et al., 2019).

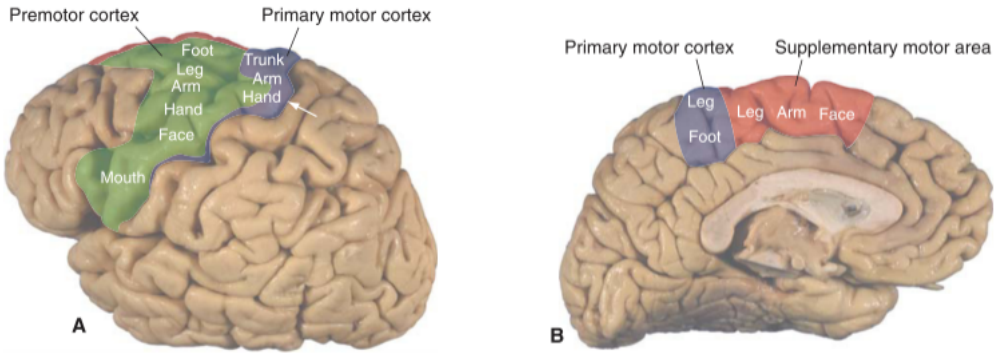


Figure 3.2: Somatotopic organization of the primary motor cortex, premotor cortex, and supplementary motor area (Nolte, 2009, p. 464).

3.3 The Role of the Motor System in Speech Production

To better understand, how language and motor areas interact, the neural correlates of speech production will be summarized here, starting with lexical-phonemic translation up to articulation. This involves a network of brain regions including the ventral primary motor cortex (M1v), ventral premotor cortex (PMv), supplementary motor area (SMA), cingulate motor area (CMA), the insula, the basal ganglia, cerebellum and the thalamus (Tremblay et al., 2016). Inferior to the motor cortex, the areas typically attributed to language are centered around the Sylvian fissure in the inferior frontal gyrus (IFG) and superior temporal gyrus (STG), also known as Broca's area and Wernicke's area respectively (Hill et al., 2019). Figure 3.3 shows a framework of the different processing components necessary for speech production.

Phonetic and phonological encoding is a crucial step in transforming thought into speech. Phonological encoding retrieves the sound structure of a word or phrase, which then gets translated into gestural commands during phonetic encoding. Both Broca's area and Wernicke's area have been associated with these processes. Indefrey and Levelt (2000) reviewed a series of imaging studies on different speech production tasks and identified the left posterior IFG and the left middle STG as locations, which were active during word repetition and pseudoword reading, though they may be differently involved in these processes. Left posterior IFG was more active in pseudoword reading, suggesting it may be more involved in phonological encoding. Non-lexical phonological encoding has been attributed mainly to Broca's area, though Papoutsi et al. (2009) argue that it is

only involved in phonetic, but not phonological encoding. Traditionally, the motor cortex has not been associated with this task, but only with the execution of the articulatory patterns.

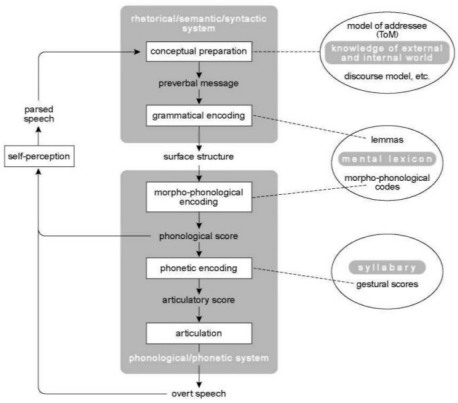


Figure 3.3: Processing components in speech production (Levelt, 1999)

In particular, activity in the cingulate motor area, pre-SMA, and the ventral PM seems to be modulated by manipulation of response selection (Tremblay and Gracco, 2006, Nagel et al., 2008). Pre-SMA stands out as it appears to be active during word selection but also non-linguistic oral gestures, as transcranial magnetic stimulation of pre-SMA impairs the selection of both words and other oral gestures (Tremblay & Gracco, 2009).

Indefrey and Levelt (2000) found that SMA was also active in non-overt production tasks like silent word generation and verbal working memory, which suggests that SMA might be involved in motor planning. SMA has also been proposed to play a role in action sequencing (Tanji and Shima, 1994, Shima and Tanji, 2000) and motor timing (Brendel et al., 2010). PM, basal ganglia, and cerebellum have also been identified to show increased activity with increasing difficulty of a produced syllable sequence (Bohland & Guenther, 2006). PM has also been associated with the learning of new motor sequences (Segawa et al., 2015).

Another aspect of speech production is the initiation of the articulatory movements. Action initiation has been associated with activity in the SMA and pre-SMA as well as the basal ganglia (Wiese et al., 2004).

Articulation forms the last step in the speech production pipeline. Here, the primary motor cortex, in particular the ventral part, controls the vocal tract. For speech production, a range of muscles in the oral cavity, larynx and pharynx are required as well as abdominal

muscles for breath control. Therefore, both the corticospinal tract, reaching the spinal cord relevant for the neck and abdominal muscles, and the corticobulbar tract, reaching the brain stem which controls the laryngeal and supralayringeal muscles, are crucial for proper articulation (Tremblay et al., 2016). But the primary motor cortex is not the only cortical area connected to the brain stem and spinal cord. The premotor cortex and supplementary motor area also contain a large amount of corticospinal and corticobulbar neurons which allows them to potentially generate movement and control it independently of the primary motor cortex. In response to electrical stimulation of the SMA, for example, vocalizations or slowed speech and speech arrest could be observed (Fried et al., 1991). Motor arrest also occurs if PMC or the somatosensory cortex is stimulated (Fornia et al., 2020).

4 Motor Cortex Activity During Passive Listening

Despite the fact, that the motor system is mainly involved in speech production, a series of studies show activity in the motor system during passive listening. Evidence of this kind comes from studies using fMRI and TMS and show that even in the absence of speech production the motor system is active to some degree.

4.1 Evidence From fMRI

Functional Magnetic Resonance Imaging offers a chance to visualize the anatomy of the human brain, as well as locate regions that are active in a specific condition by tracing blood oxygenation levels. Through this method, it has been possible to map the relevant brain areas (including motor areas) during speech production, but some studies have found a similar activation during passive listening tasks.

Benson et al. (2001) tried to distinguish brain areas involved in speech perception, by testing stimuli that contrasted in various aspects (speech vs. non-speech, acoustic vs. phonetic complexity, natural vs. synthetic speech) and found several brain regions with activity correlated to the speech vs. non-speech and complexity contrasts, including areas in the superior temporal gyrus, posterior supramarginal gyrus, and middle frontal gyrus. Activity in the premotor cortex increased in response to natural non-speech stimuli over synthetic ones. A similar region bordering PM and the middle frontal gyrus, responded to the natural speech stimuli over synthetic ones, as well as speech over non-speech.

Wilson et al. (2004) found increased activity in the PMv (BA 6 & 4a) during passive listening to meaningless monosyllables compared to resting. Compared to the regions active during the production of the same syllables, the activated areas showed significant overlap. This activity during passive listening was also larger than for listening to non-speech sounds.

Wilson et al. (2006) found activity in both auditory and motor areas (border of PM and M1) in response to passive listening to native and non-native syllables, which increased for non-native syllables. Crucially, for non-native phonemes only the activity in the auditory cortex seemed to be correlated to the producibility of the syllables, indicating that the motor system may only offer phonemic categories as candidates based on its own motor patterns.

Pulvermüller et al. (2006) also found activation in the primary motor cortex during passive listening to /p/ and /t/ syllables and found that the differential activity for /p/ vs. /t/ followed the somatotopic organization of the primary and premotor cortex. They interpret their results as evidence that motor circuits are recruited to reflect the phonetic features of the perceived phoneme.

Callan et al. (2010) used fMRI and MEG to compare correct and incorrect trials of phoneme identification (/ba/, /bo/, /da/, /do/) in noisy listening conditions and found activity in the ventral premotor cortex and neighboring pars opercularis that was greater for correct than for incorrect trials.

Lee et al. (2012) used a multivariate pattern-based analysis (MVPA) to see which brain regions would respond with distinct activity patterns to two contrasting phonemes. The two areas that responded accordingly were Broca's area and the left pre-SMA.

Grabski et al. (2013) researched whether single vowel production and perception in the absence of syllable sequencing and coarticulation induced the same sensorimotor activity. They found activity in the right ventral premotor cortex and the left superior frontal gyrus (BA 9).

Evans and Davis (2015) also found activity in the precentral and postcentral gyrus in response to listening to syllables compared to a silent rest period. These regions also responded more to degraded speech than clear speech, indicating they might get recruited when identification becomes more difficult. The areas also overlapped with those recorded for syllable production. They also developed an RSA model to check clusters for phonological identity, meaning syllables that represented the same phonemes in different versions, e.g. different speakers, and found clusters in the mid posterior superior temporal gyrus, middle temporal gyrus, superior temporal sulcus, and primary auditory cortex, but also in the precentral and postcentral gyrus.

Du et al. (2014) tested the identification of four different syllables in six different signal-to-noise ratios and found that lower accuracy correlated with higher activity in the premotor

cortex and Broca's area. They conclude that the motor system may be more robust to noisy listening conditions than the auditory cortex.

Correia et al. (2015) explored whether neural correlates for articulatory features could be found, e.g. place or manner. They trained a multivariate classifier on the syllables /pa/ vs. /ta/ and tested if it would generalize to fricative /fa/ and /sa/. They found such generalizations for place and manner in auditory, motor, and somatosensory regions. The right inferior precentral gyrus was identified as a cluster for the place of articulation.

Longcamp et al. (2019) found activity in the lateral and medial portion of the premotor cortex for both the production and perception of spoken and written language. During listening, activity from the inferior frontal gyrus extended all the way up to the ventral premotor cortex as well as the supplementary motor cortex. This was found for both listening and speaking. Similarly, for reading and listening activity overlap was found in the left ventral premotor cortex extending to the pars opercularis.

Grabski and Sato (2020) used a repetition suppression paradigm to find neural substrates of phoneme coding in listening and speaking and found a common suppressed response cluster in the pre-SMA and neighboring cingulate cortex as well as the left precentral gyrus and neighboring inferior frontal gyrus. They interpret their results as evidence that phonemic representations are coded in shared sensorimotor regions for production and perception. This activity, however, might be more pronounced during audiovisual integration, as Skipper et al. (2005) show when comparing the activity during audio-only, visual-only, and audiovisual perception. This fits with the previous findings suggesting the premotor cortex is responsible for integrating external information for response selection.

There are, however, a series of studies that could not confirm these findings (Szenkovits et al., 2012, Arsenault and Buchsbaum, 2015, Arsenault and Buchsbaum, 2016 Menenti et al., 2011). Schomers and Pulvermüller (2016) argue that these discrepancies are of methodological nature. They identify two factors that supposedly explain the differing results. The first factor is scanner noise, as the noise inherently present during MRI scans is masking the speech stimuli and making identification more difficult. Why exactly that should lead to less motor involvement is not explained, but simply based on observation and comparison of similar studies and their designs. This explanation seems a bit odd, considering that the motor system has also been proposed to be recruited specifically in noisy listening conditions and is backed by a hand-full of studies that found increased activity in motor areas in noisy listening (Murakami et al., 2012, Osnes et al., 2011,

Adank et al., 2012, Hervais-Adelman et al., 2012). The second factor that they identified is the presence of overt motor tasks during the procedure. Specifically, those studies listed by Schomers and Pulvermüller (2016) that could not find activation of motor areas would ask their participants to constantly press buttons during the procedure, which could lead to constant preparatory activity in the motor cortex. The hand and mouth representations in the primary and premotor cortex are adjacent so that the preparatory activity of the hand might transfer to the mouth representation leading to a ceiling effect which makes it hard to detect smaller speech-related activity in the articulatory motor cortex.

4.2 Evidence From TMS

But activity found in fMRI is not enough evidence to say that the motor system is involved in a causal way. Transcranial magnetic stimulation (TMS) offers more insights.

If electrical current flows through a coil, it induces a magnetic field. When such a coil is placed near a person's scalp, a strong, but brief magnetic field can be used for stimulation by inducing electrical currents in the cortical brain areas. This can have an excitatory or inhibitory effect, which, in turn, can be measured through motor-evoked potentials (MEPs). These MEPs reflect activity in the primary motor cortex and "depolarisation of neurons in the descending motor tract" (Adank et al., 2017, p. 901).

The discovery of mirror neurons in monkeys led to new interest in the motor theory and new study designs that are heavily linked to the exploration of motor simulation. TMS studies provided indirect evidence for the existence of mirror neurons in humans, e.g. Baldissera et al. (2001), who found MEPs in the finger flexor muscle during the observation of hand gestures. A similar result could be obtained for the tongue muscle during passive listening of words and non-words containing a distinct trill phoneme ("rr" as in Italian *terra*) (Fadiga et al., 2002). This was the first evidence of activation of motor areas in an articulator-specific manner.

Watkins et al. (2003) applied TMS to the cortical lip areas and measured MEPs at the lip. MEPs were enhanced when listening to speech versus non-speech. Roy et al. (2008) did a follow-up investigation to find out whether the difference found by Fadiga et al. (2002) between words and pseudowords was due to lexical frequency or meaning. They found MEPS around 100 ms after a double consonant embedded in pseudowords and also at 200-300 ms after a double consonant embedded in words with low frequency.

D'Ausilio et al. (2009) applied double pulse TMS to the lip and tongue motor cortex and found a double dissociation between the stimulation site and discrimination performance. When participants had their tongue motor area stimulated their reaction time to dental sounds improved compared to their non-TMS time, while it worsened for labial sounds and vice versa. Also the rate of labial sounds classified as dental sounds was higher for tongue motor cortex TMS than without. Also, the error rate of dental sounds classified as labial was lower and vice versa.

D'Ausilio et al. (2011) applied TMS to the tongue motor cortex and measured MEPs in the tongue for different expected double consonants. They found enhanced MEPs when participants expected tongue-related consonants compared to non-tongue-related ones, but only when the preceding syllable included tongue-related coarticulatory information.

Schmitz et al. (2019) applied TMS to the left motor cortex and measured the excitability of the lip muscle during passive listening to native and non-native vowels. The excitability showed a negative correlation to nativeness ratings, suggesting a facilitatory role of the motor system in the perception of non-native speech sounds.

Combined, the fMRI and TMS studies show activity in primary motor areas and premotor areas. While fMRI data shows more variance and can be considered simple coactivation of tightly linked brain regions, the TMS data paints a different picture. It shows that not only the level of motor activity varies in an articulator-specific manner, but also the participants' behavioral data.

5 Mechanisms Linking Language Perception and Production

But how would perception and production areas be linked? Since the discovery of mirror neurons is responsible for the revival of the motor theory and the resulting discussion, one might argue that mirror neurons are the perfect mechanism for explaining motor contribution to speech perception. After all, mirror neurons have been proposed to be the mechanism behind action understanding (Rizzolatti & Craighero, 2004), which might be exactly what is happening here for speech perception. Mirror neurons fire both during action execution and action observation. However, the whole idea that mirror neurons serve action understanding was apparently inspired by the already existing motor theory of speech perception (Rogalsky et al., 2011).: "In some ways, the existence of a well-known motor theory of perception lent indirect support for the claim in monkeys that mirror neurons were the basis of action understanding" (p.179).

Hickok (2013) argues against the interpretation that mirror neurons serve action understanding. Mirror neurons, he argues, do not seem to code the action itself but more abstractly the goal of the action. If mirror neurons respond to abstract goals of the action, how does the brain know what this goal is or when it has reached it? It receives that information through sensory input. He argues, that action understanding is therefore fundamentally a process of the perceptual system. Mirror neurons may reflect action understanding but not contribute to it.

Instead, this link between the sensory and motor areas may be due to simple associative learning (Cook et al., 2014). Since every action of the body leads to a sensory outcome, whether tactile, proprioceptive, or auditory, the motor command and its resulting sensory outcome would form a strong neural connection. Evidence for sensorimotor associative learning can be found when looking at infant studies.

Some influential studies by Kuhl and Meltzoff (1982) and Kuhl and Meltzoff (1984) found that infants aged 4-5 months prefer to look at faces that articulate a vowel congruent with

an acoustically presented vowel than faces that articulate an incongruent vowel. This finding, that infants are able to link the acoustic signal to the articulatory movement, has been replicated and seems to be robust across speaker-specific traits like gender (Patterson & Werker, 1999). It has even been observed in 2-month-old infants (Patterson & Werker, 2003), suggesting that infants learn to make this connection very quickly after birth. A study by Mugitani et al. (2008) suggests that the productive abilities of the infant could be an important factor for this type of audiovisual matching. Instead of vowels, they used a bilabial trill (vibration of both lips that creates a sound similar to *brrrr*) and a whistle. Both sounds are bilabial, can be produced for longer than typical consonants and are not part of the phoneme repertoire of most languages, so infants should not have a lot of exposure to hearing either sound. The crucial difference between the two sounds is, that at 8 months of age, infants frequently produce the bilabial trill while they do not produce the whistle. While the tested infants looked longer at matching faces for the bilabial trill, they showed no signs of matching the whistle to the articulating face.

Additionally, many of the above-mentioned studies reported infants imitating the congruent face/voice stimuli. This phenomenon of imitating facial or manual gestures can even be observed in infants as young as 12-21 days (Meltzoff & Moore, 1977). With a pacifier in their mouth, the infants would not imitate directly during stimulus presentation but rather suckle on the pacifier. When the pacifier was removed, however, the infants still showed imitative behavior despite the stimulus not being present anymore, suggesting this imitation may be more than an innate reflex and possibly an active process of matching visual and proprioceptive information.

The finding that infants perceive speech in a multimodal way is also supported by the fact that the McGurk effect[1], where a mismatch of auditory and visual information leads to an intermediate percept, can also be observed in 4.5-month-old infants (Burnham & Dodd, 2004). This also functions as evidence of proper audiovisual integration in infants, which has since also been demonstrated in 10-week-old infants (Bristow et al., 2008) Although the binding window narrows with age so that smaller levels of asynchrony are detectable (Lewkowicz & Flom, 2014), the ability to bind visual and auditory input seems to be present in the first weeks of life.

To summarize, it has been shown that infants can match and integrate auditory and

1 The McGurk effect, is a phenomenon that occurs when auditory and visual input do not match. In the original experiment, participants were shown the silent video of a woman articulating /ga/, but heard the sound /ba/. Participants report hearing /da/ (Mcgurk & Macdonald, 1976).

visual speech stimuli and that they have a tendency to mimic the facial gestures of the observed person. As Mugitani et al. (2008) suggested, productive abilities might play a big role in this integration process. A handful of studies have now found that the availability of sensorimotor information may be crucial for phoneme perception in infants. Yeung and Werker (2013) tested 4.5-month-old infants on their audiovisual matching ability while forcing their mouths into specific positions. They presented the infants with faces, one producing /i/ and one producing /u/, while either /i/ or /u/ was played in synchrony. As with previous experiments, infants of the baseline group (no manipulation of mouth position) would look longer at the face that matched the heard vowel. For the other infants, their mouth was put in a similar position to the lip position during articulation of /i/ or /u/ (chewing on a teething ring for wide-spread lips similar to producing /i/ or sucking on a finger or pacifier for rounded lips similar to producing /u/). They were put in either a lip-matching group, where the infant's lip position matched the vowel they heard, or a lip-mismatch group, where the infant's lip position did not match the heard vowel. They, too, were presented with two faces, one articulating a matching vowel and one articulating a mismatching vowel, and their looking times were recorded. The looking times of the mismatch group did not differ from the looking times of the baseline group, indicating no problems during audiovisual matching. The infants in the lip-match group, however, did not show this pattern, but instead performed below chance.

Yeung and Werker (2013) proposed two possible explanations, which they tested in their second experiment. One possibility was sensorimotor information led to a looking time bias away from the similar face, while the second option would be an interaction between motor processing and audiovisual processing. In the first case, the infants simply preferred to look at the face that was dissimilar to their own mouth position, which for the mismatch group happened to be congruent with the heard vowel, while in the second case, the sensorimotor input overrides or suppresses information about the vowel it is similar to and therefore interferes with audiovisual matching. To test this, they repeated the previous experiment but played an /a/-vowel instead of /i/ or /u/. If the sensorimotor information led to looking bias away from the matching face, then that bias should still be found when a different sound is played. If the sensorimotor information interferes with audiovisual integration, then the looking patterns should be roughly at chance level, as the vowel /a/ is unrelated to both the articulating faces and the infant's mouth position. The results support the second account, with infants' looking patterns

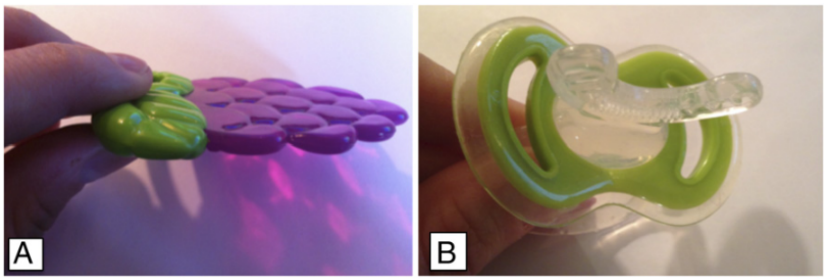


Figure 5.1: Teething toys used by Bruderer et al. (2015) and Choi et al. (2019). A) flat teething toy. B) gum teething toy

.

being different from those in the first experiment and infants' proportional looking time towards the lip-matching face being slightly above chance.

Another study by Bruderer et al. (2015) tested 6-month-old infants on their discrimination of non-native dental /d̪/ and retroflex /ɖ/. In the first months of life, infants still perceive subtle contrasts between phonemes that are not part of their native language and slowly lose that ability as they form more robust categories (Werker et al., 1981, Werker and Tees, 1984). This discrimination can be tested in an oddball paradigm, where the infant is presented with one type of stimulus multiple times until they start to lose interest, followed by a different stimulus which, if the difference is recognized, should lead to renewed interest. Here blocks of alternating or non-alternating stimuli were presented and the looking time was compared as an indicator of discrimination.

In this study, the infants received either a flat teething toy in their mouth that would block tongue movement or a different teething toy that would sit on the gums but allow the tongue to move freely (shown in figure 5.1). For both infants without a teething toy and infants with the second type of teether, the looking times differed enough to indicate discrimination. For the infants with the tongue-blocking teether, this was not the case. This experiment was replicated by Choi et al. (2019) and extended to a new phoneme contrast. This time the contrast between the native categories /b/ and /d/, one being bilabial and the other alveolar, was used to test discrimination. Infants without a teething toy in their mouth showed signs of discrimination, while those with the second teething toy (affecting lip movement, but not tongue movement) did not show signs of discrimination. Crucially, this is the same teething toy that did not affect discrimination

of /d̪/ and /ɖ/ in the study by Bruderer et al. (2015). These findings are taken by the authors to point towards an articulator-specific influence of sensorimotor information in the perception of speech sound categories in infants.

In summary, these findings not only support sensorimotor associative learning in infants but also show that their productive capabilities do have an effect on speech perception, at least in infancy. In this context, it is also interesting to see that perceptual narrowing has its onset at a similar time as babbling. This sensorimotor association might therefore be an important factor in forming phoneme categories.

6 Possible Roles of the Motor System in Speech Perception

If sensory and motor areas are linked primarily through associative learning, it remains open, to what extent the motor system should be involved in perceptual processes and in particular in speech perception. Does it merely get coactivated due to the strong connection between motor and sensory areas or could the motor system give input to the perceptual system?

6.1 Motor System Recruitment in Noisy Listening Conditions

One of the main arguments against the original motor theory of speech perception (e.g. Lotto et al., 2009) has been evidence from aphasic patients, where a loss of productive abilities often does not lead to deficits in perception (Damasio, 1992, Hillis, 2007).

One finding following a series of studies on pathological cases was impaired speech perception in impoverished listening conditions. Moineau et al. (2005), for example, find that lexical comprehension in altered listening conditions is significantly worse for patients with Broca's aphasia than for an age-matched healthy control group. The alterations used were low-pass filtering and time compression. Also, response times were slowest for Broca's aphasics compared to all other groups.

This led to the proposal that the motor system might get recruited for speech perception in cases where the auditory signal is impoverished or where cognitive demand is high.

Meister et al. (2007) applied rTMS to the premotor cortex and tested participants' discrimination of stop consonants (/pa/, /ta/, /ka/) masked by white noise. Performance under the influence of rTMS was worse than without it and performance on a non-linguistic task was also not impaired by the rTMS (color discrimination and tone discrimination).

The lack of effect on the control tasks suggests that the effect of rTMS applied to PMv is not an effect on general response selection.

D'Ausilio et al. (2012) use event-related double-pulse TMS to the primary motor areas for the lip or tongue representation and tested participants' phoneme identification on phonemes, which were either masked with noise or without noise. No effect was found for the no-noise condition, but a double dissociation was found for the noise condition. This means that for labial phonemes, participants performed worse when they received stimulation to the lip motor cortex compared to dental phonemes, and the same for dental phonemes when participants received TMS to the tongue motor cortex compared to labial phonemes. But crucially these effects were only found comparing phonemes in noisy listening conditions.

Stokes et al. (2019), however, argue against this proposal. They test phoneme perception at a continuous signal-to-noise ratio. If the motor system was recruited specifically due to noise, then performance with impaired motor capabilities should also reflect a falling signal-to-noise ratio. A psychometric curve (essentially a logistic curve) was fit to every participant's performance in a phoneme identification task on minimal pair words. Their argument is, that if you test at a specific signal-to-noise ratio you might get a big effect, while at other levels no effect might be there. Four different conditions were used: articulatory suppression, mandible movement, foot tapping, and passive listening. They found a small effect of articulatory suppression on the midpoint of the psychometric function, leading to a shift, therefore slightly improving performance in noise but only to a specific level.

6.2 Motor System Recruitment for General Speech Perception

But there are also studies supporting the involvement of the motor system in speech perception without additional noise. As described in the beginning, the classic motor theory of speech perception, which proposed that the motor system gets recruited for speech perception and the percept is the vocal gesture, does not hold up. It has become clear that speech perception does not depend on the recruitment of motor areas. Considering the recent development regarding embodiment research, however, it seems a little hasty to say, the motor system should not be involved at all. Galantucci et al. (2006), for example, argue that the core idea of the motor system being involved in

speech perception fits well with the then-blooming new research around mirror neurons. Studies like Fadiga et al. (2002) were taken as a prime example for showing that speech perception may be just another form of action understanding, which was proposed to rely on mirror neurons. Activation of motor areas during passive listening, however, does not mean that these areas actually contribute to speech perception and are not merely coactivated. A selection of studies suggests that the role of the motor system goes beyond coactivation and may actually interact with auditory processes during speech perception.

Schomers et al. (2015) applied TMS pulses to the primary motor cortex of either the lips or the tongue while participants had to listen to single words starting with either a bilabial or an alveolar consonant (e.g. "pool" vs. "tool") and found that responses were delayed for incongruent trials compared to congruent ones.

Mochida et al. (2013) compared the effect of articulatory movements of participants to the effect of watching another person articulate a syllable, that is either congruent or non-congruent with a sound being played. When participants articulated /ta/ or /ka/ it had a negative effect on the identification of /ta/ and /ka/, but not for /pa/. They argue that this is because /pa/ is associated with the lips and /ta/ and /ka/are both associated with the tongue and therefore affect each other. /ta/ and /ka/ were affected by watching another person articulate /pa/. This is the classic McGurk effect and does not seem to be articulator specific. They argue, therefore, that the effects point to two different processes, not the same one.

Möttönen and Watkins (2009) used rTMS to the lip primary motor cortex led to disruption of categorization of articulator-specific phonemes (ba-da, pa-ta, but not ga-ka or ga-da). They argue that mapping the auditory signal onto articulatory representations may aid in phoneme categorization. Möttönen et al. (2013) used rTMS and EEG to show that disruption of the lip representation suppressed responses to changes in speech sounds. The aim was to see if the motor system modulates auditory speech processing in the absence of high demand tasks. Participants would watch a silent video and listen to a sequence of speech sounds. When they received rTMS to the lip primary motor cortex, they did not notice changes in the speech sounds but did notice changes in a sequence of piano tones. The response, which can usually be seen in EEG (N400) was suppressed in the rTMS condition. When rTMS was applied to the hand representation, however, no such effect was found.

Smalle et al. (2015) tried to disambiguate the effect of motor cortex stimulation on

perceptual sensitivity and response bias. They found that stimulation of the lip led to lower sensitivity for between-category pair, while disruption of the hand led to a reduced response bias.

Berent et al. (2020) used stimuli that were ambiguous in voicing (ba/pa or da/ta) and let participants gently bite their lip or tongue tip. Congruent stimulation "shifted the voicing percept and improved discrimination sensitivity". Participants would hear the voiced phoneme more often than the unvoiced one. Because the stimulated articulator is relevant for both choices, the authors argue that the effect has to be perceptual and not response selection.

Berent and Platt (2022) used the same type of stimulation to test the effects on phonological preferences. The idea is that certain syllables are less preferred by people and less common in languages because their pronunciation is more difficult. If perception relied on simulation of articulation, then the perception of such syllables should also be affected by articulator manipulation. Stimulus perception was actually affected by the stimulation, but syllable structure was not. Therefore phonological processing is not governed by embodiment. Phonetic perception, however, seems to be subject to embodiment. They found that gently biting the tongue led to better discrimination of labial stimuli compared to coronal stimuli, but not the other way around.

Looking at these studies, an interactional model of speech perception would be necessary. Pulvermüller (2018) offers such a model by proposing action-perception circuits that get reused for language. Such circuits are formed through Hebbian learning ("cells that fire together, wire together"), as the motor and sensory areas unavoidably get activated roughly simultaneously during all actions including speech production. Once such a circuit is formed, the neurons involved can be active both during production but also perception, with mirror neurons being a result of these circuits. In short such action-perception circuits may be the basis for an interaction of motor and auditory areas.

6.3 Motor System Recruitment in Phoneme Segmentation and Identification Tasks

Crucially, these studies usually test one thing, identification or discrimination of single phonemes or syllables. Only a few (e.g. Schomers et al., 2015) use words as stimuli. Therefore some argue, that the motor system, if involved at all, only has a facilitatory role

specifically in phoneme identification or discrimination. To say, that the motor system is recruited for speech perception may therefore be wrong, because speech perception goes beyond phoneme identification and likely does usually not require identification of single phonemes, but of larger chunks (e.g. syllables).

Krieger-Redwood et al. (2013) applied TMS to three sites (PM, posterior superior temporal gyrus, and occipital pole). They gave participants tasks that would either need phoneme categorization or mapping the sounds onto meaning. They found that the premotor cortex is important for explicit phonological judgments, but not for mapping speech onto meanings.

Sato et al. (2013) applied rTMS over the ventral premotor cortex, which resulted in slower phoneme discrimination compared to sham stimulation when phonemic segmentation was required. No effect was observed in phoneme identification and syllable discrimination tasks that could be performed without the need for phonemic segmentation.

6.4 Motor System Activation as a Byproduct of Self-Monitoring

But what if the motor system did not contribute to speech perception? Since the motor theory and newer research into motor system recruitment have been heavily criticized and considered false hype by many, the findings of the studies mentioned above still remain in need of explanation. Many accounts of speech perception processes do not propose a mechanism that would allow these findings.

One such mechanism, proposed by Hickok (2013), is a self-monitoring loop that connects auditory and motor areas and monitors and potentially corrects speech output during speech production. His *State Feedback Control Model of Speech Production* is a model mainly intended to explain speech production processes but claims to also explain the phenomena seen in speech perception. The proposed feedback system calculates an internal model of the state of the articulators. For that, it uses a copy of the motor command sent to the muscles, also known as corollary discharge. This internal model can then make predictions about the sensory outcome of the action and give feedback if it does not match the intended sound. By also comparing this prediction with the actual resulting outcome, the model gets feedback on the accuracy of its prediction and is kept up to date.

The existence of a feedback loop of this kind fits well with existing data. People with severe hearing loss, for example, will notice a decline in the precision of their articulation (Waldstein, 1990). They would still have access to those motor representations that allow them to articulate, but they might be less precise because they do not get feedback and are not able to update their internal model. It would also explain how people are able to adapt to new language environments (e.g. new accents) and potentially take over some of these characteristics (Delvaux & Soquet, 2007). There are also studies that show that participants will try to correct their articulation when their own speech is presented to them in a distorted way which can actually lead to quick motor learning (e.g. Nasir & Ostry, 2009). But what Nasir and Ostry (2009) also report is that once motor learning had occurred in their participants, their perception of speech sounds also changed systematically.

And, as Hickok argues, it would explain some of the perceptual effects found in relation to the motor system. It would also mean that activity found in the motor cortex during passive listening would be due to this automatic process. But this proposed feedback loop only takes motor commands from the motor system and auditory information from the auditory system and tries to match the predictions with the actual auditory input. Its output is targeted at the motor system by giving information about the corrective measures necessary. On the one hand, passive listening does not coincide with motor commands. The feedback loop would therefore have to give feedback without matching the sound to an internal motor command and, as Hickok proposed regarding quick adaptation to new language environments, integrate this information into the model.

While this might explain the activation of the motor system, it does not account for how the motor command would shift the percept. The findings of Nasir and Ostry (2009) suggest, that once a new association between the motor command and auditory feedback has formed, this update of the model has the power to influence the perceptual process. The internal model may change based on the auditory input, but the categorization of the auditory input would have to be influenced by the prediction of the internal model. In this case, the account of the SFC model exhibits roughly the characteristics of theories proposing motor integration in speech perception.

7 Interim Summary

So far, the emerging picture is not all too clear. Almost all studies specifically tested phoneme identification or discrimination, usually using non-lexical syllables. Some of the findings related to noisy listening conditions, while others tested non-noisy listening conditions. Stimulation sites were usually either primary motor areas or premotor areas. Though a specific articulator was usually targeted, TMS in particular may stimulate adjacent areas.

Interestingly, both stimulation of primary motor areas and premotor areas have been found to have an effect usually in an articulator-specific manner, though it seems that the direction of this effect is not universal. While some studies report an improvement in discrimination for the stimulated articulator, some report the opposite. Either way, these effects need an explanation.

In the broadest sense, the debate can be broken down into models that allow interaction of motor and auditory areas in the perceptual process, and those that do not, which are often considered purely auditory. But if the motor and auditory areas really interact during the perception process, then in principle even simple manipulations of the articulators have an effect on speech perception. Forcing the articulators in a specific position, for example, should activate the corresponding primary motor cortex and primary sensory cortex. This activation should in turn get integrated into the percept. In this sense, the idea of using mechanical stimulation rather than TMS is similar to the studies performed by Berent et al. (2020) and Berent and Platt (2022). Biting of the lip or tongue may, however, activate sensory areas more than motor areas. In order to shift the activation more toward motor areas, a different stimulation technique is necessary.

Part II

Experiment

8 The Current Study

If the articulatory motor system plays a role in phoneme perception, then the stimulation of the articulators should force activation of the relevant motor areas. Through this activation, a change in phoneme perception should be observable, as the relevant motor areas are either blocked from their use in perception or the activation gets integrated into the percept. One such observable change should be the phoneme boundary, i.e. the point or area on a continuous spectrum, where perception shifts from one category to another.

In the following study, a simple wooden tongue depressor is used to force the articulators into a specific position, similar to an articulatory movement. This has two effects. On the one hand, it forces muscle engagement in a way that mimics articulation. On the other hand, it also blocks the articulators from moving freely inside the mouth, which in infant studies (Bruderer et al., 2015, Choi et al., 2019) had an effect on their discrimination abilities. Of course, adult phoneme categories should be robust enough to allow discrimination in the absence of a free range of articulatory movement, but when looking further away from the phoneme stereotype, perception might still be affected.

To find out whether articulator-specific mechanical manipulation would affect the phoneme boundary, 33 adult German native speakers were given a phoneme identification task for three different phoneme continua (/i:/-/y:/, /ba/-/da/, and /ga/-/ka/). Each continuum was rated under three different conditions: i) a baseline with no distractions present, ii) a mechanical stimulation of lips and tongue tip, and iii) a distraction task involving the foot as a control condition.

If the motor system is at least partially involved, manipulation of the articulators should have an effect on the phoneme boundary by either shifting it or widening the boundary range so that the boundary between phonemes is no longer as clear-cut as under normal listening conditions.

For /i:/ and /y:/ the tongue is in a high position, close to the hard palate. This should resemble the position the tongue has in the experimental condition. The production of /y:/ requires rounding of the lips while for /i:/ the lips are not rounded. As the tongue depressor should force the lips into a more rounded position, the articulatory movement of /y:/ resembles this position more than that of /i:/. As the phonemes /i:/ and /y:/ differ only in the rounding of the lips, the expected outcome of the experimental condition is a shift in the phoneme boundary. The direction of this shift depends on how the motor system is used in phoneme processing and can therefore give insight into which theoretical accounts should be preferred. Therefore, hypothesis 1 is as follows:

> H1: For the stimulus continuum /i:/-/y:/, the phoneme boundary is shifted for trials under the experimental condition compared to those under the baseline and the control condition within-subject.

For the phonemes /b/ and /d/, the prediction is different. For /b/, a full closure at the lips is rapidly opened while for /d/, this closure occurs by the tongue pressing against the hard palate. For production, both phonemes require the articulators that are active during the experimental condition and they should therefore be affected equally by the manipulation. It should, therefore, get more difficult to distinguish stimuli close to the phoneme boundary. The boundary should therefore show a widening of the range and should have a more shallow sloped rather than a clear vertical line. Hypothesis 2 is stated as follows:

> H2: For the stimulus continuum /b/-/d/, the phoneme boundary range is wider for trials under the experimental condition compared to those under the baseline and control condition within-subject.

The stop consonants /g/ and /k/ were chosen for the opposite reason. Their articulation requires a rapid opening of a full closure which is produced by the tongue root pressing against the velum. Neither of the articulators that are engaged in the experimental condition are used in the production of these plosives and as a result, they should not be affected by the manipulation. If those stimuli produced elsewhere show a similar boundary shift or wider boundary range, this might be an indicator that the experimental condition is more distracting than the control condition. The /g/-/k/ continuum should not show any changes, as both /g/ and /k/ are produced through a full closure at the velum and the root of the tongue. Hypotheses 3 and 4 are as follows:

> H3: For the stimulus continuum /g/-/k/, the phoneme boundary is not shifted for

trials under the experimental condition compared to those under the baseline and control condition within-subject.

H4: For the stimulus continuum /g/-/k/, the phoneme boundary range is not wider for trials under the experimental condition compared to those under the baseline and control condition within-subject.

9 Method

9.1 Participants

The target number of participants was set to 40 prior to the experiment, as this number had been used in other studies where similar effects were found, but was still considered feasible for an unpaid study with limited time. The minimum number of participants necessary for completing data collection was set to 30. Each participant would go through all three continua in all three conditions.

Participants were 33 adult German native speakers between 18 and 39 years (mean age = 25.9 years, sd = 3.54, 20 female, 13 male). Seven of the 33 participants had a second native language (1 Turkish, 2 Croatian, 2 Slovak, 1 Czech, 1 French), though all of them stated that they learned German from birth and therefore have acquired the typical German phoneme categories during the perceptual narrowing phase.

Participants were recruited through word of mouth and social media postings and had to be between 18 and 45 years old. Participants were excluded if they were over 45 (due to possible onset of age-related deterioration of hearing abilities), if German was not their native language (as the continua were designed to represent contrasts of stereotypical German phonemes), if they had hearing impairments of any sort or severe visual impairments, if they had motor impairments of the articulators, the hands or the feet, and if they had any neurological diseases that might influence their cognitive processes. Additionally, participants were excluded, if their data showed signs of poor perception of the continuum endpoints (maximum of one incorrect identification at the endpoints combined). They were not excluded based on the distribution of their button presses (e.g. imperfect boundaries)[1], as long as the endpoints were identified correctly. 11 participants had to be excluded based on the poor perception of the continuum

1 This practice had been criticized even in the early days of phoneme perception research (e.g. Lane, 1965, p. 292). If participants who do not show good boundary curves are excluded, it does not seem surprising if studies find almost perfect categorical perception as a result.

endpoints. All but one of the cases of poor endpoint perception were found for the /ba/-/da/ continuum (see section 11 for discussion).

As compensation, participants were allowed to choose from a selection of candy and enter a gift card lottery for a 50€ Amazon gift card. They were not paid and participated entirely voluntarily.

9.2 Materials

Equipment

The study was conducted in the Babelfish Lab at the University of Vienna. A PsychoPy script controlling the experiment flow was run on the stimulus presentation computer. Throughout the session, instructions in the form of text were presented on a monitor placed on a desk in front of the participant. The stimuli were played through Bose SoundLink II headphones connected via an Aux cable. Participants were given a button box shaped like a gamepad with two buttons which was connected to the stimulus presentation computer via a serial port. For the experimental condition, simple wooden tongue depressors were placed in the participants' mouths. For the control condition, a Bernina sewing machine foot pedal was placed on the floor, but the cable did not connect to another device.

Stimuli

The stimuli consisted of three 10-step continua going from one phoneme to another. The three continua consisted of the following phoneme endpoints: i) /i:/ to /y:/, ii) /b/ to /d/, and iii) /g/ to /k/. The stop consonants were presented as C-/a/ syllables. The stimuli were selected for their articulatory properties. As the previous research suggests that the effects of motor system involvement are articulator-specific, phonemes requiring the tongue tip or lips for articulation were chosen.

All stimuli were created using the open-source software Praat (Boersma & Weenink, 2022). The endpoints for each continuum were spoken by a female German native speaker and recorded using a Zoom H1 microphone. Based on these recordings, the 10-step continua were created using either a resynthesis approach (for /i:/-/y:/ and /ba/-/da/) or the cutback method (for /ga/-/ka/).

The /i:/-/y:/ continuum

For the /i:/-/y:/ continuum, a resynthesis process was used to generate the stimuli. The recordings of the female native speaker were "whitened" to strip the relevant portion of the sound of its phoneme characteristics using LPC and inverse filtering. This estimation of the glottal signal is then used as a vocal source on which an artificial vocal tract filter can be placed.

Based on the first, second, and third formant values of the endpoint recordings a series of ten equally spaced filters was created for each continuum. These filters were then placed on top of the source signal, creating the ten steps of the continuum. As /i:/ and /y:/ only differ in their formant frequencies due to the different lip positions, a modification of the first, second, and third formant frequencies is sufficient to create the continuum. All steps were conducted using a combination of self-written scripts and manual modifications in the Praat software.

The /b/-/d/ continuum

For the /b/-/d/ continuum, the main difference in the acoustic signal is the formant transition between the voice onset and the full quality of the following vowel. The average voice onset time differs slightly, but since there was no real difference in the recordings of the endpoints and also for the purpose of easier creation, voice onset time was kept stable across the continuum. Therefore, only the first, second, and third formant frequencies were modified. Due to multiple iterations of the creation of this continuum due to insufficient results, the script by Winn (2019) was used for this continuum. It uses the same technique as mentioned above for the /i:/-/y:/ continuum, but allows the automatization of most steps in the process.[2]

The /g/-/k/ Continuum

For the /g/-/k/ continuum, a different method was necessary. Several different acoustic cues distinguish voiced and voiceless stop consonants and therefore need to be manipulated. Most important for German are the voice onset time and aspiration. For this continuum, the script documented in Winn (2020) was used, which was designed specifically to create

2 For /i/-/y/ the formant values were manually extracted and inserted into a script, that would calculate the steps of the continuum and create the filters. The filters were then combined manually with the source, which had been acquired previously through LPC analysis and filtering. All of these steps required manual mouse presses in the GUI of Praat. The script by Winn (2019) goes through these steps by itself based on some input like the number of steps or the sound files acting as the continuum endpoints.

a continuum going from a voiced plosive to the corresponding voiceless plosive. This is achieved with the progressive cutback method, where from the voiceless plosive, the aspiration is cut back step by step and replaced with a time-aligned snippet from the voiced plosive. Through this, the aspiration gets progressively shorter, and gradually more of the formant transition is present. This method does not include resynthesis of the natural sounds.

Evaluation
The stimuli were evaluated by eight German native speakers, who were allowed to repeatedly listen through the stimuli in their own time and state whether they heard one of the two endpoint phonemes, both phonemes, or neither one. These judgments show whether the endpoints of the continua are clearly identifiable and if the steps are fine-grained enough to allow changes to the phoneme boundary to be observable. By giving the option to choose that both or neither of the phonemes were heard, continuum steps that would be perceived as ambiguous or as neither of the endpoint phonemes could be detected. The following criteria were chosen for a continuum to be sufficient:

- At the endpoints of the continuum, at least two steps should be clearly identifiable as the endpoint phoneme.
- At least four adjacent steps including steps 5 and 6 should not be universally rated as one of the endpoint phonemes.

The ratings for all three continua show clear identifiability at the endpoints as well as some ambiguity around steps 5 and 6. This means that they should be sufficient to cover the phoneme boundary variation between subjects but also leave some room for making identification a little challenging.

9.3 Design

This experiment used a 3x3 factorial design with the condition and the continuum category as predictor variables compared within-subject. As a measure, button presses and response times were recorded for each trial (6 repetitions per stimulus), which were then used to estimate the phoneme boundary midpoint and the boundary range for each participant, continuum, and condition. These two parameters were then compared in a two-way repeated measures ANOVA as dependent variables .

Conditions

Phoneme identification was tested under three conditions: A baseline condition, an experimental condition, and one control condition. For the baseline condition, participants were asked to sit comfortably and do nothing besides pressing the buttons on the button box to indicate their choice of category. This serves as a baseline to acquire their phoneme boundary in a non-distracted listening condition.

The experimental condition aims to engage the lips and tongue tip as well as the relevant motor areas of the brain. To achieve this, participants were asked to hold a wooden tongue depressor in their mouth, so that their lips would be in a rounded position. Additionally, they were instructed to use the tip of the tongue to gently press one end of the tongue depressor against their hard palate. Participants were asked to not use their teeth to hold the tongue depressor in place, as it would lead to less muscle engagement on the lips. The tongue depressor was held in place during the phoneme identification task. The correct position of the tongue depressor could be checked via webcam by the position of the outside part of the tongue depressor (the outer end would move upwards if participants did not press the inner end against their hard palate).

As a control condition, a motor task involving the foot was chosen, as it would also engage the motor system but not those areas relevant to articulatory processes. Unlike the hand representation, the foot representation in the motor cortex is not adjacent to the representation of the articulators. To keep the conditions comparable, the control task needed to be a simple static motor task that could be performed during trials and shows equal levels of muscle strain and discomfort as the experimental condition. This simple task was continuously pressing a pedal with their right foot (similar to the foot position when driving a car). Participants were instructed to keep the heel off the pedal and only press it with the ball and arch of their foot. Minimal adjustments to this position were made based on the participants' muscle strength to avoid mid-trial cramping. Participants were also asked to take off their shoes during the experiment, to keep the position of the foot more comparable. The pedal had to be fully pressed down during the trials and the correct position was checked visually during the experiment.

9.4 Procedure

Ahead of the experiment, participants were sent an information sheet, containing all relevant information about the cause, procedure, and potential risks of the study. While

the exact research design and hypotheses were kept from the participants, they were informed about the broad topic of the study ("testing phoneme perception under various distractions"). Once in the lab, they were shown the setup and given detailed instructions on the task. They would then fill out a short questionnaire about possible exclusion criteria and sign the information sheet to give consent to their participation and use of their data. After that, the participants were given the opportunity to test out and familiarize themselves with the wooden tongue depressor as well as the foot pedal.

The desired positions of the tongue depressor in the mouth and the foot on the pedal were practiced under the guidance of the examiner. Then, participants received a pair of headphones, and ten practice trials were used to familiarize the participants with the button box and the general procedure. The practice trials consisted of one repetition of a 10-step continuum going from /a:/ to /æ:/ presented in random order. Participants were instructed to categorize the sounds to the best of their judgment and to prioritize accuracy over speed, but not take too much time. Each stimulus would only be presented once per trial. The next stimulus would only be presented, once they had made a decision by pressing the corresponding button. When all questions of the participant were answered, they were left alone in the measuring room and all further instructions were given as text on a computer monitor controlled by the PsychoPy script.

Through webcams, the activity of the participant was documented and checked during the experiment. Each participant would go through all three continua and all three conditions in a series of nine blocks. First, they went through three blocks in one condition, then three blocks in the second condition, and finally three blocks in the third condition. The order of conditions and blocks was set in advance in a list and tied to the participant ID. Combinations of condition and continuum order were kept balanced across participants but followed the pattern shown in figure 9.1.

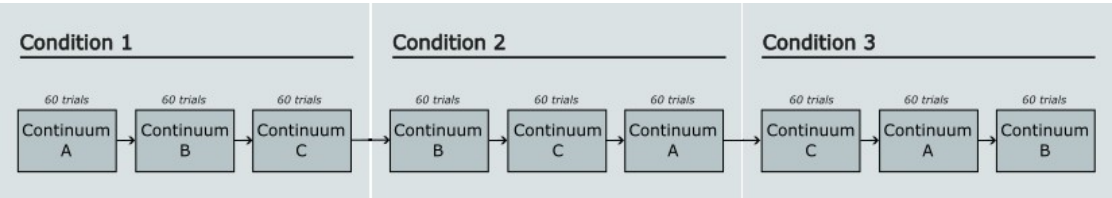


Figure 9.1: Experiment flow.

Each block contained six repetitions of one 10-step continuum (60 trials). At the beginning of the block, the continuum type was stated (e.g. "You will now hear different versions of 'ba' or 'da'. Please decide, which of the two sounds you hear by pressing the corresponding

button.") and the two contrasting phonemes were presented in the lower corner on each side corresponding to the button position (e.g. 'b' in the left corner, 'd' in the right corner). During each trial, a fixation cross was shown as well as the phonemes in each lower corner. Participants were instructed to look at the fixation cross while the auditory stimulus was played and avoid looking at the lower corners unless needed for pressing the correct button, to avoid a possible influence of the orthographic form (see also figure 9.2).

Each block took around two and a half minutes. Between each block, participants could take a break for as long as they needed. They were allowed to take the tongue depressor out of their mouth and take the foot off the pedal during those breaks. They would continue with the next block by pressing any of the two buttons. After the ninth block, the examiner entered the room again and ended the experiment. The experiment took 20-30 minutes, depending on the length of the breaks. The participant was given some candy as a gift.

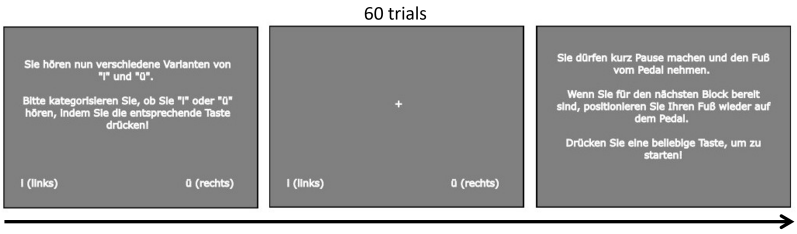


Figure 9.2: Flow for each block

.

9.5 Data Analysis

The data were analyzed by fitting a logistic curve to the phoneme ratings of each participant for each continuum and condition. From these curves, the midpoints and scale parameters were extracted and compared in a repeated measures ANOVA. This approach was used following Möttönen and Watkins (2009).

Logistic regression was chosen to model the phoneme boundaries of each participant for each continuum and condition. The logistic function is defined as $p(x) = \frac{L}{1-e^{-k(x-x_0)}}$, where L is the limit of the function, x_0 is the midpoint of the function where $p(x) = 0.5$, and k is the growth rate at the midpoint. By transforming the function into log odds, the function becomes linear and can be estimated through linear regression. The linear

function can then be written as $f(x) = log(\frac{p}{1-p}) = \beta_0 + \beta_1 x$, which can be transformed back into the logistic function of the form $p(x) = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}}$. From this, the midpoint can be extracted through $m = -\frac{\beta_0}{\beta_1}$ and the growth rate as $k = \beta_1$. Because these values for the growth rate are far from normally distributed, a scale parameter s, defined as $s = \frac{1}{\beta_1}$, can be used as an indicator of the range of the phoneme boundary.

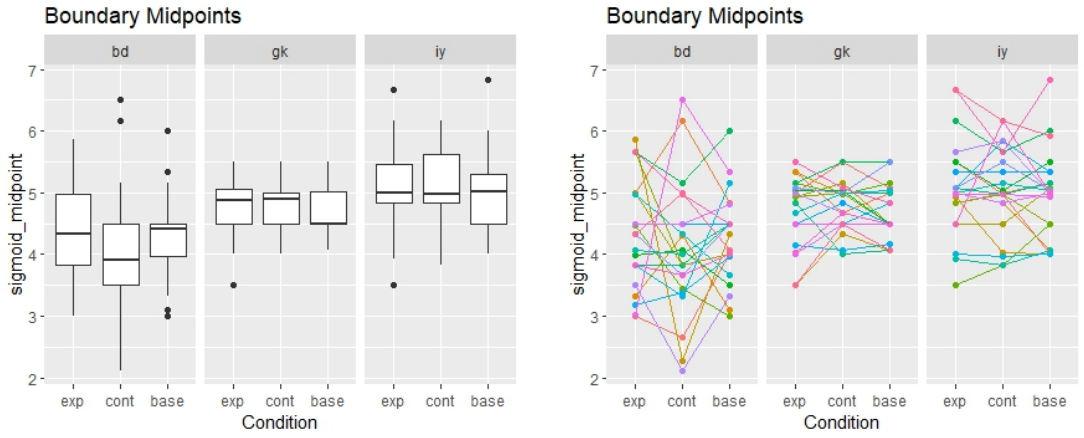
For each participant, a midpoint and scale parameter was determined for each continuum in each condition. Participants with poor perception of the continuum endpoints were removed from the data set.

These parameters were then compared using a two-way repeated measures ANOVA. Two separate ANOVAs were performed, one for the midpoints and one for the scale parameters, with the continuum type and the condition as independent variables. All analyses were conducted using R.

10 Results

10.1 Boundary Midpoints

The two-way repeated measures ANOVA revealed no statistically significant interaction of condition and continuum category on the boundary midpoints (F=0.563, df=4, p=0.690). The analysis of the main effects showed that the condition did not have an effect on the boundary midpoints (F=0.196, df=2, p=822). Only the continuum category showed a statistically significant main effect on the boundary midpoints (F=25.432, df=2, p=0.000). The distribution of boundary midpoints across conditions and continuum categories is shown in figure 10.1a and the shift of boundary midpoints between conditions is shown in figure 10.1b.

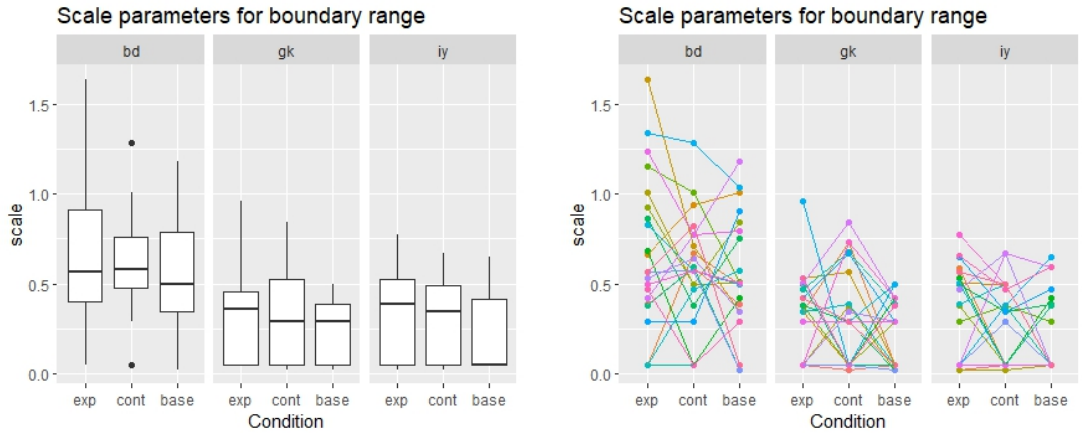


(a) **(b)**

Figure 10.1: a) Distribution of boundary midpoints for each continuum and condition. b) Individual midpoints and shifts for each continuum and condition. The color represents the individual participant.

10.2 Boundary Range

The two-way repeated measures ANOVA for the boundary scale parameters revealed, as with the boundary midpoints, no interactions of condition and continuum category (F=0.101, df=4, p=0.982). No main effect of the condition was found (F=2.404, df=2, p=0.093). Again, only the continuum category had a significant main effect on the boundary range (F=28.785, df=2, p<0.001). The distribution of range parameters across conditions and continuum categories is shown in figure 10.2a and the change of range parameter between conditions are shown in figure 10.2b.



(a) **(b)**

Figure 10.2: a) Distribution of scale parameters for boundary range (defined as $1/\beta_1$) for each continuum and condition. b) Individual scale parameters for boundary range and shifts for each continuum and condition. The color represents the individual participant.

10.3 Response Times

The response times were not specifically part of the hypotheses but were recorded as an additional measure. They were also compared in a two-way repeated measures ANOVA, which revealed no interaction of condition and continuum category (F=1.403, df=4, p=0.230). The main effects analysis, however, showed a statistically significant effect of condition on the response times (F=48.605, df=2, p=0.000, η^2=0.025), while the continuum category did not have a significant main effect (F=2.775, df=2, p=0.062).

Post-hoc analysis (pairwise t-test with Bonferroni adjustment) revealed that the re-

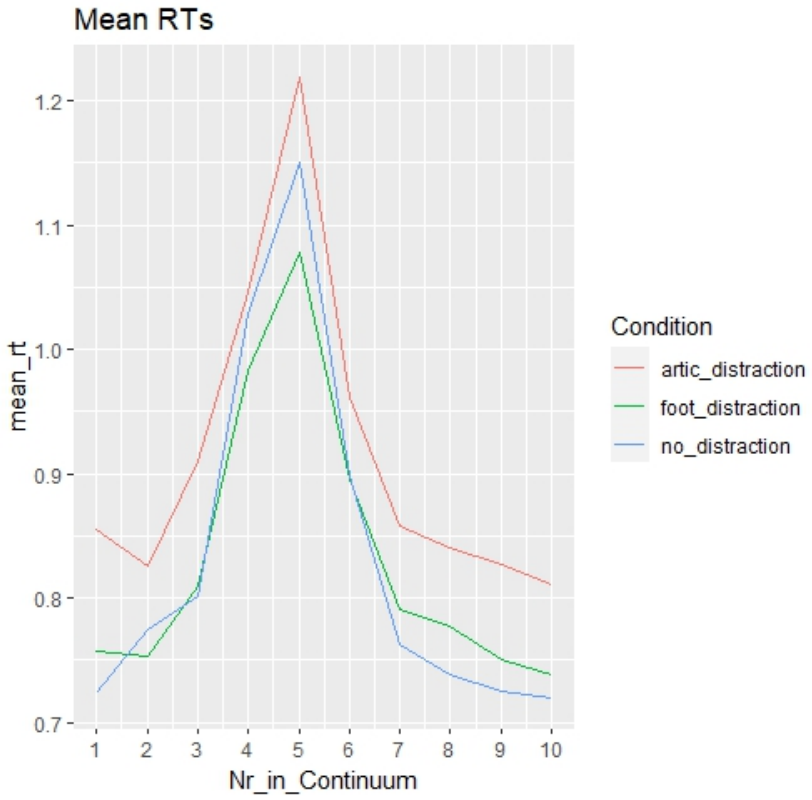


Figure 10.3: Mean response time in seconds for every step of the continuum (averaged over all three continua) for each condition.

sponse times under the experimental condition were significantly different from those in the baseline ($p<0.001$, 95% C.I.=[0.061, 0.105]) and control condition ($p<0.001$,95% C.I.=[0.062, 0.102]). The mean response time for each step of the continuum is shown for each condition in figure 10.3.

Part III

Discussion

11 Discussion

This study examined the influence of articulator-specific stimulation on phoneme perception by comparing effects on the phoneme boundary and response times during articulator stimulation and a control motor task. The results show no effect on the phoneme boundaries, neither the midpoint nor the range. Response times, however, were affected by articulator stimulation and showed an average decrease of 0.083 seconds with no correlation to the steps of the continua. While these results may be a product of the limitations of the study, some interesting conclusions can still be drawn.

11.1 The Phoneme Boundaries

Both the ANOVA on the midpoints and the ANOVA on the scale parameter revealed no significant effect of the experimental condition compared to the baseline and control conditions. This is not entirely unexpected. Effects from previous studies are small and often very specific to a task or listening condition. The data also show shifts in the midpoints and ranges but not in a uniform pattern. There are different possible reasons for that. Firstly, the experimental condition may not have been optimal. A wooden stick is not equivalent to rTMS and may therefore not cause enough activity in the motor system. It may also be, that it does not cause activity in exactly the right motor areas. Due to limited resources, an assessment of the exact activation pattern caused by the tongue depressor was not possible. It was assumed that at least the primary motor and sensory areas corresponding to the lips and the tongue tip should have been activated. To what degree a motor task like the positioning of the tongue depressor would activate premotor or supplementary motor areas is hard to tell. It may therefore be possible that the PM or SMA may be involved in phoneme perception but were simply not stimulated by the experimental condition.

The second possibility is that even if the motor system played a supplementary role in phoneme perception, the adult native speaker would have formed categories robust enough

to withstand minor deviations from the multimodal representations. Meaning, even if the motor system had a supplementary role and would in principle pass information for phoneme perception, the signal coming from the auditory system may simply be of higher weighting. This would also fit with the account of the motor system getting recruited in noisy listening conditions, where information from the motor system would only affect perception if the auditory signal is unreliable. The slower response times may also be a reflection of this, as incoming misinformation from the motor system would likely slow down the perception process.

Lastly, one must always consider that the motor system may not be involved at all, but only gets activated due to the tight connections between language and motor areas, but not actually providing input to the perception process. Though a series of studies have found indicators of motor system recruitment, this may very well be a result of the artificial contexts of a laboratory or due to response selection processes.

11.2 The Influence of Articulator Manipulation on Response Times

While the experimental condition did not show an effect on the phoneme boundaries, it did have a significant effect on response times. Response times were slowed down on average by 0.083 seconds during the experimental condition. This effect is present across the entire continuum and not just around the boundary. It is also not articulator specific as the /g/-/k/ continuum is equally affected. These can be taken as indicators that the wooden tongue depressor was most likely a greater distraction than the foot pedal. Considering many people drive a car and use their feet to press the pedal it seems reasonable that despite the effort necessary to hold the position and the slight differences in foot position, people would be more accustomed to pressing a pedal than having a wooden stick in their mouth.

It could be argued, though, that perhaps the effect may not be articulator specific because the tongue depressor makes it impossible to produce /ga/ or /ka/. In a sense, this was also argued as the reason for infants' lack of discrimination abilities in the studies by Bruderer et al. (2015) and Choi et al. (2019). To test this account, a non-speech-related task is needed. If the tongue depressor is simply more distracting, then the slowdown of responses should also affect non-linguistic tasks. If, on the other hand, the articulator

stimulation directly affects the motor areas relevant to speech perception, then response times on a non-speech task should not be delayed.

11.3 Limitations of the Study

In this study, the wooden tongue depressor may not have been sufficient for stimulation, whether it did not sufficiently stimulate the motor system in general or simply stimulated areas not relevant for speech perception. The resulting cortical activation patterns of the experimental and control condition would need to be examined, to bring clarity.

The continua may not have been ideal. Especially the /ba/-/da/ continuum showed more variation of the midpoints and higher range values than the other two continua. Eleven out of the 33 participants had to be excluded based on their /b/-/d/ performance. This raises the question if the /b/-/d/ continuum was poorly designed. All necessary steps for quality control were taken in advance. Multiple recreations had been made, to achieve the best possible version. The stimuli were also evaluated in advance and performed reasonably well. If the continuum itself had been designed so poorly that identification was too difficult, it seems strange that two-thirds of the participants did not seem to struggle too much based on their data. Though participants most commonly mentioned /b/-/d/ as the most difficult continuum, their data show a dichotomy of the identification curves, with either reasonable phoneme boundaries or no boundary at all with similar decision rates across all steps of the continuum. At this point, no pattern in the order of conditions or presentation order of the continua could be identified. Interestingly, six out of the seven participants with a second native language performed well on the /b/-/d/ continuum, the one person who didn't, being a French native speaker, while the majority of second native languages were Slavic languages. It seems likely that individual properties of the participants may lead to good or bad identification of this continuum. Such properties may be a second language or dialectal background or general auditory perception skills (e.g. musical training).

To compensate for the poor identifiability of the /b/-/d/ continuum, the analysis process was repeated without the /b/-/d/ data and all 33 participants. This analysis yielded the same results as the one reported above. The quality of the /b/-/d/ continuum and the exclusion of participants did not seem to affect the results.

12 Conclusion

The debate about the motor system's involvement in speech perception processes is ongoing. The studies conducted so far have not resulted in a clear picture of how and to what degree the motor system may be involved in speech perception.

The data collected in this study does not support motor system involvement, though the mechanical stimulation of the articulators may not have been sufficient and more intense stimulation like rTMS might be needed to induce an observable effect on perception. Overall, the auditory representations of phonemes in the adult brain seem rather robust, so any involvement of the motor system would only take a minor role. The activation of motor areas in response to speech can be explained through the tight neural connections between auditory and motor areas formed through associative learning. But such a connection does not require the motor system to return input to the perceptual process. Yet, the findings of the many studies listed above need an explanation. Some mechanism has to exist in the brain, that allows effects like those to arise. Whether this mechanism is an elaborate interaction between motor and auditory areas or simply a byproduct of a self-monitoring loop, a unified account is needed. Further studies may want to prioritize response times over identification or discrimination scores as a measure. Also, the incorporation of non-linguistic tasks can give further insight into the domain-specificity of any possible effects.

It is clear that the motor theory of speech perception is false and that the motor system is not necessary for speech perception. But how the motor system may be involved in speech perception is still not well understood. Viewing speech perception as a purely auditory process seems like an oversimplification in light of the findings mentioned above. Speech is in its essence multimodal and its neural representation should reflect that. It is clear that at this point in time, there is still a lot left to explore.

Bibliography

Adank, P., Davis, M. H., & Hagoort, P. (2012). Neural dissociation in processing noise and accent in spoken language comprehension. *Neuropsychologia*, *50*(1), 77–84. https://doi.org/10.1016/j.neuropsychologia.2011.10.024

Adank, P., Nuttall, H. E., & Kennedy-Higgins, D. (2017). Transcranial magnetic stimulation and motor evoked potentials in speech perception research. *Language, Cognition and Neuroscience*, *32*(7), 900–909. https://doi.org/10.1080/23273798.2016.1257816

Aldridge, M. A., Stillman, R. D., & Bower, T. G. R. (2001). Newborn categorization of vowel-like sounds. *Developmental Science*, *4*(2), 220–232. https://doi.org/10.1111/1467-7687.00167

Angeli, A., Davidoff, J., & Valentine, T. (2008). Face familiarity, distinctiveness, and categorical perception. *Quarterly Journal of Experimental Psychology*, *61*(5), 690–707. https://doi.org/10.1080/17470210701399305

Arsenault, J. S., & Buchsbaum, B. R. (2015). Distributed neural representations of phonological features during speech perception. *The Journal of Neuroscience*, *35*(2), 634–642. https://doi.org/10.1523/JNEUROSCI.2454-14.2015

Arsenault, J. S., & Buchsbaum, B. R. (2016). No evidence of somatotopic place of articulation feature mapping in motor cortex during passive speech perception. *Psychonomic Bulletin & Review*, *23*(4), 1231–1240. https://doi.org/10.3758/s13423-015-0988-z

Baldissera, F., Cavallari, P., Craighero, L., & Fadiga, L. (2001). Modulation of spinal excitability during observation of hand actions in humans: Spinal excitability during observation of hand actions. *European Journal of Neuroscience*, *13*(1), 190–194. https://doi.org/10.1046/j.0953-816x.2000.01368.x

Benson, R. R., Whalen, D., Richardson, M., Swainson, B., Clark, V. P., Lai, S., & Liberman, A. M. (2001). Parametrically dissociating speech and nonspeech perception in the brain using fMRI. *Brain and Language*, *78*(3), 364–396. https://doi.org/10.1006/brln.2001.2484

Berent, I., & Platt, M. (2022). Is phonology embodied? evidence from mechanical stimulation. *Journal of Psycholinguistic Research*, *51*, 597–626. https://doi.org/10.1007/s10936-022-09871-x

Berent, I., Platt, M., Theodore, R., Balaban, E., Fried, P. J., & Pascual-Leone, A. (2020). Speech perception triggers articulatory action: Evidence from mechanical stimulation. *Frontiers in Communication*, *5*, 34. https://doi.org/10.3389/fcomm.2020.00034

Binkofski, F., & Buccino, G. (2006). The role of ventral premotor cortex in action execution and action understanding. *Journal of Physiology-Paris*, *99*(4), 396–405. https://doi.org/10.1016/j.jphysparis.2006.03.005

Boersma, P., & Weenink, D. (2022, December 9). *Praat: Doing phonetics by computer* (Version 6.2.19). https://www.praat.org

Bohland, J. W., & Guenther, F. H. (2006). An fMRI investigation of syllable sequence production. *NeuroImage*, *32*(2), 821–841. https://doi.org/10.1016/j.neuroimage.2006.04.173

Brendel, B., Hertrich, I., Erb, M., Lindner, A., Riecker, A., Grodd, W., & Ackermann, H. (2010). The contribution of mesiofrontal cortex to the preparation and execution of repetitive syllable productions: An fMRI study. *NeuroImage*, *50*(3), 1219–1230. https://doi.org/10.1016/j.neuroimage.2010.01.039

Bristow, D., Dehaene-Lambertz, G., Mattout, J., Soares, C., Gliga, T., Baillet, S., & Mangin, J.-F. (2008). Hearing faces: How the infant brain matches the face it sees with the speech it hears. *Journal of Cognitive Neuroscience*, *21*(5), 905–921. https://doi.org/10.1162/jocn.2009.21076

Bruderer, A. G., Danielson, D. K., Kandhadai, P., & Werker, J. F. (2015). Sensorimotor influences on speech perception in infancy. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(44), 13531–13536. https://doi.org/10.1073/pnas.1508631112

Buccino, G., Binkofski, F., Fink, G. R., Fadiga, L., Fogassi, L., Gallese, V., Seitz, R. J., Zilles, K., Rizzolatti, G., & Freund, H.-J. (2001). Action observation activates premotor and parietal areas in a somatotopic manner: An fMRI study. *European Journal of Neuroscience*, *13*(2), 400–404. https://doi.org/10.1046/j.1460-9568.2001.01385.x

Burnham, D., & Dodd, B. (2004). Auditory-visual speech integration by prelinguistic infants: Perception of an emergent consonant in the McGurk effect. *Developmental Psychobiology*, *45*(4), 204–220. https://doi.org/10.1002/dev.20032

Callan, D., Callan, A., Gamez, M., Sato, M.-a., & Kawato, M. (2010). Premotor cortex mediates perceptual performance. *NeuroImage*, *51*(2), 844–858. https://doi.org/10.1016/j.neuroimage.2010.02.027

Choi, D., Bruderer, A. G., & Werker, J. F. (2019). Sensorimotor influences on speech perception in pre-babbling infants: Replication and extension of bruderer et al. (2015). *Psychonomic Bulletin & Review*, *26*(4), 1388–1399. https://doi.org/10.3758/s13423-019-01601-0

Chomsky, N. (1968). *Language and mind*. Harcourt, Brace & World.

Cook, R., Bird, G., Catmur, C., Press, C., & Heyes, C. (2014). Mirror neurons: From origin to function. *Behavioral and Brain Sciences*, *37*(2), 177–192. https://doi.org/10.1017/S0140525X13000903

Cooper, F. S., Gaitenby, J. H., & Nye, P. W. (1984). Evolution of reading machines for the blind: Haskins laboratories' research as a case history. *Journal of Rehabilitation Research and Development*, *21*(1), 51–87.

Correia, J. M., Jansma, B. M., & Bonte, M. (2015). Decoding articulatory features from fMRI responses in dorsal speech regions. *The Journal of Neuroscience*, *35*(45), 15015–15025. https://doi.org/10.1523/JNEUROSCI.0977-15.2015

Cross, D. V., Lane, H. L., & Sheppard, W. C. (1965). Identification and discrimination functions for a visual continuum and their relation to the motor theory of speech perception. *Journal of Experimental Psychology*, *70*(1), 63–74. https://doi.org/10.1037/h0021984

Damasio, A. R. (1992). Aphasia. *New England Journal of Medicine*, *326*(8), 531–539. https://doi.org/10.1056/NEJM199202203260806

D'Ausilio, A., Bufalari, I., Salmas, P., & Fadiga, L. (2012). The role of the motor system in discriminating normal and degraded speech sounds. *Cortex*, *48*(7), 882–887. https://doi.org/10.1016/j.cortex.2011.05.017

D'Ausilio, A., Jarmolowska, J., Busan, P., Bufalari, I., & Craighero, L. (2011). Tongue corticospinal modulation during attended verbal stimuli: Priming and coarticulation effects. *Neuropsychologia*, *49*(13), 3670–3676. https://doi.org/10.1016/j.neuropsychologia.2011.09.022

D'Ausilio, A., Pulvermüller, F., Salmas, P., Bufalari, I., Begliomini, C., & Fadiga, L. (2009). The motor somatotopy of speech perception. *Current Biology*, *19*(5), 381–385. https://doi.org/10.1016/j.cub.2009.01.017

Delvaux, V., & Soquet, A. (2007). The influence of ambient speech on adult speech productions through unintentional imitation. *Phonetica*, *64*(2), 145–173. https://doi.org/10.1159/000107914

Dent, M. L., Brittan-Powell, E. F., Dooling, R. J., & Pierce, A. (1997). Perception of synthetic /ba/–/wa/ speech continuum by budgerigars (*Melopsittacus undulatus*). *The Journal of the Acoustical Society of America*, *102*(3), 1891–1897. https://doi.org/10.1121/1.420111

Dooling, R. J., Okanoya, K., & Brown, S. D. (1989). Speech perception by budgerigars (melopsittacus undulatus): The voiced-voiceless distinction. *Perception & Psychophysics*, *46*(1), 65–71. https://doi.org/10.3758/BF03208075

Du, Y., Buchsbaum, B. R., Grady, C. L., & Alain, C. (2014). Noise differentially impacts phoneme representations in the auditory and speech motor systems. *Proceedings of the National Academy of Sciences*, *111*(19), 7126–7131. https://doi.org/10.1073/pnas.1318738111

Eimas, P. D., Siqueland, E. R., Jusczyk, P., & Vigorito, J. (1971). Speech perception in infants. *Science*, *171*(3968), 303–306. https://doi.org/10.1126/science.171.3968.303

Etcoff, N. L., & Magee, J. J. (1992). Categorical perception of facial expressions. *Cognition*, *44*(3), 227–240. https://doi.org/10.1016/0010-0277(92)90002-Y

Evans, S., & Davis, M. H. (2015). Hierarchical organization of auditory and motor representations in speech perception: Evidence from searchlight similarity analysis. *Cerebral Cortex*, *25*(12), 4772–4788. https://doi.org/10.1093/cercor/bhv136

Fadiga, L., Craighero, L., Buccino, G., & Rizzolatti, G. (2002). Speech listening specifically modulates the excitability of tongue muscles: A TMS study. *European Journal of Neuroscience*, *15*(2), 399–402. https://doi.org/10.1046/j.0953-816x.2001.01874.x

Fitch, H. L., Halwes, T., Erickson, D. M., & Liberman, A. M. (1980). Perceptual equivalence of two acoustic cues for stop-consonant manner. *Perception & Psychophysics*, *27*(4), 343–350. https://doi.org/10.3758/BF03206123

Fodor, J. A. (1983). *The modularity of mind: An essay on faculty psychology*. MIT Press.

Fornia, L., Puglisi, G., Leonetti, A., Bello, L., Berti, A., Cerri, G., & Garbarini, F. (2020). Direct electrical stimulation of the premotor cortex shuts down awareness of voluntary actions. *Nature Communications*, *11*(1), 705. https://doi.org/10.1038/s41467-020-14517-4

Fox, R. J., Kasner, S. E., Chatterjee, A., & Chalela, J. A. (2001). Aphemia: An isolated disorder of articulation. *Clinical Neurology and Neurosurgery*, *103*(2), 123–126. https://doi.org/10.1016/S0303-8467(01)00126-3

Fried, I., Katz, A., McCarthy, G., Sass, K., Williamson, P., Spencer, S., & Spencer, D. (1991). Functional organization of human supplementary motor cortex studied

by electrical stimulation. *The Journal of Neuroscience*, *11*(11), 3656–3666. https://doi.org/10.1523/JNEUROSCI.11-11-03656.1991

Galantucci, B., Fowler, C. A., & Turvey, M. T. (2006). The motor theory of speech perception reviewed. *Psychonomic Bulletin & Review*, *13*(3), 361–377. https://doi.org/10.3758/BF03193857

Geyer, S., Luppino, G., & Rozzi, S. (2012). Motor cortex. In J. K. Mai & G. Paxinos (Eds.), *The human nervous system* (pp. 1012–1035). Elsevier. https://doi.org/10.1016/B978-0-12-374236-0.10027-6

Goldstone, R. L., Lippa, Y., & Shiffrin, R. M. (2001). Altering object representations through category learning. *Cognition*, *78*(1), 27–43. https://doi.org/10.1016/S0010-0277(00)00099-8

Grabski, K., & Sato, M. (2020). Adaptive phonemic coding in the listening and speaking brain. *Neuropsychologia*, *136*, 107267. https://doi.org/10.1016/j.neuropsychologia.2019.107267

Grabski, K., Schwartz, J.-L., Lamalle, L., Vilain, C., Vallée, N., Baciu, M., Le Bas, J.-F., & Sato, M. (2013). Shared and distinct neural correlates of vowel perception and production. *Journal of Neurolinguistics*, *26*(3), 384–408. https://doi.org/10.1016/j.jneuroling.2012.11.003

Hervais-Adelman, A. G., Carlyon, R. P., Johnsrude, I. S., & Davis, M. H. (2012). Brain regions recruited for the effortful comprehension of noise-vocoded words. *Language and Cognitive Processes*, *27*(7), 1145–1166. https://doi.org/10.1080/01690965.2012.662280

Hickok, G. (2013). Do mirror neurons subserve action understanding? *Neuroscience Letters*, *540*, 56–58. https://doi.org/10.1016/j.neulet.2012.11.001

Hickok, G., Houde, J., & Rong, F. (2011). Sensorimotor integration in speech processing: Computational basis and neural organization. *Neuron*, *69*(3), 407–422. https://doi.org/10.1016/j.neuron.2011.01.019

Hill, V., Cankurtaran, C., Liu, B., Hijaz, T., Naidich, M., Nemeth, A., Gastala, J., Krumpelman, C., McComb, E., & Korutz, A. (2019). A practical review of functional MRI anatomy of the language and motor systems. *American Journal of Neuroradiology*, *40*(7), 1084–1090. https://doi.org/10.3174/ajnr.A6089

Hillis, A. E. (2007). Aphasia: Progress in the last quarter of a century. *Neurology*, *69*(2), 200–213. https://doi.org/10.1212/01.wnl.0000265600.69385.6f

Iacoboni, M., Molnar-Szakacs, I., Gallese, V., Buccino, G., Mazziotta, J. C., & Rizzolatti, G. (2005). Grasping the intentions of others with one's own mirror neuron system. *PLoS Biology*, *3*(3), e79. https://doi.org/10.1371/journal.pbio.0030079

Indefrey, P., & Levelt, W. J. M. (2000). The neural correlates of language production. In M. S. Gazzaniga (Ed.), *The new cognitive neurosciences* (pp. 845–865). MIT Press.

Johnson, K., Sherman, V. C., & Sherman, S. G. (2011). *Acoustic and auditory phonetics*. John Wiley & Sons, Incorporated.

Kluender, K. R., & Lotto, A. J. (1994). Effects of first formant onset frequency on [-voice] judgments result from auditory processes not specific to humans. *The Journal of the Acoustical Society of America*, *95*(2), 1044–1052. https://doi.org/10.1121/1.408466

Kopp, J., & Udin, H. (1969). Identification and discrimination functions for pure tone auditory frequencies. *Psychonomic Science*, *16*(2), 95–96. https://doi.org/10.3758/BF03336638

Krieger-Redwood, K., Gaskell, M. G., Lindsay, S., & Jefferies, E. (2013). The selective role of premotor cortex in speech perception: A contribution to phoneme judgements but not speech comprehension. *Journal of Cognitive Neuroscience*, *25*(12), 2179–2188. https://doi.org/10.1162/jocn_a_00463

Kronrod, Y., Coppess, E., & Feldman, N. H. (2016). A unified account of categorical effects in phonetic perception. *Psychonomic Bulletin & Review*, *23*(6), 1681–1712. https://doi.org/10.3758/s13423-016-1049-y

Kuhl, P. K. (1981). Discrimination of speech by nonhuman animals: Basic auditory sensitivities conducive to the perception of speech-sound categories. *The Journal of the Acoustical Society of America*, *70*(2), 340–349. https://doi.org/10.1121/1.386782

Kuhl, P. K., & Meltzoff, A. N. (1982). The bimodal perception of speech in infancy. *Science*, *218*(4577), 1138–1141. https://doi.org/10.1126/science.7146899

Kuhl, P. K., & Meltzoff, A. N. (1984). The intermodal representation of speech in infants. *Infant Behavior and Development*, *7*(3), 361–381. https://doi.org/10.1016/S0163-6383(84)80050-8

Kuhl, P. K., & Miller, J. D. (1975). Speech perception by the chinchilla: Voiced-voiceless distinction in alveolar plosive consonants. *Science*, *190*(4209), 69–72. https://doi.org/10.1126/science.1166301

Kuhl, P. K., & Padden, D. M. (1983). Enhanced discriminability at the phonetic boundaries for the place feature in macaques. *The Journal of the Acoustical Society of America*, *73*(3), 1003–1010. https://doi.org/10.1121/1.389148

Lane, H. (1965). The motor theory of speech perception: A critical review. *Psychological Review*, *72*(4), 275–309. https://doi.org/10.1037/h0021986

Lee, Y.-S., Turkeltaub, P., Granger, R., & Raizada, R. D. S. (2012). Categorical speech processing in broca's area: An fMRI study using multivariate pattern-based analysis. *The Journal of Neuroscience*, *32*(11), 3942–3948. https://doi.org/10.1523/JNEUROSCI.3814-11.2012

Levelt, W. J. M. (1999). Producing spoken language: A blueprint of the speaker. In C. M. Brown & P. Hagoort (Eds.), *The neurocognition of language* (Repr, pp. 83–122). Oxford University Press.

Lewkowicz, D. J., & Flom, R. (2014). The audiovisual temporal binding window narrows in early childhood. *Child Development*, *85*(2), 685–694. https://doi.org/10.1111/cdev.12142

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, *74*(6), 431–461. https://doi.org/10.1037/h0020279

Liberman, A. M., Delattre, P. C., & Cooper, F. S. (1958). Some cues for the distinction between voiced and voiceless stops in initial position. *Language and Speech*, *1*(3), 153–167. https://doi.org/10.1177/002383095800100301

Liberman, A. M. (1957). Some results of research on speech perception. *The Journal of the Acoustical Society of America*, *29*(1), 117–123. https://doi.org/10.1121/1.1908635

Liberman, A. M., Delattre, P., & Cooper, F. S. (1952). The role of selected stimulus-variables in the perception of the unvoiced stop consonants. *The American Journal of Psychology*, *65*(4), 497. https://doi.org/10.2307/1418032

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition*, *21*(1), 1–36. https://doi.org/10.1016/0010-0277(85)90021-6

Longcamp, M., Hupé, J.-M., Ruiz, M., Vayssière, N., & Sato, M. (2019). Shared premotor activity in spoken and written communication. *Brain and Language*, *199*, 104694. https://doi.org/10.1016/j.bandl.2019.104694

Lotto, A. J., Hickok, G., & Holt, L. L. (2009). Reflections on mirror neurons and speech perception. *Trends in Cognitive Sciences*, *13*(3), 110–114. https://doi.org/10.1016/j.tics.2008.11.008

MacNeilage, P. F., Rootes, T. P., & Chase, R. A. (1967). Speech production and perception in a patient with severe impairment of somesthetic perception and motor control. *Journal of Speech and Hearing Research*, *10*(3), 449–467. https://doi.org/10.1044/jshr.1003.449

Mann, V., & Liberman, A. M. (1983). Some differences between phonetic and auditory modes of perception. *Cognition*, *14*(2), 211–235. https://doi.org/10.1016/0010-0277(83)90030-6

Mcgurk, H., & Macdonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*(5588), 746–748. https://doi.org/10.1038/264746a0

Meister, I. G., Ingo G. Meister, Wilson, S. M., Stephen M. Wilson, Deblieck, C., Wu, A. D., & Iacoboni, M. (2007). The essential role of premotor cortex in speech perception. *Current Biology*, *17*(19), 1692–1696. https://doi.org/10.1016/j.cub.2007.08.064

Meltzoff, A., & Moore, M. (1977). Imitation of facial and manual gestures by human neonates. *Science*, *198*(4312), 74–78. https://doi.org/10.1126/science.897687

Menenti, L., Gierhan, S. M. E., Segaert, K., & Hagoort, P. (2011). Shared language: Overlap and segregation of the neuronal infrastructure for speaking and listening revealed by functional MRI. *Psychological Science*, *22*(9), 1173–1182. https://doi.org/10.1177/0956797611418347

Mochida, T., Kimura, T., Hiroya, S., Kitagawa, N., Gomi, H., & Kondo, T. (2013). Speech misperception: Speaking and seeing interfere differently with hearing. *PLOS ONE*, *8*(7). https://doi.org/10.1371/journal.pone.0068619

Moineau, S., Dronkers, N. F., & Bates, E. (2005). Exploring the processing continuum of single-word comprehension in aphasia. *Journal of Speech, Language, and Hearing Research*, *48*(4), 884–896. https://doi.org/10.1044/1092-4388(2005/061)

Morse, P. A., & Snowdon, C. T. (1975). An investigation of categorical speech discrimination by rhesus monkeys. *Perception & Psychophysics*, *17*(1), 9–16. https://doi.org/10.3758/BF03203991

Möttönen, R., Dutton, R., & Watkins, K. E. (2013). Auditory-motor processing of speech sounds. *Cerebral Cortex*, *23*(5), 1190–1197. https://doi.org/10.1093/cercor/bhs110

Möttönen, R., & Watkins, K. E. (2009). Motor representations of articulators contribute to categorical perception of speech sounds. *The Journal of Neuroscience*, *29*(31), 9819–9825. https://doi.org/10.1523/jneurosci.6018-08.2009

Mugitani, R., Kobayashi, T., & Hiraki, K. (2008). Audiovisual matching of lips and non-canonical sounds in 8-month-old infants. *Infant Behavior and Development*, *31*(2), 307–310. https://doi.org/10.1016/j.infbeh.2007.12.002

Murakami, T., Restle, J., & Ziemann, U. (2012). Effective connectivity hierarchically links temporoparietal and frontal areas of the auditory dorsal stream with the motor cortex lip area during speech perception. *Brain and Language*, *122*(3), 135–141. https://doi.org/10.1016/j.bandl.2011.09.005

Nachev, P., Kennard, C., & Husain, M. (2008). Functional role of the supplementary and pre-supplementary motor areas. *Nature Reviews Neuroscience*, *9*(11), 856–869. https://doi.org/10.1038/nrn2478

Nagel, I. E., Schumacher, E. H., Goebel, R., & D'Esposito, M. (2008). Functional MRI investigation of verbal selection mechanisms in lateral prefrontal cortex. *NeuroImage*, *43*(4), 801–807. https://doi.org/10.1016/j.neuroimage.2008.07.017

Nasir, S. M., & Ostry, D. J. (2009). Auditory plasticity and speech motor learning. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(48), 20470–20475. https://doi.org/10.1073/pnas.0907032106

Nolte, J. (2009). *The human brain: An introduction to its functional anatomy* (6th ed). Mosby/Elsevier.

Osnes, B., Hugdahl, K., & Specht, K. (2011). Effective connectivity analysis demonstrates involvement of premotor cortex during speech perception. *NeuroImage*, *54*(3), 2437–2445. https://doi.org/10.1016/j.neuroimage.2010.09.078

Papoutsi, M., De Zwart, J. A., Jansma, J. M., Pickering, M. J., Bednar, J. A., & Horwitz, B. (2009). From phonemes to articulatory codes: An fMRI study of the role of broca's area in speech production. *Cerebral Cortex*, *19*(9), 2156–2165. https://doi.org/10.1093/cercor/bhn239

Patterson, M. L., & Werker, J. F. (1999). Matching phonetic information in lips and voice is robust in 4.5-month-old infants. *Infant Behavior and Development*, *22*(2), 237–247. https://doi.org/10.1016/S0163-6383(99)00003-X

Patterson, M. L., & Werker, J. F. (2003). Two-month-old infants match phonetic information in lips and voice. *Developmental Science*, *6*(2), 191–196. https://doi.org/10.1111/1467-7687.00271

Pulvermüller, F. (2018). Neural reuse of action perception circuits for language, concepts and communication. *Progress in Neurobiology*, *160*, 1–44. https://doi.org/10.1016/j.pneurobio.2017.07.001

Pulvermüller, F., Huss, M., Kherif, F., del Prado Martín, F. M., Hauk, O., & Shtyrov, Y. (2006). Motor cortex maps articulatory features of speech sounds. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(20), 7865–7870. https://doi.org/10.1073/pnas.0509989103

Repp, B. H. (1984). Categorical perception: Issues, methods, findings. In N. J. Lass (Ed.), *Speech and language* (pp. 243–335). Elsevier. https://doi.org/10.1016/B978-0-12-608610-2.50012-1

Rizzolatti, G., & Craighero, L. (2004). The mirror neuron system. *Annual Review of Neuroscience*, *27*(1), 169–192. https://doi.org/10.1146/annurev.neuro.27.070203.144230

Rogalsky, C., Love, T., Tracy Love, Driscoll, D., Anderson, S. W., & Hickok, G. (2011). Are mirror neurons the basis of speech perception? evidence from five cases

with damage to the purported human mirror system. *Neurocase*, *17*(2), 178–187. https://doi.org/10.1080/13554794.2010.509318

Roland, P. E., Larsen, B., Lassen, N. A., & Skinhoj, E. (1980). Supplementary motor area and other cortical areas in organization of voluntary movements in man. *Journal of Neurophysiology*, *43*(1), 118–136. https://doi.org/10.1152/jn.1980.43.1.118

Roy, A. C., Craighero, L., Fabbri-Destro, M., & Fadiga, L. (2008). Phonological and lexical motor facilitation during speech listening: A transcranial magnetic stimulation study. *Journal of Physiology-Paris*, *102*(1), 101–105. https://doi.org/10.1016/j.jphysparis.2008.03.006

Sato, M., Troille, E., Ménard, L., Cathiard, M.-A., & Gracco, V. L. (2013). Silent articulation modulates auditory and audiovisual speech perception. *Experimental Brain Research*, *227*(2). https://doi.org/10.1007/s00221-013-3510-8

Schmitz, J., Bartoli, E., Maffongelli, L., Fadiga, L., Sebastian-Galles, N., & D'Ausilio, A. (2019). Motor cortex compensates for lack of sensory and motor experience during auditory speech perception. *Neuropsychologia*, *128*, 290–296. https://doi.org/10.1016/j.neuropsychologia.2018.01.006

Schomers, M. R., Kirilina, E., Weigand, A., Bajbouj, M., & Pulvermüller, F. (2015). Causal influence of articulatory motor cortex on comprehending single spoken words: TMS evidence. *Cerebral Cortex*, *25*(10), 3894–3902. https://doi.org/10.1093/cercor/bhu274

Schomers, M. R., & Pulvermüller, F. (2016). Is the sensorimotor cortex relevant for speech perception and understanding? an integrative review. *Frontiers in Human Neuroscience*, *10*, 435. https://doi.org/10.3389/fnhum.2016.00435

Segawa, J. A., Tourville, J. A., Beal, D. S., & Guenther, F. H. (2015). The neural correlates of speech motor sequence learning. *Journal of Cognitive Neuroscience*, *27*(4), 819–831. https://doi.org/10.1162/jocn_a_00737

Shima, K., & Tanji, J. (2000). Neuronal activity in the supplementary and presupplementary motor areas for temporal organization of multiple movements. *Journal of Neurophysiology*, *84*(4), 2148–2160. https://doi.org/10.1152/jn.2000.84.4.2148

Skipper, J. I., Nusbaum, H. C., & Small, S. L. (2005). Listening to talking faces: Motor cortical activation during speech perception. *NeuroImage*, *25*(1), 76–89. https://doi.org/10.1016/j.neuroimage.2004.11.006

Smalle, E. H. M., Rogers, J., & Möttönen, R. (2015). Dissociating contributions of the motor cortex to speech perception and response bias by using transcranial magnetic stimulation. *Cerebral Cortex*, *25*(10), 3690–3698. https://doi.org/10.1093/cercor/bhu218

Stokes, R. C., Venezia, J. H., & Hickok, G. (2019). The motor system's [modest] contribution to speech perception. *Psychonomic Bulletin & Review*, *26*(4), 1354–1366. https://doi.org/10.3758/s13423-019-01580-2

Szenkovits, G., Peelle, J. E., Norris, D., & Davis, M. H. (2012). Individual differences in premotor and motor recruitment during speech perception. *Neuropsychologia*, *50*(7), 1380–1392. https://doi.org/10.1016/j.neuropsychologia.2012.02.023

Tanji, J., & Shima, K. (1994). Role for supplementary motor area cells in planning several movements ahead. *Nature*, *371*(6496), 413–416. https://doi.org/10.1038/371413a0

Tremblay, P., Deschamps, I., & Gracco, V. L. (2016). Neurobiology of speech production: A motor control perspective. In G. Hickok & S. L. Small (Eds.), *Neurobiology of language* (pp. 741–750). Elsevier/AP, Academic Press is an imprint of Elsevier.

Tremblay, P., & Gracco, V. L. (2006). Contribution of the frontal lobe to externally and internally specified verbal responses: fMRI evidence. *NeuroImage*, *33*(3), 947–957. https://doi.org/10.1016/j.neuroimage.2006.07.041

Tremblay, P., & Gracco, V. L. (2009). Contribution of the pre-SMA to the production of words and non-speech oral motor gestures, as revealed by repetitive transcranial magnetic stimulation (rTMS). *Brain Research*, *1268*, 112–124. https://doi.org/10.1016/j.brainres.2009.02.076

Waldstein, R. S. (1990). Effects of postlingual deafness on speech production: Implications for the role of auditory feedback. *The Journal of the Acoustical Society of America*, *88*(5), 2099–2114. https://doi.org/10.1121/1.400107

Watkins, K. E., Antonio P. Strafella, Antonio P. Strafella, Strafella, A. P., & Paus, T. (2003). Seeing and hearing speech excites the motor system involved in speech production. *Neuropsychologia*, *41*(8), 989–994. https://doi.org/10.1016/s0028-3932(02)00316-0

Weeks, R., Honda, M., Catalan, M.-J., & Hallett, M. (2001). Comparison of auditory, somatosensory, and visually instructed and internally generated finger movements: A PET study. *NeuroImage*, *14*(1), 219–230. https://doi.org/10.1006/nimg.2001.0780

Werker, J. F., Gilbert, J. H. V., Humphrey, K., & Tees, R. C. (1981). Developmental aspects of cross-language speech perception. *Child Development*, *52*(1), 349. https://doi.org/10.2307/1129249

Werker, J. F., & Tees, R. C. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development*, *7*(1), 49–63. https://doi.org/10.1016/S0163-6383(84)80022-3

Wiese, H., Stude, P., Nebel, K., De Greiff, A., Forsting, M., Diener, H. C., & Keidel, M. (2004). Movement preparation in self-initiated versus externally triggered movements: An event-related fMRI-study. *Neuroscience Letters*, *371*(2), 220–225. https://doi.org/10.1016/j.neulet.2004.08.078

Wilson, S. M., Stephen M. Wilson, & Iacoboni, M. (2006). Neural responses to non-native phonemes varying in producibility: Evidence for the sensorimotor nature of speech perception. *NeuroImage*, *33*(1), 316–325. https://doi.org/10.1016/j.neuroimage.2006.05.032

Wilson, S. M., Stephen M. Wilson, Saygin, A. P., Sereno, M. I., & Iacoboni, M. (2004). Listening to speech activates motor areas involved in speech production. [MAG ID: 1984689963]. *Nature Neuroscience*, *7*(7), 701–702. https://doi.org/10.1038/nn1263

Winn, M. B. (2019). *GUI-based wizard for creating realistic vowel formant continua from modified natural speech* (Version 38). http://www.mattwinn.com/praat.html

Winn, M. B. (2020). Manipulation of voice onset time in speech stimuli: A tutorial and flexible praat script. *The Journal of the Acoustical Society of America*, *147*(2), 852–866. https://doi.org/10.1121/10.0000692

Yeung, H. H., & Werker, J. F. (2013). Lip movements affect infants' audiovisual speech perception. *Psychological Science*, *24*(5), 603–612. https://doi.org/10.1177/0956797612458802