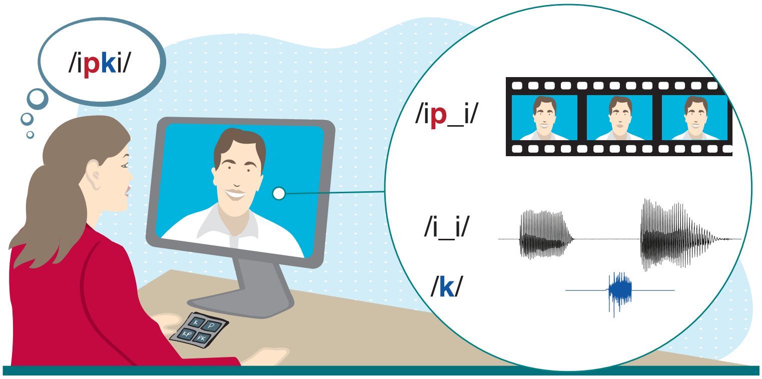
CONTRIBUTIONS TO
HEARING RESEARCH

Volume 44

Juan Camilo Gil-Carvajal

Towards a feature-based theory of audiovisual integration of speech



DTU Health Tech
Department of Health Technology

Towards a feature-based theory of audiovisual integration of speech

PhD thesis by
Juan Camilo Gil-Carvajal

Preliminary version: July 31, 2020

CHESS Oticon Centre of Excellence
for Hearing and Speech Sciences

Technical University of Denmark

2020

Preprint version for the assessment committee.
Pagination will differ in the final published version.
Cover picture: © Ana Sofía Patiño-Peláez, 2020

This PhD dissertation is the result of a research project carried out at the Cognitive Systems Section, Department of Applied Mathematics and Computer Science and the Hearing Systems Section, Department of Health Technology, Technical University of Denmark.

The project was partly financed by the Oticon Centre of Excellence for Hearing and Speech Sciences (2/3) and by the Technical University of Denmark (1/3).

Supervisors

Assoc. Prof. Tobias S. Andersen
Cognitive Systems Section
Department of Applied Mathematics and Computer Science
Technical University of Denmark
Kgs. Lyngby, Denmark

Prof. Torsten Dau
Hearing Systems Section
Department of Health Technology
Technical University of Denmark
Kgs. Lyngby, Denmark

Abstract

Speech perception is facilitated by seeing the mouth movements of the talker, which is particularly useful in noisy listening environments. The mouth gestures of the talker can also modify the auditory phonetic percept. This is evidenced in the McGurk effect, a well-known audiovisual phenomenon that demonstrates the audiovisual nature of speech perception. The McGurk effect occurs when a speech sound is presented simultaneously with incongruent articulatory mouth movements corresponding to another speech token. Sometimes the McGurk effect results in a fusion of the two consonants presented. For example, dubbing auditory /aba/ onto visual /aga/ usually produces the audiovisual perception of a third consonant /ada/. In contrast, when auditory /aga/ is dubbed onto visual /aba/, the McGurk effect leads to a combination illusion of hearing the two consonants presented, typically /abga/ or /agba/. Despite decades of research in audiovisual speech, it has remained unclear why some audiovisual stimuli elicit McGurk fusions while others produce McGurk combinations, and which are the audiovisual phonetic features that affect the perceived consonant order in the latter case.

This PhD project investigated audiovisual integration of speech, with a particular focus on providing behavioral evidence for a featured-based model of audiovisual integration of speech. The role of timing of phonetic features on audiovisual integration, as well as the perceived consonant order in the McGurk combination illusion, were the key aspects addressed in this thesis. In one study, the integration of audiovisual speech features was tested using stimuli that consisted of consonant clusters and single consonants, which produced novel illusory percepts. For example, auditory /abga/ dubbed onto visual /aga/ was mostly perceived as /adga/, which indicated a partial fusion illusion between the initial auditory consonant /b/ and the initial visual gesture for /g/, while the perception of the subsequent auditory consonant /g/ was unaffected. Thus, the results suggested the existence of sequential audiovisual features that are integrated separately.

In the second study, the audiovisual perception of phonetic features was investigated in McGurk combination illusions. The effect of timing on the perceived consonant order was investigated by varying the audiovisual stimulus onset asynchrony (SOA) or the syllabic context by articulating the consonant in the syllable onset or offset. While varying SOA mostly affected the strength of audiovisual integration, the syllabic context mostly influenced the perceived consonant order. Notably, the asymmetry of the audiovisual temporal integration window was found to be the opposite for vowel-consonant and consonant-vowel stimuli. These results supported the existence of articulatory constraints on audiovisual integration, which are imposed by the visual speech gestures of the talker. The third study further explored whether the effect of syllabic context on the perceived consonant order could be explained by a feature-based approach. The findings showed that the perceived consonant order in McGurk combinations is driven by the timing between the acoustic release burst and the mouth movements of the talker, which provides further support for a feature-based model of audiovisual integration of speech. Overall, this thesis provided experimental evidence that constitutes a valuable foundation for the development of a feature-based model of audiovisual integration of speech.

Resumé

Talopfattelse faciliteres af at se talerens mundbevægelser, hvilket er særlig nyttigt i støjende miljøer. Talerens mundbevægelser kan også ændre den auditive fonetiske opfattelse. Dette fremgår af McGurk-effekten, et velkendt audiovisuelt fænomen, der demonstrerer den audiovisuelle karakter af taleopfattelse. McGurk-effekten opstår, når en talelyd gengives samtidigt med uoverensstemmende mundbevægelser, der svarer til en anden talelyd. Nogle gange resulterer McGurk-effekten i en fusion af de to præsenterede konsonanter. For eksempel skaber dubbing af auditivt /aba/ på visuelt /aga/ ofte den audiovisuelle opfattelse af en tredje konsonant /ada/. I modsætning hertil, når auditivt /aga/ dubbes på visuelt /aba/, fører McGurk-effekten til opfattelsen af en kombinationsillusion af at høre de to præsenterede konsonanter, typisk /abga/ eller /agba/. På trods af årtier med forskning i audiovisuel tale er det fortsat uklart, hvorfor nogle audiovisuelle stimuli fremkalder McGurk-fusioner, mens andre producerer McGurk-kombinationer, og hvilke audiovisuelle fonetiske features der påvirker den opfattede konsonantrækkefølge i sidstnævnte tilfælde.

Dette ph.d.-projekt undersøgte audiovisuel integration af tale, med et særligt fokus på at give evidens fra adfærdsstudier for en feature-baseret model af audiovisuel integration af tale. Timing af fonetiske features, såvel som den opfattede konsonantrækkefølge i McGurk-kombinationsillusionen, var de vigtigste aspekter der blev behandlet i denne afhandling. I det første studie blev integrationen af audiovisuelle features testet ved hjælp af stimuli, der bestod af konsonantklynger og enkeltkonsonanter, hvilket frembragte nye illusoriske opfattelser. F.eks. blev auditiv /abga/ dubbet på visual /aga/ ofte opfattet som /adga/, hvilket indikerede en delvis fusionsillusion mellem den indledende auditive konsonant /b/ og den indledende visuelle gestus for /g/, mens opfattelsen af den efterfølgende auditive konsonant /g/ ikke blev påvirket. Resultaterne antydede således eksistensen af sekventielle audiovisuelle fonetiske features, der integreres separat.

I det andet studie blev den audiovisuelle opfattelse af fonetiske features undersøgt i McGurk-kombinationsillusioner. Virkningen af timing på den opfattede konsonantorden blev undersøgt ved at variere den audiovisuelle asynkroni (SOA) eller den syllabiske kontekst ved at artikulere konsonanten i starten eller slutningen af stavelsen. Mens ændring af SOA mest påvirkede styrken af den audiovisuelle integration, påvirkede den syllabiske kontekst mest den opfattede konsonantrækkefølge. Bemærkelsesværdigt viste det sig, at asymmetrien i det audiovisuelle tidslige integrationsvindue var modsat for vokal-konsonant

og konsonant-vokal stimuli. Disse resultater understøtter eksistensen af artikulationsbegrænsninger for audiovisuel integration, der pålægges af talerens visuelle talebevægelse. Den tredje undersøgelse undersøgte yderligere, om virkningen af den syllabiske kontekst på den opfattede konsonanttrækkefølge kunne forklares med en feature-baseret tilgang. Resultaterne viste, at den opfattede konsonantrækkefølge i McGurk-kombinationer er drevet af timingen mellem det aspirationslyden og talerens mundbevægelser, hvilket yderligere støtter en feature-baseret model af audiovisuel integration af tale. Samlet set leverede denne afhandling empirisk evidens, der udgør et værdifuldt fundament for udviklingen af en feature-baseret model for audiovisuel integration af tale.

Resumen

La percepción del habla se facilita al ver los movimientos de la boca del interlocutor, lo cual es particularmente útil en ambientes de escucha ruidosos. Los gestos de la boca del interlocutor también pueden modificar la percepción fonética auditiva. Esto se evidencia en el efecto McGurk, un conocido fenómeno audiovisual que demuestra la naturaleza audiovisual de la percepción del habla. El efecto McGurk se produce cuando un sonido del habla se reproduce simultáneamente con movimientos discrepantes de la boca correspondientes a otro sonido del habla. A veces el efecto McGurk da lugar a una fusión de las dos consonantes presentadas. Por ejemplo, el doblaje del componente auditivo /aba/ en el componente visual /aga/ suele producir la percepción audiovisual de una tercera consonante /ada/. Por el contrario, cuando el componente auditivo /aga/ es doblado en el componente visual /aba/, el effecto McGurk produce la ilusión de oir una combinación de las dos consonantes presentadas, usualmente /abga/ o /agba/. A pesar de décadas de investigación en el habla audiovisual, todavía no esta completamente claro por qué algunos estímulos audiovisuales provocan fusiones de McGurk mientras que otros producen combinaciones de McGurk, y cuáles son las características fonéticas audiovisuales que afectan el orden percibido de las consonantes en este último caso.

En este proyecto de doctorado se investigó la integración audiovisual del habla, con un enfoque particular en proporcionar evidencias de comportamiento perceptual para un modelo de integración audiovisual del habla basado en sus elementos característicos. Los aspectos clave abordados en esta tesis fueron el papel de la sincronización temporal de los elementos phonéticos en la integración audiovisual, así como el orden percibido de las consonantes en la ilusión de combinación del efecto McGurk. En el primer estudio se probó la integración de los elementos característicos del habla audiovisual utilizando estímulos que consistían en consonantes agrupadas y consonantes únicas, los cuales produjeron percepciones ilusorias novedosas. Por ejemplo, el componente auditivo /abga/ doblado en el componente visual /aga/ se percibió principalmente como /adga/, indicando una ilusión de fusión parcial entre la consonante auditiva inicial /b/ y el gesto visual inicial para /g/, mientras que la percepción de la consonante auditiva posterior /g/ no se vio afectada. Por lo tanto, los resultados sugieren la existencia de elementos característicos audiovisuales secuenciales que se integran por separado.

En el segundo estudio, se investigó la perception audiovisual de los elementos característicos fonéticos en las ilusiones de combinación de McGurk. El efecto de la sincronización temporal en el orden percibido de las consonantes se investigó variando la asincronía de inicio del estímulo audiovisual (SOA) o el contexto silábico mediante la articulación de la consonante en el inicio o el final de la sílaba. Si bien la variación de la SOA afectó principalmente la fuerza de la integración audiovisual, el contexto silábico influyó sobre todo en el orden de las consonantes percibidas. Cabe señalar que la asimetría de la ventana temporal de integración audiovisual fue la contraria para los estímulos de combinación McGurk con vocal-consonante y consonante-vocal. Estos resultados respaldaron la existencia de limitaciones articulatorias en la integración audiovisual, las cuales son impuestas por los gestos visuales del interlocutor. En el tercer estudio se estudió en mayor detalle si el efecto del contexto silábico en el orden percibido de las consonantes podía explicarse con un enfoque basado en los elementos característicos audiovisuales del estímulo. Los resultados mostraron que el orden percibido de las consonantes en las combinaciones de McGurk se define por la sincronización temporal entre el estallido de liberación acústica y los movimientos de la boca del interlocutor, lo que, en consecuencia, proporciona un mayor apoyo a un modelo de integración audiovisual del habla basado en sus elementos característicos. En general, esta tesis aportó pruebas experimentales que establecen un valioso fundamento para el desarrollo de un modelo de integración audiovisual del habla basado en sus elementos característicos.

Acknowledgments

First, I would like to thank my supervisors without whom this project would not have been possible. Tobias thank you for your support, guidance and for being so passionate about the project. You were really inspiring and motivating at all times. Thank you also Torsten for your continued support and encouragement since my Master's thesis. You have always inspired me to reach new goals.

I would also like to thank Wanja Andersen and Caroline van Oosterhout, for helping me with all administrative aspects of the project during all these years. Special thanks go to Johannes Zaar for many insightful discussions and support at the beginning of this project.

Further, I would like to thank Jean-Luc Schwartz for welcoming me in his lab and for creating a great work environment during my external research stay. Thank you also Christophe Savariaux for helping me recording the audiovisual stimuli of study 3, the lip trajectories presented in the introduction and for the fruitful discussions. Thank you to all the people at GIPSA-LAB who made my external stay a memorable experience.

I want to thank also my colleagues and friends at the Cognitive Systems Section, Hearing Systems Section and the Acoustic Technology Group. It has been a great experience to go through this journey with you. Thanks for being so supportive, for the great work atmosphere and for all the fun.

Finally, thanks to my family for their unconditional love and support despite the distance, and to my friends in Denmark who always made me feel at home. And specially, thank you Valen for your tireless patience and support, you made all this possible.

Related publications

Journal papers

- Gil-Carvajal, J. C., and Andersen, T. S. (**2020**). "Audiovisual integration of single consonants and consonant clusters," unsubmitted manuscript.
- Gil-Carvajal, J. C., Dau, T., and Andersen, T. S. (**2020**). "Order matters: timing and syllabic context influence audiovisual integration of speech," Journal of Experimental Psychology: Human Perception and Performance, submitted.
- Gil-Carvajal, J. C., Schwartz, J.-L., Dau, T., and Andersen, T. S. (**2020**). "Feature-based audiovisual speech integration of multiple sequences," Trends in Hearing, under review.

Conference papers

- Gil-Carvajal, J., Schwartz, J-L., Dau, T., and Andersen, T. (**2020**). "Feature-based audiovisual speech integration of multiple streams," Proceedings of the International Symposium on Auditory and Audiological Research. 7, 333-340.

Published abstracts

- Gil-Carvajal, J. C., Schwartz, J-L., Dau, T., and Andersen, T. S. (**2019**). "Featured-based audiovisual speech integration of multiple streams," Proceedings of the International Symposium on Audiology and Audiological Research, Nyborg, Denmark, August 2019.
- Gil-Carvajal, J. C., Dau, T., and Andersen, T. S. (**2018**). "Consonant-order reversals in the McGurk combination illusion," Proceedings of the 19th

Annual International Multisensory Research Forum, Toronto, Canada, June 2018.

- Andersen, T. S., and Gil-Carvajal, J. C. (**2018**). “Audiovisual integration of consonant clusters,” Proceedings of the 19th Annual International Multisensory Research Forum, Toronto, Canada, June 2018.
- Gil-Carvajal, J. C., Dau, T., and Andersen, T. S. (**2017**). “Audiovisual illusions in consonant perception and implications for speech perception,” Pire Workshop/Summer School, Girona, Spain, June 2017.

Datasets

- Gil-Carvajal, J. C., Lindborg, A.C., and Andersen, T. S. (**2020**). “Audiovisual integration of single consonants and consonant clusters (audiovisual dataset),” Zenodo. http://doi.org/10.5281/zenodo.3969033.
- Gil-Carvajal, J. C., Dau, T., and Andersen, T. S. (**2020**). “Order matters: timing and syllabic context influence audiovisual integration of speech (audiovisual dataset),” Zenodo. http://doi.org/10.5281/zenodo.3970148.
- Gil-Carvajal, J. C., Schwartz, J-L., Savariaux, C., and Andersen, T. S. (**2020**). “Feature-based audiovisual speech integration of multiple sequences (audiovisual dataset),” Zenodo. http://doi.org/10.5281/zenodo.3970152.

Contents

1

General introduction

Speech perception is naturally associated with the processing of the voice. There is, however, a great deal of information stemming from a talker's visual speech gestures that facilitates the perception of spoken language. when we find ourselves in face-to-face settings, it is natural to direct the gaze towards the interlocutor's face when they speak. This spontaneous reaction seems trivial, but in challenging (i.e., noisy) situations, it could make the difference between understanding a message or not. After decades of research, it is still unclear which and how auditory and visual speech features are integrated. This thesis thus aims to contribute to a better understanding of how humans combine phonetic information from sight and hearing to comprehend speech.

1.1 Advantage of audiovisual speech perception

The benefit of visual speech gestures to speech perception has long been known. This was first demonstrated by the pioneering studies of Sumby and Pollack (1954) as well as O'Neill (1954), which showed that speech intelligibility improves when hearing *and* seeing the talker, as opposed to listening alone. These studies showed a greater contribution of visual speech at low signal-to-noise ratios in which there was more "room" for improvement. The advantage of seeing the talker's articulatory gestures for speech comprehension has been shown in many other studies in which the acoustic speech signal was noisy (Miller and Nicely, 1955; Binnie et al., 1974; Erber, 1975; MacLeod and Summerfield, 1987; Helfer, 1997; Ross et al., 2007; Schwartz et al., 2004) or clean but semantically complex (Reisberg et al., 1987). Visual speech information is also useful for speech detection, resulting in 1-3 dB improvement in masked thresholds (Grant and Seitz, 2000). Furthermore, visible speech gestures contribute to speech development in young children (Erber, 1972; Kuhl and Meltzoff, 1982; Legerstee, 1990; Arnold and Köpsel, 1996; Schorr et al., 2005), provide substantial benefit to speech recognition in hearing-impaired as well as cochlear implant listeners

(Grant et al., 1998; Rouger et al., 2007; Kaiser et al., 2012), and have been shown to accelerate the cortical processing of the auditory speech signal (Van Wassenhove et al., 2005).

1.2 Auditory and visual speech cues

Speech perception is a remarkable ability that often entails parsing a continuous speech stream into discrete sound events (i.e., words, syllables, segments), which are easier to decode by the human perceptual system. Speech perception is then mediated by language rules that should be common to the talker and listener for an effective communication. Phonemes are the smallest units of sound that enable the distinction of one word from another (Reisberg, 2010). For example, replacing the initial phoneme /d/ with /b/ in the English word "dad" produces a different word "bad". Phonetic perception relies on the auditory cues resulting from the movement of the speech articulators, such as the lips, tongue and teeth. For instance, when producing stop consonants (i.e., /b/ and /p/) the articulators completely occlude the airflow from the lungs. This builds up pressure inside the vocal tract, which is subsequently released as a consonantal burst after the articulators move apart. The period of occlusion can be heard as a silence interval or can be accompanied by voicing. This is because during this period, the vocal folds may vibrate or not, which produces either voiced (i.e., /b/, /d/, /g/) or unvoiced stop consonants (i.e., /p/, /t/, /k/), respectively.

The vibration of the vocal folds is not the only distinction between voiced and unvoiced consonants, as the latter tend to be produced with higher intensity, which in turn gives rise to a period of aspiration that follows the release burst (Halle et al., 1957; Roach, 2000). The opening and closing movements of the articulators during consonant production can also influence the perception of the adjacent vowel formant frequencies. These auditory cues are known as formant transitions, and their frequency content depends on the preceding and subsequent vowels (Stevens and Blumstein, 1978). The consonantal burst, aspiration, silence and formant transitions are all important cues for the auditory perception of stop consonants (Halle et al., 1957; Menon et al., 1974; Dorman and Raphael, 1980; Li et al., 2010). These cues are displayed in the spectrograms of Figure 1.1 for /p/ (left bottom panel) and /k/ (right bottom panel), respectively. For these two consonants, the formant transitions (outlined with blue lines) as well as the burst and aspiration (highlighted in green) have different acoustic content, and hence, either of these cues can serve to distinguish one consonant from the other.

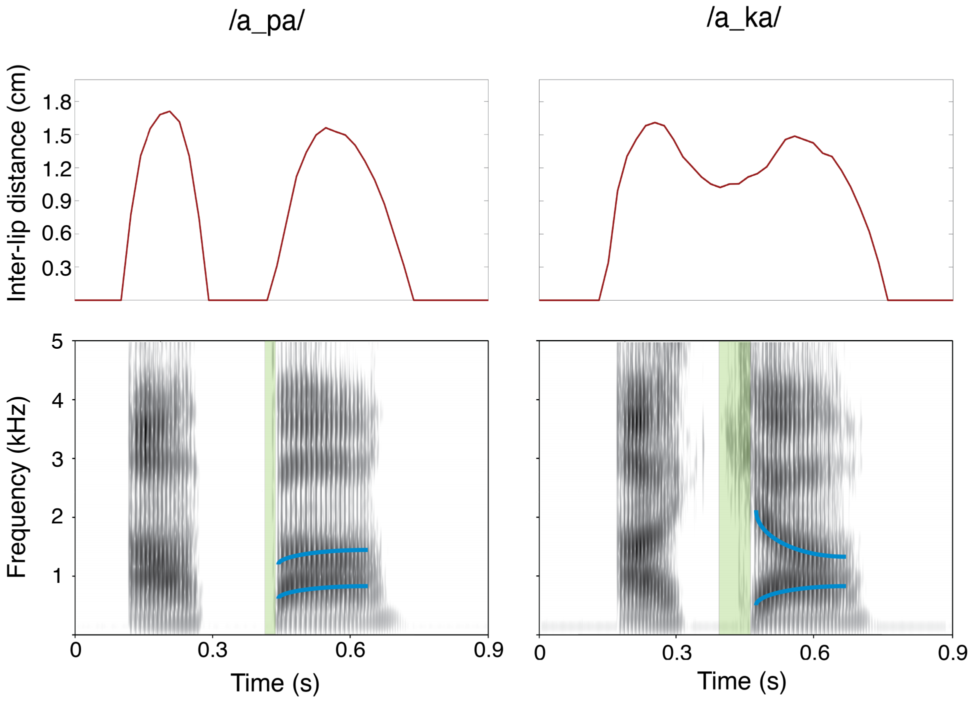


Figure 1.1: Lip trajectories (top) and spectrograms (bottom) for /p/ (left) and /k/ (right), respectively. The release burst and aspiration are highlighted in green, while the formant transitions (F1 and F2) are outlined with blue lines. The underscore in the articulations represents an intersyllabic silence.

In the visual domain, the lips, tongue and teeth also provide visual cues for phonetic perception (Summerfield, 1992). The opening and closing gestures of the speech articulators are usually visible, as illustrated by the lip trajectories in Figure 1.1 for /p/ (top left panel) and /k/ (top right panel), respectively. The movement of the speech articulators provide information about the identity of the consonant articulated as well as temporal information that can serve to direct the listener's attention, which in turn enhances speech processing (Peelle and Sommers, 2015). Almost all available visual speech cues from a talker's face can be extracted from the movements of the oral area alone (Ijsseldijk, 1992; Marassa and Lansing, 1995; Thomas and Jordan, 2004). The opening area of the mouth movements has indeed been shown to correlate with the amplitude of the auditory envelope, such that large opening movements usually correspond to louder voices (Chandrasekaran et al., 2009). While seeing only the lip movements of the speaker enhances speech recognition (Summerfield, 1979), adding the visibility of the teeth provides a greater enhancement (Summerfield, 1989). Moreover, lip rounding affects the identification of phonetic contrasts (i.e., /s/ vs. /ʃ/ and /i/ vs. /u/, Winn et al., 2013). The visual influence of a talker's facial movements on speech perception is robust. This is maintained across a broad range of viewing angles (Jordan and Thomas, 2001), after presenting only illuminated moving dots attached to the talker's articulators (Rosenblum and Saldaña, 1996), and does not require direct visual fixation of the listener (Smeele et al., 1998). Extra oral facial regions (i.e., cheeks, eyes, etc.) can also influence speech perception (Scheinberg, 1980; Preminger et al., 1998; Thomas and Jordan, 2004) which, could be due to the high correlation of the movement of these areas with the displacement of the oral articulators (Munhall and Vatikiotis-Bateson, 2013). This intrinsic relationship between visible articulatory gestures and the voice could partly explain the visual enhancement of speech perception observed in behavioral experiments.

Another reason why seeing a talker aids speech understanding is that visible speech gestures often complement acoustic speech information. While some speech features are difficult to hear, they are generally easy to see and vice versa (Miller and Nicely, 1955; Walden et al., 1977; Rosenblum and Saldaña, 1996). For instance, the information about the place of articulation, which helps to distinguish where in the vocal track a consonant is produced, is easily degraded by acoustic noise (Miller and Nicely, 1955) but is visually prominent (i.e., one can see the difference between a labial consonant /p/ and a non-labial consonant /k/, as evidenced by the different lip trajectories in Figure 1.1). In contrast, information about manner of articulation and voicing, which help to distinguish the way a consonant is articulated, is not very informative visually (Green and Kuhl, 1991; Brancazio, 2004) but is acoustically robust even under noisy conditions (i.e., one can hear the difference between a voiced consonant /b/ and an unvoiced consonant /p/).

1.3 The McGurk effect

The audiovisual nature of speech perception is convincingly demonstrated by the McGurk effect (McGurk and MacDonald, 1976). This was discovered by McGurk and MacDonald in 1976 while preparing stimuli to investigate whether incongruent audiovisual speech (i.e., dubbing the sound /baba/ onto visual articulatory gestures /gaga/) could disrupt infant's visual attention (MacDonald, 2018). To their surprise, when seeing and hearing the incongruent stimuli, they perceived speech sounds different than the recorded. In contrast, while hearing with closed eyes, the recorded speech sounds could be correctly heard. The McGurk effect was then taken as evidence that auditory speech can be categorically altered by a simultaneous presentation of incongruent visual speech (MacDonald and McGurk, 1978; Rosenblum and Saldaña, 1996; Brancazio et al., 2003; Tiippana, 2014). After its discovery, the McGurk effect has contributed greatly to the field of speech perception, with more than 7000 citations of the original nature manuscript at present (source: Google scholar, July 2020).

The McGurk effect has been exploited as a research tool for understanding the audiovisual integration of speech. This is done by comparing the rate of responses obtained through the presentation of McGurk stimuli to those obtained through the unimodal auditory presentation (Tiippana, 2014; Alsius et al., 2017). If the listeners are instructed to respond according to what they heard, any deviation from the expected phonetic categorization can then be attributed to audiovisual speech integration. Thus, the unimodal perception of the visual component also needs to be considered to ensure the correct perception of the visual information, and hence, that the audiovisual response can be attributed to a multisensory integration process. It has been indicated, however, that the perception and processing of McGurk stimuli differs to that of natural (congruent) speech (for a review, see Alsius et al., 2017). This highlights the need for obtaining a more clear understanding of the factors that influence the perception of McGurk stimuli.

The McGurk effect occurs even when the listeners are aware of the artificial manipulation that gives rise to the illusion (Rosenblum and Saldaña, 1996), despite a mismatch in the gender of the talker's voice and face (Green et al., 1991), and even after degrading the visual speech information (i.e., Rosenblum and Saldaña, 1996; MacDonald et al., 2000). However, the reported rate of McGurk responses has also varied substantially across studies (Alsius et al., 2017). This has been attributed to several factors, such as differences in the quality of stimuli, the diversity of experimental designs and the differences in the audiovisual stimulus pairings (Sekiyama, 1993; Basu Mallick et al., 2015; Alsius et al., 2017). This indicates that there is still a need for standardizing the methodology for recording, preparing, presenting and reporting the McGurk stimuli, which could improve the reliability and reproducibility across studies. The implementation of such standardization might be difficult without a clear comprehension of the stimulus features that are most important for eliciting the McGurk effect.

Another known source of variability in the McGurk effect is the inter-individual sensitivity to stimuli, as some listeners almost always perceive the McGurk effect, whereas others almost never experience it (Basu Mallick et al., 2015). Despite this variability, the susceptibility to the illusion is maintained over a one-year test-retest interval (Basu Mallick et al., 2015), which makes the McGurk effect a reliable tool for investigating audiovisual integration of speech. The individual perceptual differences in the McGurk effect have been associated with the listener's native language and cultural background (Sekiyama, 1993). However, in a large sample study, Magnotti et al. (2015) showed no significant differences in the rate of McGurk responses between native Mandarin Chinese and native American English speakers. The inter-individual differences reported in the McGurk effect could reflect the existence of individual processing differences, as evidenced in audiovisual studies with clinical populations (Hamilton et al., 2006; Rouger et al., 2007), which is also supported by the fact that not all listeners exhibit the same benefit from seeing a talker's speech gestures (Grant et al., 1998; Tye-Murray et al., 2007).

1.3.1 Variants of the McGurk effect

The McGurk effect can elicit different audiovisual percepts depending on the specific audiovisual stimulus pairings, as shown in Figure 1.2. The classical and most investigated percept is the so-called *fusion* illusion. This is typically produced when an acoustic labial consonant (i.e., /aba/) is dubbed onto the visual articulatory gestures of a non-labial consonant (i.e., /aga/), which is perceived as a third consonant (i.e., /ada/) that reflects the fusion of the two places of articulation (Van Wassenhove, 2013). Another possible outcome of the McGurk effect are the *visually-driven* (or "visual dominance") responses (Rosenblum and Saldaña, 1992; Van Wassenhove, 2013; Tiippana, 2014), which occur when the percept only accounts for the visual place of articulation (i.e., auditory /aba/ + visual /ada/ heard as /ada/). Furthermore, the McGurk effect can elicit percepts that combine the place of articulation of the acoustically and visually presented consonants, and hence, have been coined as McGurk *combinations* (McGurk and MacDonald, 1976; Green and Norrix, 1997; Soto-Faraco and Alsius, 2009). McGurk combinations typically occur with the reversed audiovisual pairing than McGurk fusions. That is, when a non-labial acoustic consonant (i.e., /aga/) is dubbed onto a visual labial consonant (i.e., /aba/), which results in a consonant cluster percept that contains the two consonants presented (i.e., either /abga/ or /agba/). Unlike fusions, the McGurk combinations have received little scientific attention, and therefore, have remained poorly understood.

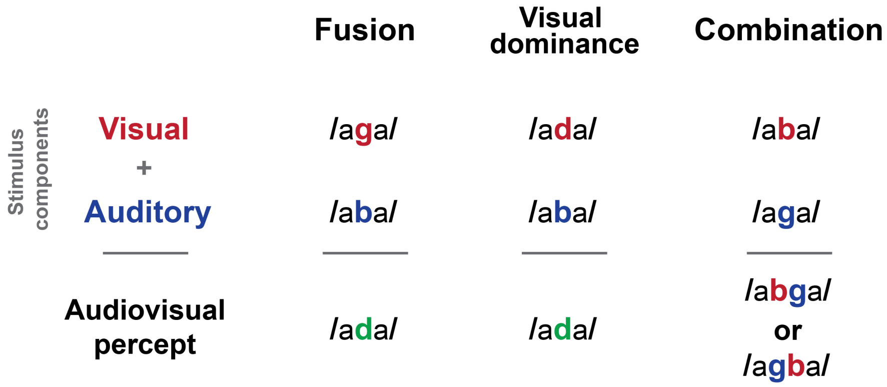


Figure 1.2: Example of the stimulus pairings and typical audiovisual percepts for the different variants of the McGurk effect.

McGurk fusions have been explained as the best plausible solution that reconciles the visual place cues and the auditory manner (and voicing) cues (MacDonald and McGurk, 1978; Goldstein and Fowler, 2003), or as the integration of the visual and auditory place cues (Van Wassenhove, 2013; Tiippana, 2014). In contrast, in the case of McGurk combinations, it has been suggested that the phonetic incongruence between the auditory and visual information cannot be fit into a unified single consonant percept (McGurk and MacDonald, 1976). Despite previous attempts at explaining the nature of McGurk combination responses, it is still unclear how the auditory and visual components are integrated to produce these type of audiovisual percepts. This thesis has been particularly focused on investigating the perception of the McGurk combination illusion. The following section describes the hypotheses tested in this thesis, which are pertinent to the study of McGurk combinations.

1.3.2 McGurk combination illusion: background and hypotheses

Another puzzling aspect of McGurk combinations is that the perceived consonant order seems to be more commonly reported with the bilabial consonant leading, whereas the reverse consonant order tends to be perceived sporadically (i.e., MacDonald and McGurk, 1978; Massaro and Cohen, 1993; Jiang and Bernstein, 2011; Soto-Faraco and Alsius, 2007; Soto-Faraco and Alsius, 2009; Jiang and Bernstein, 2011). The fact that the visual consonant tends to lead the auditory consonant in McGurk combinations might be due to the temporal processing differences across modalities (Massaro et al., 1996), which is referred to as the timing hypothesis throughout this thesis. Previous studies that investigated this effect of timing on the McGurk combination indicated that varying the stimulus onset asynchrony (SOA) between the visual and auditory speech signals modulates the rate of cluster responses, but does not affect the perceived order of consonants (Massaro and Cohen, 1993; Soto-Faraco and Alsius, 2009). For instance, using consonant-vowel (CV) stimuli, Massaro and Cohen (1993) showed that irrespective of the audiovisual SOA, auditory /da/ paired with visual /ba/ was almost always perceived as /bda/ but not as /dba/. The reason for the imbalance in favor of responses with the bilabial consonant leading was attributed to the high similarity of the visual component to only one cluster articulation (i.e., visual /ba/ is similar to /bda/ but not to /dba/). In contrast, the study of Hampson et al. (2003) reported an effect of varying audiovisual SOA on the perceived consonant order of vowel-consonant (VC) McGurk combinations. However, since varying audiovisual SOA can also affect the strength of the audiovisual integration, this strategy alone could be problematic for investigating the perceived consonant order of the cluster response. Thus, although a few studies have addressed the effect of timing on the McGurk combination, their results might be inconclusive. To understand how the brain integrates auditory and visual speech components in the McGurk combination, it is then important to further explore the timing hypothesis by manipulating the stimulus components while maintaining a similar strength of integration of the audiovisual cluster response, as investigated in Chapter 3.

Another hypothesis for the perception of McGurk combinations postulates that the release burst and aspiration of the acoustic consonant contribute to the occurrence of combination responses. This was tested by Green and Norrix (1997), who showed that removing the consonantal burst and aspiration reduced the rate of McGurk combinations but not the frequency of fusion responses. The study of Colin et al. (2002) further supported this hypothesis by showing that McGurk combinations occur more frequently with unvoiced consonants (i.e., /p/ and /k/), in which the burst and aspiration tend to have a higher intensity than in voiced consonants. Although these studies indicated an effect of the release burst and aspiration on the strength of the audiovisual integration for McGurk combinations, the role of these acoustic features on the perceived consonant order has not been established yet. This hypothesis is thus further tested in Chapter 4.

The articulatory constraints imposed by the visual speech gestures might also influence the perceived McGurk combination. For CV combination stimuli, the strong visual bilabial constraint imposed at the initial syllable boundary could affect the perceived consonant order. For example, in the combination illusion of auditory /da/ and visual /ba/, the bilabial visual constraint could impose a strong perceptual evidence for the percept /bda/, since the auditory component could only be naturally produced after the bilabial consonant is released. The strong visual bilabial constraint could also explain the increase in the rate of combination responses when the visual /ba/ is presented ahead of the auditory component, as reported in previous studies (Massaro and Cohen, 1993; Soto-Faraco and Alsius, 2007 ;Soto-Faraco and Alsius, 2009). The articulatory constraints hypothesis is tested and discussed in Chapter 3 in light of the behavioral evidence.

1.4 Timing and audiovisual window of integration

Early studies on audiovisual speech perception noticed that visual information was beneficial for speech perception even when the auditory signal was delayed with respect to the visual speech signal by as much as 400 ms (Dodd, 1977) and that the benefit was stable until the asynchrony exceeded 240 ms (Massaro et al., 1996). Subsequent studies have shown that auditory and visual speech signals are perceived as synchronous over a so-called temporal window of integration, which spans a range of audiovisual asynchronies from about 50 ms audio-leads to 200 ms visual-leads (i.e., Van Wassenhove et al., 2007; Venezia et al., 2016). The temporal window of integration has been tested for congruent speech (for a review, see Venezia et al., 2016), and its width has been found to coincide with the range of audiovisual asynchronies for which the McGurk effect occurs, including fusions (Van Wassenhove et al., 2007) and combinations (Massaro and Cohen, 1993; Soto-Faraco and Alsius, 2009). It has been suggested that the reason for the asymmetry of the audiovisual temporal window of integration is that in natural speech the visual signal tends to lead the acoustic signal (Chandrasekaran et al., 2009). However, more recently, it has been shown that in connected speech the auditory and visual signals are more or less synchronous, varying within a range that includes both auditory leads and visual leads (Schwartz and Savariaux, 2014). If the natural timing of audiovisual speech determines the characteristic asymmetry of the audiovisual temporal window of integration, it would then be expected that in some cases the window would exhibit the opposite asymmetry. This interesting aspect of audiovisual speech perception is further discussed in Chapter 3.

1.5 Overview of the thesis

The purpose of this thesis was to expand the knowledge about how humans combine auditory and visual phonetic information, with a particular focus on presenting evidence for a feature-based integration of speech. The work presented in this thesis provides behavioral evidence as well as novel audiovisual stimuli that challenge the current understanding of audiovisual speech perception. The role of timing of the speech features on audiovisual integration, as well as the perceived consonant order in the understudied McGurk combination effect, were the key aspects addressed here. The work consists of three studies that constitute the three main chapters of the thesis.

Chapter 2 provides the first step towards the development of a feature-based theory of audiovisual speech perception by investigating the audiovisual integration of consonant segments. The study tested a sequential cues hypothesis, which assumes that the audiovisual features of the initial and subsequent segments of consonants are integrated separately. The audiovisual stimuli were then created by pairing auditory and visual utterances of single consonant and consonant cluster, and the tested audiovisual pairings were either phonetically congruent or incongruent (McGurk). A key motivation for studying the audiovisual integration of consonant clusters is that these type of percepts appear in response to McGurk combinations, even though only one consonant is presented acoustically. Understanding how consonant clusters are integrated audiovisually could thus lead to a better comprehension of the perception of McGurk combinations. Specifically, the study sought to determine: (1) whether the auditory perception of acoustic consonant clusters could be influenced by visual articulatory gestures of single consonants and consonant clusters; and (2) whether visual articulatory gestures of consonant clusters could influence the auditory perception of consonant segments. The analysis then focused on whether the response to each stimulus pairings reflected either visual enhancement or visual interference (McGurk effect), and which consonant segments (the initial or subsequent) would be affected.

Chapter 3 examines the integration of audiovisual speech features by presenting a thorough investigation of the perception of McGurk combinations. To determine whether the perceived consonant order in this variant of the McGurk effect could be related to the temporal differences between the auditory and visual consonants, the effect of cross-modal timing of McGurk combination stimuli was varied in two ways. First, the cross-modal timing of the consonants was altered while the vowels were kept synchronous. This was done by pairing auditory and visual utterances recorded with different vowel-consonant-vowel (VCV) articulations, in which the consonant was pronounced either during the syllable onset (V_CV) or the offset (VC_V). Second, the cross-modal timing of both vowels and consonants was altered by varying the audiovisual SOA, as has been done in previous studies (Massaro and Cohen, 1993; Soto-Faraco and Alsius, 2007; Soto-Faraco and Alsius, 2009). Besides the timing hypothesis, the similarity and articulatory constraints hypotheses were tested to determine whether they could account for the obtained cluster percepts. This was done by further investigating consonant-vowel (CV) and vowel-consonant (VC) McGurk combination stimuli. The analysis was then focused on the perceived consonant order, the strength of the audiovisual integration as well as the asymmetry of the audiovisual temporal window of integration.

Chapter 4 investigates whether the perceived consonant order of McGurk combinations could be explained in terms of the temporal differences between the audiovisual speech features, which would provide evidence for the feature-based integration of speech. The results of Green and Norrix (1997), which showed that removing the acoustic release burst and aspiration decreased the proportion of combination illusions served as a basis for this investigation. Specifically, this study evaluates whether varying the timing of the release burst and aspiration relative to the mouth closing gestures could change the perceived consonant order of McGurk combinations. The stimuli consisted of an acoustic continuum created with the articulation of /i_i/ and the added burst (and aspiration) of another naturally recorded consonant /k/, which were aligned at different timings. Two additional continua were produced by presenting the acoustic continuum with the articulatory gestures for /ip_i/ or /i_pi/. The results were analyzed in terms of the perceived consonant order and the strength of the audiovisual integration.

Finally, *Chapter 5* summarizes the main findings, discusses their implications, and provide perspectives on future work on audiovisual speech perception.

2

Audiovisual integration of single consonants and consonant clusters[a]

Abstract

The visual articulatory gestures of a speaker aid speech understanding, particularly in noisy listening situations. They can also change the auditory phonetic percept when paired with discrepant acoustic speech tokens, eliciting illusory percepts as demonstrated by the McGurk effect. In the McGurk fusion illusion, the illusory percept consists of a consonant that is different from the presented auditory and visual consonants. In the McGurk combination illusion, the auditory percept contains both consonants. Despite the substantial research on the McGurk effect, it is still unclear why some audiovisual speech stimuli produce single consonant percepts, whereas others produce consonant clusters. Here, we investigated the audiovisual integration of single consonants and consonant clusters for the congruent and incongruent audiovisual pairings of the disyllables /aba/, /aga/, /ada/, /abga/ and /abda/. The audiovisual perception of these stimuli was tested in the framework of a *sequential speech cues* hypothesis, which assumes that the initial and subsequent audiovisual features of consonant segments are integrated separately. We found novel illusory percepts, such as partial fusions and visual dominance illusions. For example, dubbing auditory /abga/ onto visual /aga/ produced mostly /adga/ percepts, which indicated a partial fusion of the initial acoustic consonant with the initial visual gesture, whereas the subsequent consonant was unaffected. Furthermore, dubbing auditory /aba/ onto visual /abga/ was perceived as /abda/, which reflected a partial fusion illusion between the subsequent consonant segments. Thus, the

---

[a] This chapter is based on Gil-Carvajal and Andersen (2020a).

McGurk effect occurred in the initial or subsequent segments of the stimuli, which supported the existence of sequential audiovisual features in consonant segments that are integrated separately.

Keywords: Consonant clusters, speech perception, audiovisual integration, McGurk effect.

2.1 Introduction

Visual articulatory gestures influence what we hear. They can be beneficial for speech comprehension in face-to-face communication, particularly when the acoustic speech signal is noisy (Sumby and Pollack, 1954; Binnie et al., 1974). Visual articulatory gestures can also affect auditory perception if discrepant visual and auditory utterances are reproduced simultaneously, which produces illusory percepts as demonstrated by the McGurk effect (McGurk and MacDonald, 1976; Van Wassenhove et al., 2007). In the McGurk fusion illusion, a bilabial auditory consonant (i.e., /aba/) dubbed onto a non-labial visual consonant (i.e., /aga/) appears to perceptually fuse into another single consonant (i.e., /ada/). Interestingly, in the McGurk combination illusion, the conversed audiovisual pairing of a non-labial auditory consonant and a bilabial visual consonant produces a cluster percept of the two consonants (i.e., /abga/). Sometimes the McGurk effect takes the form of visually-driven responses, or visual "dominance" illusions, when the consonant heard corresponds to the visually presented consonant (Rosenblum and Saldaña, 1992; Tiippana, 2014). Although the McGurk effect has been extensively studied to explore the mechanisms underlying the integration of audiovisual speech (Alsius et al., 2017), the fundamental question of why some stimuli produce single consonant percepts whereas others elicit consonant clusters has remained unclear.

The single consonant percept in fusion illusions has been explained as the merging of the place information in the auditory and visual speech signals (Green and Norrix, 1997; Colin et al., 2002; Van Wassenhove, 2013; Tiippana, 2014). In the case of combination illusions, the two consonants are perceived due to their large perceptual saliency. While the bilabial consonant is visually prominent, the non-labial consonant (typically /d/ or /g/) contains a strong release burst that is not perceptually fused with the visual gesture (Colin et al., 2002). After the first description of the McGurk combination (McGurk and MacDonald, 1976), several studies have replicated the combination illusion (i.e., Massaro and Cohen, 1993; Walker et al., 1995; Hampson et al., 2003; Soto-Faraco and Alsius, 2009; Jiang and Bernstein, 2011). However, it is still unclear how the audiovisual speech features are integrated to produce the characteristic consonant cluster percept elicited by combination illusions. Understanding how consonant clusters are integrated audiovisually might lead to a better comprehension of the audiovisual integration in the McGurk effect.

When consonants are uttered in the medial position, such as in a vowel-consonant-vowel (VCV) stimulus, a closing gesture typically links the articulation of the initial vowel with the oral constriction as preparation for the subsequent consonant release. Following the consonant release, an opening gesture leads to the articulation of the adjacent final vowel. In the acoustic domain, the place information is signaled by the closing and opening formant transitions that accompany the visual articulatory gestures along with the release burst (Dorman and Raphael, 1980). The idea that the speech features occurring in the initial consonant segment are integrated separately from the audiovisual speech features in the subsequent consonant segment, is here referred to as the *sequential speech cues* hypothesis.

Consonant clusters exist in many languages (Saporta, 1955; Greenberg, 1965; McLeod et al., 2001). They appear in syllables at the initial or the final position, and sometimes at the medial position as they can be formed with adjacent consonants of two different syllables or words (Hardcastle and Roach, 1979; Byrd, 1996). Previous research has shown that when the consonants are uttered in sequences, their articulations are less accurate, i.e., they are often of shorter duration and smaller magnitude than when they are produced in isolation (Kerswill, 1985; Nolan, 1992; Byrd, 1996). Also, due to gestural overlap, the articulation of one consonant could interfere with the acoustic perception of the other (Byrd, 1992), although the visual articulatory gesture would remain present (Browman and Goldstein, 1990; Nolan, 1992). Consonant clusters could, therefore, be particularly susceptible to visual influence due to their perceptual auditory ambiguity.

A consonant geminate is a special case of sequence formed with two identical consonants (Lahiri and Hankamer, 1988; Pickett et al., 1999). Arai et al. (2017) showed that Japanese subjects were able to recognize unimodally the geminates (i.e., /atta/) and their corresponding single consonant stimuli (i.e., /a_ta/, where the underscore represents a silent interval). Interestingly, an incongruent pairing of an auditory consonant with its corresponding visual geminate gesture restored the auditory perception of gemination. This finding was also reported by Scott and Idrissi (2018) for Arabic native speakers. In both experiments, the effect was attributed to the fact that the subjects detected the longer duration of the visual closing gesture, which is crucial for the perception of consonant geminates (Lahiri and Hankamer, 1988; Pickett et al., 1999; Arai et al., 2017; Scott and Idrissi, 2018). The results of these studies suggested a visual influence of consonant geminates on auditory speech perception. It remained unclear, however, how the visual articulatory gestures of sequences of different consonants could influence auditory speech perception.

In this study, we investigated the integration of audiovisual single consonants and consonant clusters for the phonetically congruent and incongruent audiovisual pairings of the dysillables /aba/, /aga/, /ada/, /abga/, and /abda/. According to the *sequential speech cues* hypothesis, the initial speech features should be integrated separately from the subsequent speech features of audiovisual consonant segments. Our investigation had two main purposes. The first was to determine whether the perception of acoustic consonant clusters can be influenced by visual articulatory gestures. We explored what type of percepts are elicited in the case of audiovisual integration, and whether the visual influence would be reflected in the initial bilabial consonant, the subsequent non-labial consonant, or both. The second purpose was to investigate whether visual consonant clusters affect auditory speech perception, and if so, which segment of the acoustic consonant would be influenced, the initial consonant segment, the subsequent segment, or both. The audiovisual perception of the stimuli was then expected to reflect cross-modal interactions according to the time course of the audiovisual speech features.

2.2 Methods

2.2.1 Stimuli

The utterances /aba/, /aga/, /ada/, /abga/ and /abda/ were naturally produced by a female native speaker of Swedish. The consonants /b/, /g/ and /d/ were chosen because they have been shown to elicit strong McGurk fusions (Van Wassenhove et al., 2007; Schwartz, 2010) and combinations (Massaro and Cohen, 1993; Green and Norrix, 1997). Auditory feedback was provided to the speaker via earphones during the recordings. For this, a metronome was set at 170 beats per minute. The beats of the metronome lasted 40 ms and were reproduced every 353 ms. The purpose was to keep a similar duration and speaking rate across utterances, which has been suggested to induce stronger McGurk effects (Munhall et al., 1996).

The video recordings were made with a video camera Canon EOS Rebel T5i at a resolution of 1920 x 1080 and 30 frames per second. The sound was captured with a B&K condenser microphone at a bit depth of 24 bits and a sampling rate of 48 kHz. The recordings took place in a sound-proof booth with a black background behind the speaker. Offline time-alignment of the acoustic and video signals was made in the software Adobe Premiere Pro to create the audiovisual stimuli. All visual and audiovisual stimuli started with a still face of the speaker with a closed mouth, lasting at least ten frames. The duration of all stimuli was 1.7 s.

2.2.2 Subjects

Twenty subjects (19 – 29 years of age, mean age 24) participated in the experiment. They all reported to have normal hearing and normal or corrected-to-normal vision. All experiments were carried out in the same sound-proof booth in which the recordings were made, and were approved by the Science-Ethics Committee for the Capital Region of Denmark (reference H-16036391). All subjects provided written consent and were financially compensated.

2.2.3 Experimental procedure

The stimuli were displayed on a Dell 27-inch color screen placed at 75 cm in front of the subject. The sound was reproduced at 65 dBA through a loudspeaker Genelec 8020C situated on a desk right below the screen. Three types of stimuli were presented, audiovisual (congruent or McGurk), visual only, and auditory only. The subjects were asked to choose the consonant, or the consonant sequence they heard in each trial. In the case of visual trials, the subjects were asked to perform speechreading. Nine possible response options were provided, which were labeled on a computer keyboard in a 3x3 arrangement. The single consonants /b/, /g/, and /d/ appeared in the first row, followed by the cluster responses /bg/ /gd/ and /db/ in the second row, and the corresponding reverse cluster responses in the third row /gb/ /dg/ and /bd/.

The experiment was divided into two blocks of ten trials each. In each trial, all audiovisual, visual, and auditory stimuli appeared, and the order of presentation was randomized. Before the experiment, the subjects received written instructions. The experiment started with a training test that lasted one trial, which served to familiarize the subject with the task. After the first experimental block, the subjects took a five-minute break. The total duration of the experiment was about 60 minutes.

2.2.4 Data analysis

To determine how acoustic consonant clusters could be affected by visual articulatory gestures, we analyzed the responses to audiovisual stimuli that contained a bilabial visual consonant. These stimuli were expected to induce a strong visual effect due to the high perceptual saliency of their visual bilabial component. We also examined the responses to audiovisual stimuli that contained non-labial visual consonants, which despite being less visually salient than their bilabial counterpart (Colin et al., 2002), when paired with incongruent single acoustic consonants can elicit McGurk fusions (McGurk and MacDonald, 1976; Van Wassenhove et al., 2007; Tiippana, 2014). Furthermore, to determine whether visual consonant clusters influence auditory speech perception, we examined the responses to audiovisual stimuli that contained either an acoustic consonant cluster or a single acoustic consonant. The statistical comparisons between the mean response proportions from the audiovisual stimuli and each corresponding auditory stimulus were performed with one-sided paired t-tests, unless stated otherwise. A significance level of $\alpha = 0.05$ was considered in all statistical analyses. The statistical overview of comparisons is presented in Tables 2.1, 2.2, and 2.3 of the supplementary material.

2.3 Results

2.3.1 Effect of bilabial visual stimuli on the perception of acoustic consonant clusters

Figure 2.1 shows the results for the audiovisual stimuli that contained an acoustic consonant cluster and a visual bilabial consonant (green bars). The figure also shows the results obtained with the corresponding unimodal auditory (blue) and visual stimuli (red). One-sided paired t-tests revealed a significant enhancement in the perception of the acoustic consonant clusters due to congruent visual speech. For auditory /abga/, the mean proportion correct was 0.88 and increased to 0.95 for congruent audiovisual /abga/ ($p = .008$). For auditory /abda/, the mean proportion correct was 0.86 compared to 0.97 for congruent audiovisual /abda/ ($p < .0001$).

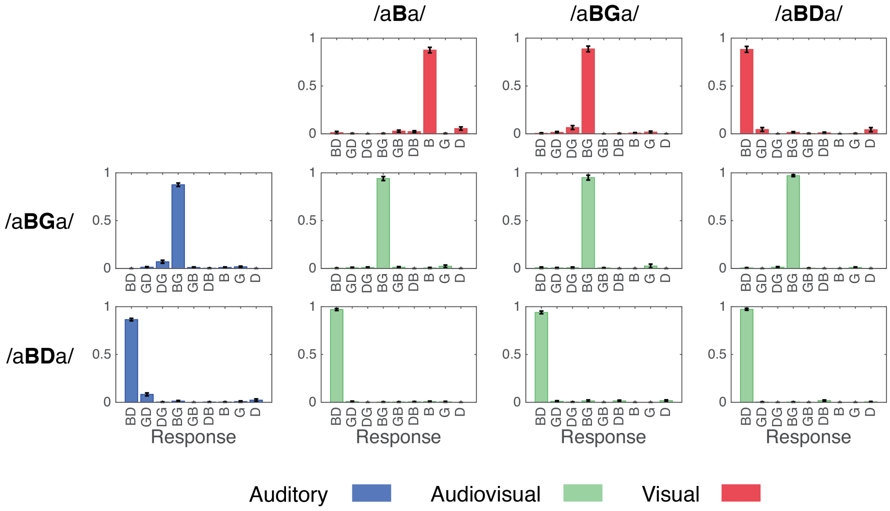


Figure 2.1: Effect of bilabial visual stimuli on the perception of acoustic consonant clusters. Mean response proportions across subjects. The leftmost column and top row represent the auditory (blue) and visual (red) stimuli, respectively. The audiovisual stimuli (green) were produced with all possible pairings, congruent and incongruent, of the unimodal auditory and visual stimuli. Error bars represent the standard error of the mean.

More generally, we found that *all* visual speech tokens that contained a bilabial closing gesture (/aba/, /abga/ and /abda/) facilitated the perception of both acoustic consonant clusters. For auditory /abga/, the proportion correct was 0.97 when dubbed onto visual /abda/, and 0.94 when dubbed onto visual /aba/. Both proportions were significantly higher than for auditory /abga/ alone ($p < .006$). For auditory /abda/, the proportion correct was 0.94 when dubbed onto visual /abga/ and 0.97 when dubbed onto visual /aba/. Both proportions were also significantly higher than for the corresponding auditory stimulus /abda/ alone ($p < .001$). Notably, the facilitation provided by the partly incongruent visual speech stimuli was not significantly different from the facilitation obtained with the congruent visual speech stimuli ($p > .33$, two-sided paired t-tests). Thus, the partly incongruent visual speech tokens that contained a bilabial consonant enhanced the perception of the acoustic consonant clusters to the same extent as the congruent visual speech tokens. This was unexpected since incongruent visual speech have been reported to usually induce illusory percepts rather than an enhancement (McGurk and MacDonald, 1976; Tiippana, 2014; Alsius et al., 2017).

The visual stimuli containing a bilabial consonant were clearly perceived. The proportion correct was 0.87 for visual /aba/, 0.89 for visual /abga/ and 0.88 for visual /abda/. Therefore, the similar facilitation obtained with these three visual stimuli cannot be due to them being indistinguishable. The facilitation provided by the partly incongruent visual stimuli might be due to a strong visual influence of the bilabial consonant and only little or no visual influence of the non-labial consonant. This should be reflected in an increased proportion of responses with a bilabial consonant. Hence, the facilitation effects should remain when examining the perception of the bilabial gesture of the cluster by counting all responses as correct if they contain a bilabial consonant. We tested the facilitation after pooling response across the response categories /b/, /bg/, /gb/, /bd/ and /db/. The facilitation was significant for both consonant clusters and all three visual stimuli containing a bilabial component ($p < .021$). The bilabial visual gesture is, thus, a strong contributor to the facilitation effects for consonant clusters.

2.3.2 Effect of non-labial visual stimuli on the perception of the bilabial acoustic consonant

Figure 2.2 shows the results for the audiovisual stimuli that contained a bilabial acoustic consonant and a non-labial visual consonant (green bars), as well as their corresponding unimodal auditory (blue) and visual stimuli (red). The proportion of /ada/ responses for auditory /aba/ was only 0.04. This proportion significantly increased to 0.59, when auditory /aba/ was dubbed onto visual /aga/ ($p < .0001$). This was expected, as the visual consonant /g/ is known to induce the fusion illusion of changing the percept of auditory /b/ to /d/. The proportion of /ada/ responses also increased to 0.71 when auditory /aba/ was dubbed onto visual /ada/ ($p < .0001$), reflecting a visual dominance illusion.

Similarly, the non-labial visual consonants influenced the perception of acoustic consonant clusters, producing partial fusions and visual dominance illusions. Dubbing auditory /abga/ onto visual /aga/ or /ada/ produced the illusion of hearing /adga/, with a response proportion of 0.65 and 0.69, respectively. This illusory percept indicates a partial fusion or a visual dominance illusion between the initial auditory consonant /b/ and the initial visual gesture for /g/

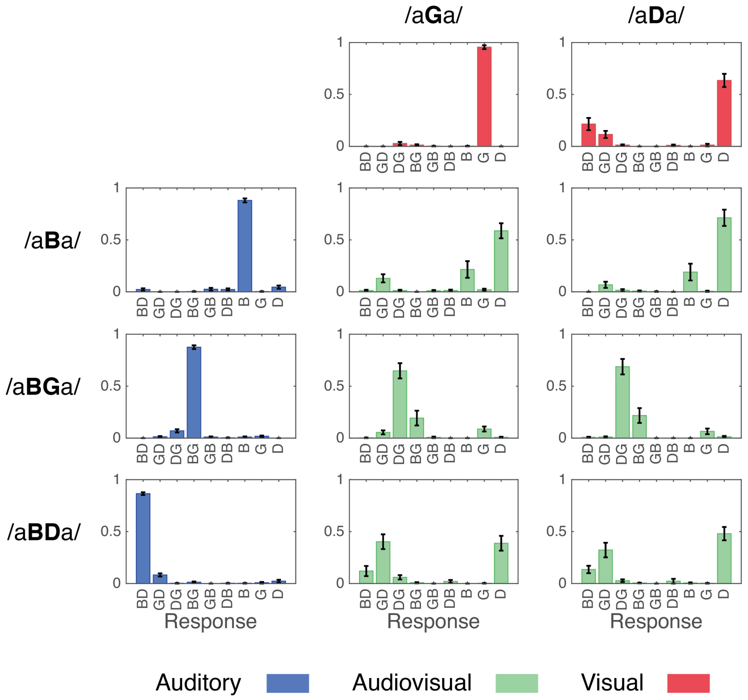


Figure 2.2: Effect of non-labial visual stimuli on the perception of the bilabial acoustic consonant. Mean response proportions across subjects. The leftmost column and top row represent the auditory (blue) and the visual (red) stimuli, respectively. The audiovisual stimuli (green) were produced with all possible pairings of the unimodal auditory and visual stimuli. Error bars represent the standard error of the mean.

or /d/, respectively, while the subsequent auditory consonant /g/ remained unchanged. In both cases, the response proportions were significantly higher than those obtained for auditory /abga/ alone ($p < .0001$). For auditory /abda/, the proportion of illusory /agda/ responses increased to 0.40 when dubbed onto visual /aga/ and to 0.32 when dubbed onto visual /ada/ ($p < .0001$). Likewise, the proportion of /ada/ responses for auditory /abda/ increased to 0.39 when dubbed onto visual /aga/ and to 0.48 when dubbed onto visual /ada/ ($p < .0001$). Thus, these results indicate that the initial visual gesture of the non-labial consonants influenced the initial consonant of the acoustic cluster, which produced partial fusions or visual dominance illusions. In contrast, the subsequent acoustic consonant remained unchanged.

To determine whether the fusions and visual dominance illusions occurred in the initial segment, we pooled across all responses that contained a non-labial consonant in the initial segment of the articulation (/db/, /dg/, /gb/, /gd/, /d/ and /g/). These response proportions were higher for both acoustic consonant clusters when they were dubbed onto the non-labial visual consonants than for their corresponding auditory stimuli alone ($p < .0001$). Furthermore, we analyzed any changes on the second consonant of the cluster due to the non-labial visual stimuli by pooling across responses that contained the auditory consonant in the subsequent segment of the articulation. The analysis revealed that the non-labial visual gesture did not significantly change the response proportions containing /g/ or /d/ in the subsequent consonant of the acoustic clusters ($p > .21$, two-sided paired *t*-test), except for auditory /abda/ dubbed onto visual /aga/ ($p = .01$). This indicates that the non-labial visual consonants affected the perception of the initial bilabial acoustic consonant of the cluster, while the subsequent non-labial acoustic consonant remained mostly unaffected.

2.3.3 Effect of bilabial visual stimuli on the perception of acoustic non-labial consonants

Figure 2.3 shows the results for the audiovisual stimuli that contained a single acoustic consonant and a bilabial visual consonant (green bars), as well as the corresponding unimodal auditory (blue) and visual stimuli (red). The visual consonant clusters induced partial fusions and visual dominance illusions in the acoustic stimulus with a single bilabial consonant. This was reflected in the mean proportion of illusory /abda/ responses, which significantly increased to 0.50 when auditory /aba/ was dubbed onto visual /abga/ ($p < .0001$) and to 0.59 when auditory /aba/ was dubbed onto visual /abda/ ($p < .0001$). The effect cannot be attributed to perceptual ambiguity since the mean proportion of confusions of auditory /aba/ with /abda/ was negligible. In comparison, visual /aba/, as expected, did not significantly increase the proportion of /abda/ confusions when paired with auditory /aba/ ($p = .96$). This suggests that the illusory percepts were due to the non-labial gesture of the visual consonant clusters, which influenced the subsequent acoustic consonant segment while the initial acoustic segment remained unchanged.

Visual consonant clusters also elicited combination illusions, which were similar to those induced by the visual stimulus with a bilabial single consonant. Dubbing auditory /aga/ onto a visual consonant cluster increased the proportion of /abga/ responses to 0.71 and to 0.69 for /abga/ and /abda/, respectively ($p < .0001$). For auditory /ada/, the proportion of illusory /abda/ responses increased to 0.83 and 0.88 when dubbed onto visual /abga/ and /abda/, respec-

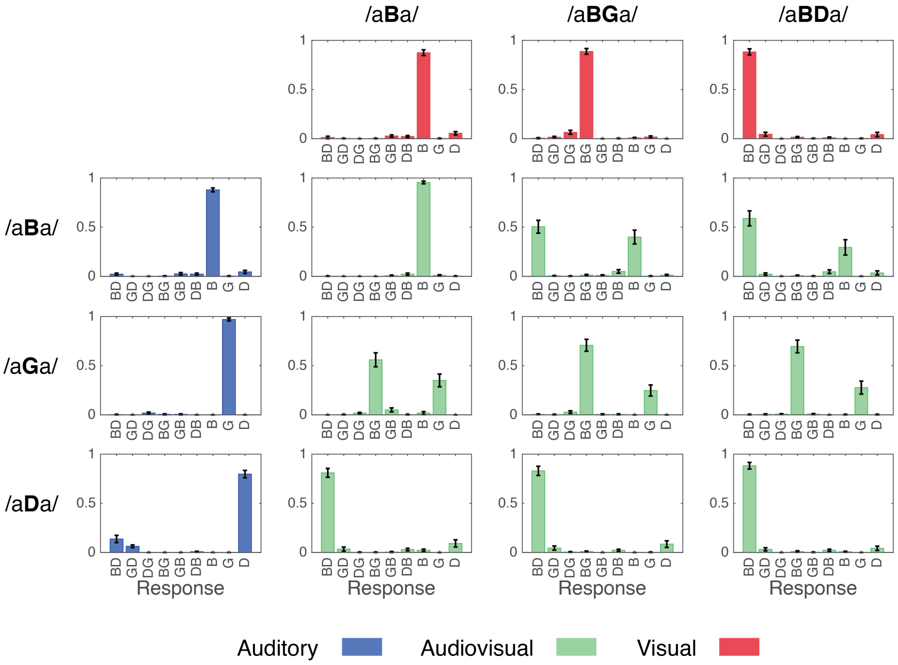


Figure 2.3: Effect of bilabial visual stimuli on the perception of acoustic non-labial consonants. Mean response proportions across subjects. The leftmost column and top row represent the auditory (blue) and visual (red) stimuli, respectively. The audiovisual stimuli (green) were produced with all possible pairings of the unimodal auditory and visual stimuli. Error bars represent the standard error of the mean.

tively, ($p < .0001$). Similarly, in the case of single consonant stimuli, adding the bilabial visual component /aba/ increased the proportion of /abga/ responses to 0.56 for auditory /aga/, as well as the proportion of /abda/ responses to 0.81 for auditory /ada/ ($p < .0001$). This indicates that the initial bilabial gesture influenced the initial segment of the non-labial acoustic consonant, whereas the subsequent acoustic segment was unchanged.

To assess the audiovisual integration in the initial consonant segment, we pooled across all responses containing the bilabial consonant in the initial segment of the articulation (/b/, /bg/ and /bd/). For auditory /aga/ and /ada/, the response proportions were significantly higher when dubbed onto the bilabial visual stimuli compared to the corresponding auditory stimuli alone ($p < .0001$). Furthermore, to determine whether the subsequent non-labial visual consonant also influenced the perception of the acoustic consonants, we analyzed the pooled response proportions that contained a non-labial consonant in the subsequent segment of the articulation (/bd/, /gd/, /dg/, /bg/, /g/, /d/). For auditory /aba/ dubbed onto visual /abga/ or /abda/, the analysis revealed an increased proportion of responses with non-labial consonants in the subsequent segment of the articulation ($p < .0001$). This is consistent with the obtained partial fusions and visual dominance illusions in the subsequent consonant segment. In contrast, for auditory /aga/ and /ada/, the visual stimuli did not increase the proportion of non-labial responses in the subsequent consonant segment ($p > .83$). This suggests that the combination responses were produced by the visual influence of the bilabial consonant on the initial segment of the non-labial acoustic consonants, while the subsequent segment was unaffected.

2.4 Discussion

The results of the current study support the *sequential speech cues* hypothesis, which assumes that the audiovisual speech features in the initial and subsequent segments of single consonants and consonant clusters integrate separately. Such audiovisual speech processing seems to be reflected in either an audiovisual enhancement of the bilabial consonant or illusory percepts. The latter took the form of combinations, novel partial fusions and visual dominance illusions. Moreover, the above results demonstrated that the visual articulatory gestures for each of the consonants in the cluster affect speech perception in the initial and subsequent segments of consonants.

Our findings show an audiovisual enhancement even in the case of partly incongruent audiovisual speech tokens, and not only in the case of congruent audiovisual speech as has been reported in earlier studies (i.e., Sumby and Pollack, 1954; MacLeod and Summerfield, 1987; Arnold and Hill, 2001). A likely explanation of this effect is that the initial visual bilabial component enhanced the initial bilabial consonant of the acoustic cluster, whereas the subsequent acoustic consonant was unaffected by the speech gestures. This is reasonable considering that the subsequent acoustic consonant was either a velar or alveolar consonant, and hence, contained a perceptually stronger acoustic cue (release burst) than the initial acoustic bilabial consonant (Green and Norrix, 1997; Colin et al., 2002).

Our results also show novel McGurk partial fusions as well as visual dominance illusions in the initial segment of the stimulus, as shown in Figure 2.4. Unlike the facilitation observed when dubbing visual bilabial gestures onto acoustic consonant clusters, the non-labial visual gestures interfered with the perception of the bilabial acoustic component of the cluster. For instance, the audiovisual pairing of auditory /abga/ and visual /aga/ produced mostly /dg/ responses (Figure 2.4), which indicates a partial fusion illusion between the cross-modal speech features in the initial gesture of the utterance, while the second acoustic consonant remained unchanged by the subsequent visual gesture. This further supports the finding that the bilabial component of the acoustic consonant clusters is prone to visual influence, as shown by the audiovisual enhancement.

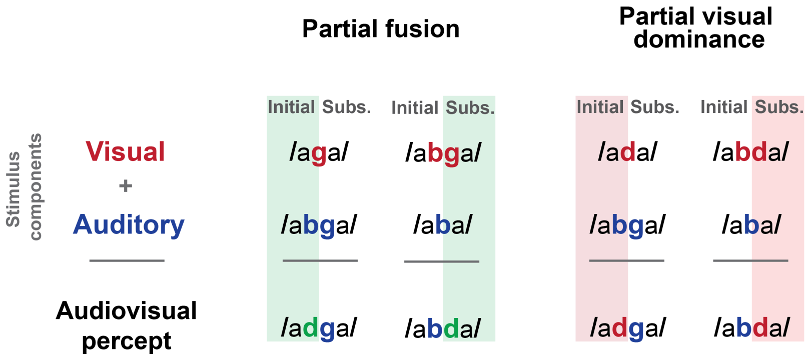


Figure 2.4: Examples of partial illusory percepts. Partial fusions occurred in the initial or subsequent consonant segments (highlighted in green), while the other segment was unchanged. Partial visual dominance illusions also occurred in the initial or subsequent consonant segments (highlighted in red), while the other segment was unchanged.

Visual consonant clusters also affected speech perception. This is in agreement with the results of Arai et al. (2017) as well as Scott and Idrissi (2018), which showed a visual influence of the closing gesture of geminate consonants on auditory perception. In our results, the visual influence of the initial consonant of the cluster on the closing stage of the auditory non-labial consonant produced mostly McGurk combinations. In contrast, the visual influence of the second consonant of the cluster resulted in mostly partial fusions or visual dominance illusions occurring in the subsequent consonant segment, as shown in Figure 2.4. This is the case for the partly fused percept /bd/ that occurred in response to the audiovisual pairing of auditory /aba/ and visual /abga/. Thus, while previous studies reported a visual influence in the closing gesture of the consonant sequence, here we show that such visual effect can occur for both consonants, in the initial closing and the subsequent mouth opening gesture.

Importantly, the time course of the audiovisual speech features seems to underlie the integration in the McGurk effect. A McGurk fusion illusion appears when the audiovisual speech features in the initial consonant segment and the speech features in the subsequent consonant segment are integrated into one and the same consonant. In contrast, a McGurk combination illusion occurs if the earlier audiovisual speech features are integrated into one consonant and the subsequent speech features integrate into another consonant. The same concept would account for the illusory percepts occurring for McGurk vowel-consonant and consonant-vowel stimuli reported in previous studies (Massaro and Cohen, 1993; Hampson et al., 2003; Van Wassenhove et al., 2007), for which the initial and subsequent speech features might only occur in one stage. The above results highlight the need for control of the timing of the audiovisual speech stimuli to ensure that the acoustic speech features are integrated with the desired visual gestures. The present study serves as a framework for an audiovisual speech model, in which the cross-modal speech features of the initial consonant segments are predicted separately from the speech features of the subsequent segments.

Acknowledgments

This research was supported by the Oticon Centre of Excellence for Hearing and Speech Sciences (CHeSS) at the Technical University of Denmark (DTU).

Supplementary material

Table 2.1: Summary results of the paired t-test performed to determine the effect of bilabial visual stimuli on the perception of acoustic consonant clusters.

Contrast			**On /BG/ responses**	
Auditory	vs	**Audiovisual (A+V)**	**$t(19)$**	***p***
/BG/		/BG/ + /B/	-2.830	.0053
/BG/		/BG/ + /BG/	-2.615	.0085
/BG/		/BG/ + /BD/	-4.711	<.0001
			On /BD/ responses	
			$t(19)$	***p***
/BD/		/BD/ + /B/	-6.489	<.0001
/BD/		/BD/ + /BD/	-9.245	<.0001
/BD/		/BD/ + /BG/	-4.567	.0001
Contrast[*]			**On /BG/ responses**	
Audiovisual (A+V)	vs	**Audiovisual (A+V)**	**$t(19)$**	***p***
/BG/+/BG/		/BG/ + /B/	0.809	.4283
/BG/+/BG/		/BG/ + /BD/	-1.017	.3220
/BD/+/BD/		/BD/ + /B/	0.237	.5923
/BD/+/BD/		/BD/ + /BG/	2.942	.9958
Contrast			**Over pooled responses with a bilabial consonant**	
Auditory	vs	**Audiovisual (A+V)**	**$t(19)$**	***p***
/BG/		/BG/ + /B/	-3.005	.0036
/BG/		/BG/ + /BG/	-2.188	.0207
/BG/		/BG/ + /BD/	-4.213	.0002
/BD/		/BD/ + /B/	-5.385	<.0001
/BD/		/BD/ + /BD/	-7.443	<.0001
/BD/		/BD/ + /BG/	-4.341	.0001

[*] Two-sided t-test

Table 2.2: Summary results of the paired *t*-test performed to determine the effect of non-labial visual stimuli on the perception of the bilabial acoustic consonant.

Contrast			**On /D/ responses**	
Auditory	vs	**Audiovisual (A+V)**	***t*(19)**	***p***
/B/		/B/ + /G/	-2.830	<.0001
/B/		/B/ + /D/	-9.141	<.0001
			On /DG/ responses	
			***t*(19)**	***p***
/BG/		/BG/ + /G/	-8.283	<.0001
/BG/		/BG/ + /D/	-8.741	<.0001
			On /GD/ responses	
			***t*(19)**	**p**
/BD/		/BD/ + /G/	-4.297	<.0001
/BD/		/BD/ + /D/	-3.302	.0019
			On /D/ responses	
			***t*(19)**	***p***
/BD/		/BD/ + /G/	-5.536	<.0001
/BD/		/BD/ + /D/	-7.721	<.0001
			Over pooled responses with a non-labial consonant in the initial segment	
			***t*(19)**	***p***
/BG/		/BG/ + /G/	-10.186	<.0001
/BG/		/BG/ + /D/	-9.614	<.0001
/BD/		/BD/ + /G/	-14.576	<.0001
/BD/		/BD/ + /D/	-16.834	<.0001
			Over pooled responses containing the acoustic consonant in the subsequent segment[*]	
			***t*(19)**	***p***
/BG/		/BG/ + /G/	1.288	.2131
/BG/		/BG/ + /D/	-0.484	.6342
/BD/		/BD/ + /G/	2.854	.0102
/BD/		/BD/ + /D/	1.084	.2918

[*] Two-sided *t*-test

Table 2.3: Summary results of the paired *t*-test performed to determine the effect of the visual stimuli on the perception of acoustic non-labial consonants.

Contrast			**On /BD/ responses**	
Auditory	vs	**Audiovisual (A+V)**	***t*(19)**	***p***
/B/		/B/ + /B/	1.798	.956
/B/		/B/ + /BG/	-7.676	<.0001
/B/		/B/ + /BD/	-7.696	<.0001
			On /B/ responses	
			***t*(19)**	***p***
/B/		/B/ + /B/	-4.4591	.0001
			On /BG/ responses	
			***t*(19)**	***p***
/G/		/G/ + /B/	-7.690	<.0001
/G/		/G/ + /BG/	-11.172	<.0001
/G/		/G/ + /BD/	-10.430	<.0001
			On /BD/ responses	
			***t*(19)**	***p***
/D/		/D/ + /B/	-11.509	<.0001
/D/		/D/ + /BG/	-11.444	<.0001
/D/		/D/ + /BD/	-14.983	<.0001
			Over pooled responses with a bilabial consonant in the initial segment	
			***t*(19)**	***p***
/B/		/B/ + /B/	-2.761	.0124
/B/		/B/ + /BG/	-0.376	.7109
/B/		/B/ + /BD/	0.374	.7125
/G/		/G/ + /B/	-8.067	<.0001
/G/		/G/ + /BG/	-11.435	<.0001
/G/		/G/ + /BD/	-10.505	<.0001
/D/		/D/ + /B/	-12.140	<.0001
/D/		/D/ + /BG/	-11.864	<.0001
/D/		/D/ + /BD/	-15.936	<.0001
			Over pooled responses with a non-labial consonant in the subsequent segment	
			***t*(19)**	***p***
/B/		/B/ + /B/	3.093	.9970
/B/		/B/ + /BG/	-7.124	<.0001
/B/		/B/ + /BD/	-7.697	<.0001
/G/		/G/ + /B/	3.115	.9971
/G/		/G/ + /BG/	1.000	.8351
/G/		/G/ + /BD/	0.809	.7859
/D/		/D/ + /B/	3.162	.9974
/D/		/D/ + /BG/	1.371	.9068
/D/		/D/ + /BD/	1.949	.9669

3

Order matters: timing and syllabic context influence perceived consonant order in audiovisual speech[a]

Abstract

Visibility of the talker's mouth movements facilitates speech perception in face-to-face communication. The McGurk illusion convincingly demonstrates that speech perception involves the integration of audiovisual phonetic information for speech comprehension. Despite decades of research, it is still unclear how phonetic information is perceptually integrated. Here, we study the case of the McGurk combination illusion, in which the perception of an auditory non-labial consonant, i.e., /aga/ dubbed onto a visible labial consonant, i.e., /aba/ is perceived as a sequence of the two consonants presented, i.e., /abga/. In most studies, the visual labial consonant is heard as the leading consonant; while the reverse perceived consonant order tends to occur sporadically. A few studies have suggested that the perceived consonant order varies with the audiovisual stimulus onset asynchrony (SOA), but the results have been inconsistent. Here, we varied either the SOA, as has been done in previous studies, or the audiovisual "internal" timing of the stimuli for which we used two different syllabic contexts. With the latter approach, the consonants were made cross-modally asynchronous while the vowels were kept synchronous. The results showed that the syllabic context strongly influenced the perceived consonant order, whereas audiovisual SOA mostly affected the strength of the illusion. We also found that the asymmetry of the temporal window of integration depended on the syllabic context. We propose that, in

---

[a] This chapter is based on Gil-Carvajal et al. (2020b).

addition to timing, articulatory constraints influenced the perceive consonant order and hence, audiovisual integration.

Keywords: Audiovisual speech perception, cross-modal synchrony, syllabic context, articulatory constraints.

3.1 Introduction

Combining visual cues from a talker's mouth with auditory cues from some talker's voice facilitates speech perception (Sumby and Pollack, 1954; MacLeod and Summerfield, 1987; Arnold and Hill, 2001). This process of audiovisual integration occurs in natural face-to-face communication and is particularly useful in complex settings, such as a conversation in a noisy environment (Binnie et al., 1974).

The phenomenon of audiovisual integration has been the topic of various investigations. In particular, the McGurk illusion has proven a useful tool to study audiovisual speech integration. It is produced by dubbing phonetically incongruent acoustic speech onto a video of an articulating face (McGurk and MacDonald, 1976; Alsius et al., 2017). The illusion comes in several variants. The McGurk fusion illusion can be created by dubbing the nonsense speech sound /aba/ onto articulatory gestures for /aga/ leading to the audiovisually fused percept /ada/ (McGurk and MacDonald, 1976; Van Wassenhove et al., 2007). With the converse pairing of auditory /aga/ with visual /aba/, the illusion takes the form of a cluster percept of either /abga/ or /agba/. These types of cluster responses are called McGurk combinations (McGurk and MacDonald, 1976; Green and Norrix, 1997; Tiippana, 2014) since they contain the two consonants presented. Despite the substantial research on the McGurk illusion, little is still known about how speech features are integrated across modalities to create illusory percepts in these two variants of the illusion. The current study investigated factors influencing the combination illusion in an attempt to improve our understanding of audiovisual integration of speech.

Several hypotheses have been proposed to explain why certain audiovisual stimuli produce fusion illusions, whereas others produce combination illusions. The *manner-place* hypothesis proposed by MacDonald and McGurk assumes that the place of articulation is the most salient visual phonetic feature (MacDonald and McGurk, 1978), while the manner of articulation provides a stronger acoustic cue (Miller and Nicely, 1955; Binnie et al., 1974). Consequently, the fused percept has been argued to be a “best-fit solution” to the place features of the visual information and the manner features of the auditory signal. In contrast, the combination illusion occurs if the phonetic incongruence between the auditory and the visual information is too large such that no fused percept fits well (McGurk and MacDonald, 1976).

An alternative explanation proposed by Green and Norrix is here referred to as the *burst and aspiration* hypothesis (Green and Norrix, 1997). They suggested that combination illusions result, at least partially, from release bursts or aspirations in the acoustic signal that are not affected by audiovisual integration. When removing these stimulus features, Green and Norrix demonstrated that the occurrence of cluster responses decreased while fusion responses were much less affected (Green and Norrix, 1997). The authors also argued that the bilabial consonants (i.e., /b/) typically used in fusion illusions exhibit less pronounced bursts and aspirations, which could explain why the acoustic consonant is not perceived in fusion illusions as it is in combination illusions. Consistent with this view, Colin and colleagues showed that combinations are more frequent with voiceless consonants (i.e., /k/), which typically contain stronger release bursts and more aspirations than voiced consonants (Colin et al., 2002).

In many studies, the perceived order of the consonants in the combination illusion was not assessed, and the responses (i.e., /abga/ and /agba/) were pooled into only one response category (Green and Norrix, 1997; Colin et al., 2002; Aruffo and Shore, 2012; Baart et al., 2017). In other studies where the order of the consonants was reported explicitly, the visual bilabial consonant was commonly found in the lead position of the consonant cluster (MacDonald and McGurk, 1978; Walker et al., 1995), although not always (Massaro and Cohen, 1993; Soto-Faraco and Alsius, 2007; Soto-Faraco and Alsius, 2009; Jiang and Bernstein, 2011). The reason for more frequent responses with a leading visual consonant could be that visual speech information naturally precedes the auditory signal because of the anticipatory mouth movements (Sams et al., 1991; Schwartz et al., 2004; Van Wassenhove et al., 2005; Venezia et al., 2016). The idea that differences in arrival times between the auditory and the visual component could determine the perceived consonant order in the combination illusion is referred to here as the *timing* hypothesis. This hypothesis was tested by Massaro and Cohen (Massaro and Cohen, 1993), who manipulated the stimulus onset asynchrony (SOA) between auditory /da/ and visual /ba/. Their findings were consistent with the timing hypothesis in that the proportion of /bda/ responses increased, as expected when the SOA was varied from 200 ms auditory-lead to 200 ms visual-lead. However, the reverse consonant order, /dba/, was almost never perceived, even when the stimulus onset for auditory /da/ was ahead of the visual /ba/.

The idea that visual speech information naturally precedes the auditory signal because of the anticipatory mouth movements has been tested experimentally by analyzing natural audiovisual speech (Chandrasekaran et al., 2009; Schwartz and Savariaux, 2014; Venezia et al., 2016). Whereas Chandrasekaran et al. (2009) confirmed this temporal relationship, Schwartz and Savariaux (2014) found that it only held consistently either at the beginning of a speech sequence or in the case of consonant-vowel (CV) stimuli, in which visible preparatory movements precede the consonant sound. For vowel-consonant-vowel (VCV) sequences as well as more complex stimuli, the range of asynchronies includes both auditory-lead and visual-lead dynamics (Schwartz and Savariaux, 2014). Hence, the fact that most studies (MacDonald and McGurk, 1978; Green and Norrix, 1997; Colin et al., 2002; Soto-Faraco and Alsius, 2009; Baart et al., 2017) employed CV sequences to produce the combination stimuli may explain why a perceived visual lead of the consonant order has been most commonly reported. Interestingly, using vowel-consonant (VC) McGurk combinations, Hampson et al. (2003) found several cluster responses containing the auditory non-labial consonant first. This indicates that the syllabic context is important for the perceived order of the consonants.

Perceptual similarity between the unisensory stimuli and the audiovisual speech percept is another factor that may influence the type of illusion that the observer perceives (Massaro and Palmer Jr, 1998). According to Massaro's Fuzzy Logical Model of Perception (FLMP), the combination percept /bda/ arises due to the similarity between the unimodal components (auditory /da/ and visual /ba/, respectively) and the articulatory features of the cluster response. The reverse cluster /dba/, on the other hand, is visually dissimilar with the articulatory gestures for /ba/, which leads to the typical finding of fewer /dba/ percepts than /bda/ percepts in response to this combination (Massaro and Cohen, 1993; Massaro et al., 1996; Soto-Faraco and Alsius, 2009).

Articulatory constraints could also affect the perceived order of consonants. Depending on the syllabic structure of the combination stimulus, the syllable boundaries might impose articulatory constraints on the cluster percept. For instance, in the combination visual /ba/ with auditory /ga/, the velar plosive burst can only be produced after the opening of the mouth, which is clearly seen as a bilabial opening, and /bga/ is therefore the only possible consonant sequence. This constraint could apply even if neither visual /ba/ nor auditory /ga/ were similar to /bga/, and would thus affect multisensory integration directly and not through similarity. Following this reasoning, one would expect the perceived consonant order for VC stimuli to be the reverse of that for CV stimuli, since the terminal bilabial closure would not allow any following consonants. The results from Hampson et al. (2003) partly support this view, as they found this reverse order for VC stimuli to be the most frequent percept but only when a visual lag was added to the stimulus. In contrast, visual VCV stimuli might not impose the same constraints as the mouth is seen opening both before and after the visual consonant.

In the present study, we tested the timing hypothesis by varying audiovisual synchrony in two ways. In Experiment 1, we varied the “internal” audiovisual timing by using two different syllabic contexts. The consonant was either pronounced in the syllable offset (VC_V) or the syllable onset (V_CV) of the VCV sequence. Thus, the consonants were cross-modally asynchronous while the vowels were kept synchronous. In this way, we attempted to minimize the effect of cross-modal asynchrony on the strength of audiovisual integration in order to isolate the effect of cross-modal asynchrony on perceived consonant order. Additionally, we manipulated the audiovisual SOA in order to compare the two approaches of varying cross-modal synchrony. In Experiment 2, we changed the syllabic context to determine whether the results obtained for VCV sequences in Experiment 1 generalize to CV and VC stimuli, which have simpler syllabic structure, and hence, possibly different perceptual constraints. With this approach, we aimed to determine how syllabic context and audiovisual SOA affect the perceived combination illusion.

3.2 Material and methods

3.2.1 Stimuli

For Experiment 1, we recorded the utterances /aga/ and /aba/. These were produced by a male native speaker of Spanish (author JCGC) using two different syllabic contexts. The VC_V was produced by articulating the consonant during the offset (the coda) of the first syllable (Kessler and Treiman, 1997). The V_CV context was formed producing the consonant during the onset of the second syllable. The underscore represents the stop in the articulation of the sequence that was approximately 128 ms for VC_V and 143 ms for V_CV stimuli.

A total of 20 stimuli was used for testing. The congruent audiovisual stimuli corresponded to the four originally recorded utterances, /ag_a/, /a_ga/, /ab_a/ and /a_ba/. Another four stimuli corresponded to the unimodal conditions: Auditory /ag_a/ and /a_ga/ for which the visual information was a still image of the talker with closed mouth, and visual /ab_a/ and /a_ba/, which did not contain sound. The remaining 12 stimuli corresponded to those used to study the effect of the syllabic context (2) and the effect of audiovisual SOAs (10) on VCV combination stimuli.

To investigate the syllabic context, McGurk combination stimuli were created dubbing auditory /ag_a/ onto visual /a_ba/, and auditory /a_ga/ onto visual /ab_a/. Therefore, one combination stimulus contained a leading auditory consonant and the other a leading visual consonant. Both stimuli, however, had cross-modal synchronous vowels. To test the effect of audiovisual SOAs on VCV combination illusions, two additional stimuli were created using the same syllabic context across modalities, auditory /ag_a/ paired with visual /ab_a/, and auditory /a_ga/ with visual /a_ba/. For these two stimuli, audiovisual SOAs were artificially manipulated in order to obtain five different audiovisual SOAs, –200, –100, 0, 100 and 200 ms. The negative sign represents auditory leads, the positive sign indicates visual leads, and 0 ms represents the synchronous condition.

In Experiment 2, a total of 22 stimuli was considered for testing VC and CV sequences. For this purpose, the same talker uttered the syllables /ag/, /ab/, /ga/ and /ba/. The original recordings of these utterances corresponded to the four congruent audiovisual stimuli presented. The four unimodal stimuli consisted of auditory /ag/ and /ga/, and visual /ab/ and /ba/. The remaining 14 stimuli were the combination illusions used for testing the effect of audiovisual asynchrony. Seven stimuli were used to test VC combinations at the asynchronies of –400, –200, –100, 0, 100, 200 and 400 ms. Another seven stimuli were used to investigate CV combinations for the same SOAs.

In both experiments, a metronome was used as a reference for timing during all recordings to keep a similar pace and duration across the speech tokens. The tempo was set at 170 beats per minute. This was the speaking rate at which the talker felt more comfortable following the metronome. The speaker heard the metronome through earphones, and articulated all speech tokens starting and ending with a closed mouth. Each VCV utterance was spoken synchronously with two consecutive beats of the time reference. The first beat marked the start of the first syllable, and the second beat the onset of the second syllable. The beats of the metronome consisted of transient sounds with a duration of 40 ms. Inspection of the waveform of the recorded utterances revealed that the time delay between the onset of the two syllables was 330 ms and 390 ms, for VC_V and V_CV stimuli, respectively. In the case of the monosyllabic stimuli, the first beat of the metronome indicated the start of the utterance, and the second beat the end of the utterance. The mean duration of the utterance was 388 ms and 375 ms, for VC and CV stimuli, respectively. Therefore, for all stimuli, the difference with respect to the time reference was within ± 37 ms; roughly the duration of a video frame.

The speech material was recorded using a video camera Canon EOS Rebel T5i with a resolution of 1920 x 1080 at 30 frames per second. The audio was captured at a sampling rate of 48 kHz, and a resolution of 24 bits, using a BK condenser microphone placed at the position of the camera (50 cm in front of the talker). At least five takes were made for each utterance, and the one with the highest quality (more natural and synchronous with the beat) was chosen. All acoustic speech tokens were normalized to have the same root mean square average level. The separate audio and video streams were time-aligned offline in the software Adobe Premiere Pro. The recorded articulations started and ended in a closed-mouth that lasted at least ten frames. In total, the duration was 2 s for the VCV and 1.7 s for VC stimuli.

3.2.2 Experimental procedure

In Experiment 1, 20 subjects were tested with VCV sequences (19 – 29 years of age, mean age 24, nine male). One subject who failed to correctly identify the unimodal and the congruent audiovisual stimuli was discarded, leaving 19 subjects. In Experiment 2, 11 subjects (20 – 27 years of age, mean age 23, three female) who did not participate in Experiment 1 were tested with VC and CV stimuli, and none of them was excluded from the experiment. All subjects reported normal hearing and normal or corrected-to-normal vision. All participants provided informed consent and were financially compensated for their time. All experiments were approved by the Science-Ethics Committee for the Capital Region of Denmark (reference H-16036391). The experimental data were collected in a soundproof booth. The stimuli were presented via a Dell 27-inch color screen, installed at 75 cm from the seating position. The sound was reproduced through a Genelec 8020C loudspeaker placed right below the screen. The sound level was set at a comfortable level of 65 dBA (40 dB above the background noise), which was measured using a BK 2250 sound level meter on a fast scale.

Subjects were instructed to report what they heard, and to speechread when the stimuli did not contain sound. Each stimulus presentation was followed by a screen with four response options available, /b/, /g/, /bg/ or /gb/. The subjects could select one option by pressing the corresponding response labeled on a computer keyboard. The tests were carried out in two experimental blocks of ten trials each. A short break of at least five minutes was encouraged after the first block was finished. Each trial contained all stimuli in random order. Before the actual experiment, there was a training session to familiarize the subjects with the task. The total duration of the test was between 45 and 60 minutes.

3.2.3 Data analysis

Audiovisual integration

The integration index was calculated to investigate the effects of syllabic context and audiovisual SOA on the integration of the combination stimuli. The integration index was computed by subtracting the sum of /bg/, /gb/, and /b/ response percentages, $P_{bg}^{A}+P_{gb}^{A}+P_{b}^{A}$, for the auditory stimulus from the corresponding sum of these response proportion, $P_{bg}^{AV}+P_{gb}^{AV}+P_{b}^{AV}$, in the audiovisual stimuli. Therefore, the index assigns lower values of integration to the audiovisual stimuli with more ambiguous auditory components.

Perceived consonant order

The BG-index, $\frac{P_{bg}}{P_{bg}+P_{gb}}$, was calculated to study the effect of syllabic context on the perceived consonant order. The index estimates the conditional probability of perceiving a leading bilabial consonant (/bg/) given that a consonant cluster is perceived. There were some instances in which the index could not be estimated due to the lack of a cluster response. These data were not considered when computing the mean index, nor for the statistical analyses.

Statistical analysis

The statistical tests were carried out with linear mixed-effects model analysis of variance (ANOVA), which were fitted to either the response percentages, the BG-index, or the integration index computed for each subject. Subjects were treated as random effects, whereas the fixed effects are specified in the results section for each analysis. All pairwise comparisons were performed with post-hoc analyses using Tukey's honest significant difference (HSD) test, unless stated otherwise. A significance level of 0.05 was used in all tests.

3.3 Results

3.3.1 Experiment 1: The effect of audiovisual timing on VCV combination stimuli

Effect of syllabic context

Figure 3.1 shows the mean response percentages obtained for the combination stimuli in which the "internal" audiovisual timing was varied by syllabic context while maintaining synchrony of the overall temporal structure of the tokens. The internally synchronous stimuli contained the consonant in either the offset (VC_V) or in the onset (V_CV) for both the auditory and visual VCV utterance. The internally asynchronous stimuli were either auditory-leading (auditory VC_V dubbed onto visual V_CV) or visual-leading (auditory V_CV dubbed onto visual VC_V).

The integration index was used to assess audiovisual integration while correcting for the ambiguity of the auditory stimulus (see section 3.2 Material and methods). To determine whether audiovisual integration influenced the responses to audiovisual stimuli, one-sample t-tests comparisons were performed on the integration index of the audiovisual stimuli. The analyses revealed that for each of the audiovisual stimuli tested, internally synchronous (one-sided t-test, consonant offset: $t(18) = 7.20$, $p < .0001$, $\mathrm{d} = 1.65$; consonant onset: $t(18) = 12.13$, $p < .0001$, $\mathrm{d} = 2.78$) as well as asynchronous (auditory-leading: $t(18) = 8.18$, $p < .0001$, $\mathrm{d} = 1.88$; visual-leading: $t(18) = 12.66$, $p < .0001$, $\mathrm{d} = 2.90$), the integration index was significantly greater than zero. This indicates that the observed cluster percepts were due to audiovisual integration.

To assess how syllabic context influenced audiovisual integration, a linear-

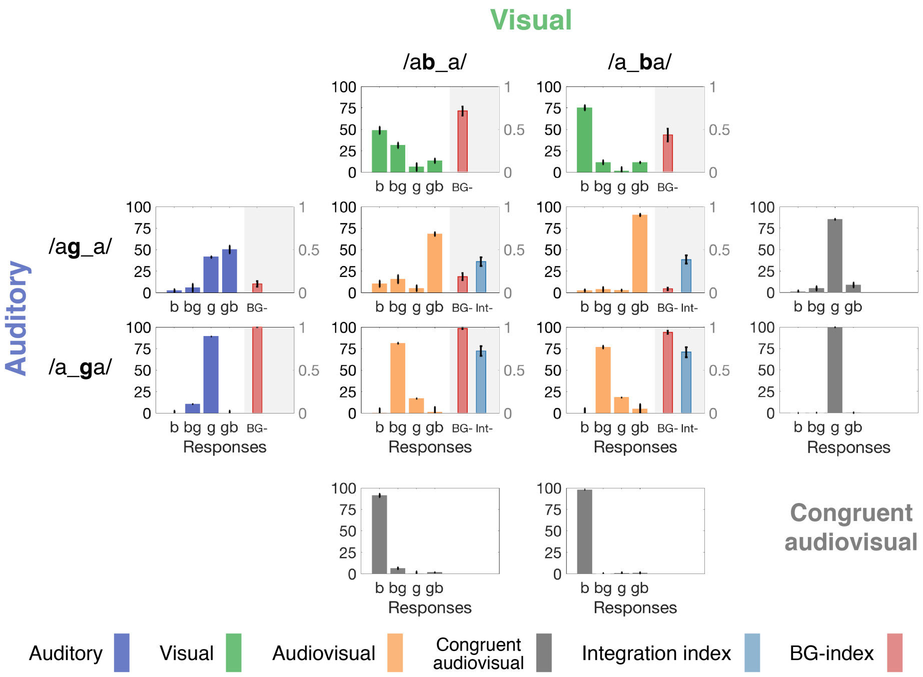


Figure 3.1: Effect of varying the syllabic context across modalities on the combination illusion. Mean response percentages across subjects. The leftmost column and top row represent the unisensory auditory (blue), and visual components (green), respectively. These components were produced with two different syllabic contexts. The audiovisual stimuli (orange) corresponded to all possible pairings of the unisensory components, resulting in either auditory-leading, visual-leading or synchronous stimulus pairings. The congruent audiovisual stimuli (gray) are shown in the rightmost column and lower row. The integration index (blue) and BG-index (red) are depicted within the shaded region. Error bars represent the standard error of the mean.

mixed-effects-model ANOVA was performed on the integration index with fixed effects of auditory (two levels: consonant offset or onset) and visual stimulus (two levels: consonant offset or onset), and subject as a random effect. The analysis showed a significant main effect of the auditory stimulus ($F(1,54) = 55.13$, $p < .0001$), whereas the effect of the visual stimulus ($F(1,54) = 0.013$, $p = .909$) as well as the interaction ($F(1,54) = 0.159$, $p = .691$) were insignificant. The significant effect of the auditory stimulus is reflected by the fact that the integration index was lower for the audiovisual stimuli that contained the auditory consonant in the offset than in the onset (Figure 3.1), as revealed by post-hoc comparisons using Tukey's HSD test ($p < .0001$). Hence, when corrected for the similarity between the audiovisual and the auditory stimuli, integration was weaker for the stimuli that contained auditory /ag_a/ than for auditory /a_ga/.

Since the integration index, and hence, audiovisual integration was significantly affected by the auditory stimulus, it was not possible to directly assess the perceived consonant order using the response percentages. This is because cluster responses would be influenced by both integration and perceived consonant order. Instead, we used the BG-index to quantify the perceived consonant order (see section 3.2 Material and methods). For the auditory-leading and the synchronous V_CV stimuli, the BG-index was undefined for one subject in both cases, due to zero cluster responses. For auditory /a_ga/, the index was undefined for eight subjects, whereas for visual stimuli only for /a_ba/ one subject did not report cluster responses, which made the estimation of the index infeasible.

To test the effect of syllabic context on the perceived consonant order, a linear-mixed-effects-model ANOVA was performed on the BG-index for the audiovisual stimuli with auditory (two levels: consonant offset, or onset) and visual stimulus (two levels: consonant offset, or onset) as fixed effects, and subjects as a random effect. The analysis yielded significant main effects for the auditory ($F(1{,}53.91) = 1179.09$, $p < .0001$) and the visual stimulus ($F(1{,}52.50) = 14.28$, $p = .0004$), whereas their interaction was marginally not significant ($F(1{,}52.50) = 3.87$, $p = .054$).

The main effect of the auditory stimulus reflects that /gb/ was the strongest illusion when the auditory /g/ was in the offset (with 91% and 68% /gb/ responses for the auditory-leading and synchronous VC_V stimuli, respectively). Conversely, /bg/ was the strongest when auditory /g/ was in the onset (with 81% and 77% /bg/ responses for the visual-leading and synchronous V_CV stimuli, respectively). The main effect of the visual stimulus, on the other hand, reflects the stronger /gb/ percept elicited when the visual consonant was in the onset (/a_ba/) than in the offset (/ab_a/). This indicates a strong effect of “internal timing” on perceived consonant order. However, as will be discussed further below, the similarity between the audiovisual and the unisensory stimuli needs to been taken into account.

The unisensory auditory stimuli differed in their similarity to consonant clusters. Auditory /ag_a/ was confused with /agba/ in 50% of the trials and correctly identified in 41% of the trials. In contrast, the consonant in unisensory /a_ga/ was correctly identified in 89% of the trials, and never confused with /agba/. The BG-index was, accordingly, lower when the consonant was in the offset than in the onset of the auditory stimulus (one-sided t-test, $t(10) = -29.68$, $p < .0001$, $d = -9.51$). This indicates that the main effect of the auditory stimulus on the BG-index for audiovisual stimuli could be related to the difference in the auditory unimodal confusions.

The visual condition also showed confusions for the VC_V context. Visual /ab_a/ was correctly identified in 49% of the trials and was confused with /abga/ in 32% of the presentations. The consonant in the visual /a_ba/ was correctly identified in 75% of the trials, and in the remaining trials it was confused almost equally with the two consonant cluster responses. This is reflected in the higher BG-index observed when the consonant was in the offset than in the onset of the visual stimulus (one-sided *t*-test, $t(17) = 2.86$, $p = .0054$, $\mathrm{d} = 0.99$). Thus, the similarity between the unisensory stimuli and consonant clusters may underlie the main effect of the visual stimulus on the perceived consonant order for audiovisual stimuli. Importantly, the trend towards a significant interaction effect on the BG-index of audiovisual stimuli was unlikely due to the similarity of the unisensory stimuli to the cluster articulations, but could reflect the asynchronous "internal timing" of the auditory and visual consonants.

For the unimodal presentations, the results suggest that the stimuli with the consonant in the offset (VC_V) were more ambiguous than those with the consonant in the onset (V_CV). However, it is noteworthy that the target consonants presented, auditory /g/ and visual /b/, were almost always perceived in the audiovisual stimuli in the two VCV contexts. This is evident if one adds, for each condition, the single consonant responses that match the consonant presented and the cluster responses that contain the consonant in the correct VCV context. For example, auditory /ag_a/ shows 42% /g/ responses and 50% /gb/ responses, thus representing 92% of the trials where /g/ was perceived. Likewise, visual /ab_a/ presents 49% /b/ responses and 32% /bg/ responses, thus representing 81% correct identifications of /b/. Thus, it is the intersyllabic stop that seems to be misperceived as a consonant. This is perhaps not surprising as it has been shown that a silence interval can induce the auditory perception of a stop consonant (Bastian, 1962; Dorman et al., 1979).

Given that we used non-conventional syllabic structures for our stimuli, we tested the congruent audiovisual conditions to confirm that the articulations recorded could be correctly perceived. These congruent conditions showed that the consonants were unambiguously perceived in both the VC_V (with 85% and 91%, for /g/ and /b/, respectively) and the V_CV (with 99% and 98%, for /g/ and /b/, respectively) contexts. Paired *t*-test between the auditory and the corresponding congruent audiovisual conditions showed that the percentage of correct responses was significantly higher for the congruent audiovisual conditions in the VC_V (one-sided, $t(18) = 10.27$, $p < .0001$, d = 2.68), and the V_CV context (one-sided, $t(18) = 5.26$, $p < .0001$, d = 1.84). These results, as expected, suggest that the integration of congruent audiovisual speech information enhanced the perception of the VCV stimuli.

Effect of audiovisual SOA

Figure 3.2A shows the response percentages for the audiovisual stimuli as a function of the audiovisual SOA. The left panel represents the results obtained for the stimuli with the consonants in the offset. When the auditory speech signal led the visual information by –200 ms, /gb/ was perceived in 91% of the trials. As the asynchrony went towards the visual-lead side, the /gb/ response tended to decrease. At synchrony (0 ms), /gb/ was reported 68% of the time, while the minimum of 47% was reached at 200 ms visual-lead. The right panel of Figure 3.2A shows the responses for the stimuli with the consonants in the onset. When the auditory signal led the visual component (-200 ms), the unisensory /g/ was perceived in 80% of the trials. Unlike VC_V stimuli, in this context the combination /bg/ was the most frequent percept for all other asynchronies, reaching a maximum of 78% responses at 100 ms visual-lead. To examine the effects of integration and perceived consonant order we used the integration- and BG-indices.

The integration index was calculated to investigate whether there was an influence of audiovisual SOAs on the integration of VCV combination stimuli (Figure 3.2B). A linear-mixed effects model ANOVA with fixed effects of audiovisual SOA (five levels) and syllabic context (two levels), and subjects as random effects was performed on the integration index. The outcome revealed a significant main effect of SOA ($F(4,162) = 16.96$, $p < .0001$), syllabic context ($F(1,162) = 58.01$, $p < .0001$) and their interaction ($F(4,162) = 20.92$, $p < .0001$). This significant main effect of syllabic context reflects that integration was weaker when the stimulus contained the consonants in the offset (VC_V) than when they were in the onset (V_CV).

As the syllabic context influenced the effect of audiovisual SOA on integration, we further examined the integration index across SOA for each syllabic context. For the stimuli with the consonants in the offset (VC_V), no significant differences on integration were found across SOA (the largest difference between –200 and 200 ms with $p = 0.98$, Tukey's HSD test). In contrast, for

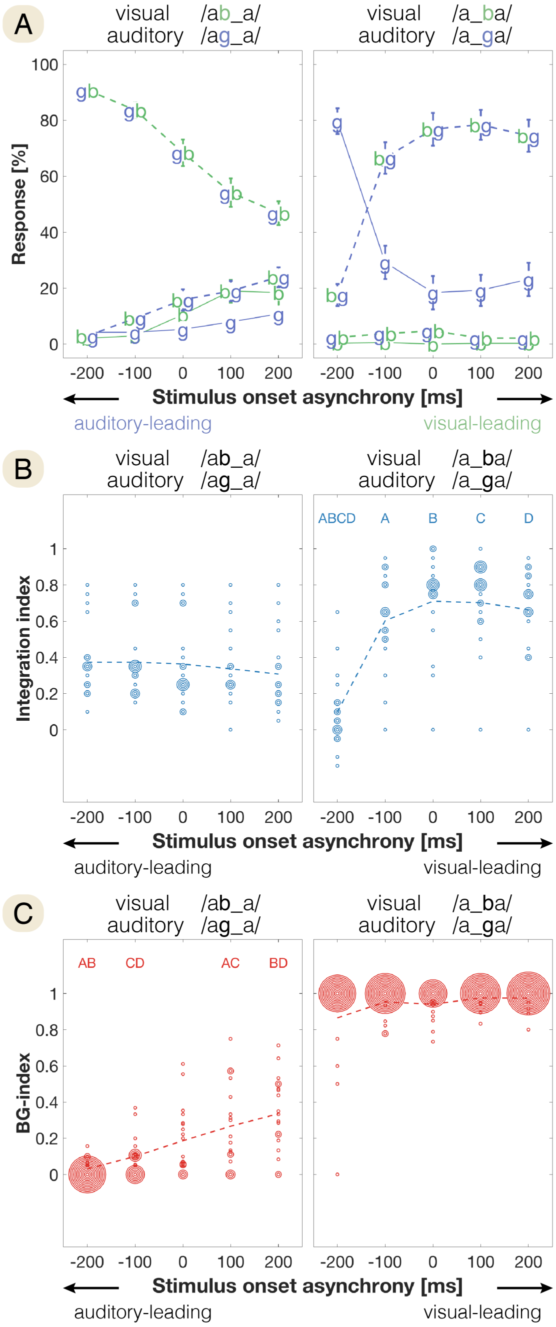


Figure 3.2: Effect of varying audiovisual SOA on the combination illusion for VCV stimuli. **A)** Mean response percentages. The responses are presented for the VC_V (left panel) and V_CV (right panel) stimuli. Single consonant responses are displayed alone in either green for /b/ or in purple for /g/. These are connected with solid lines of the same color. Cluster responses are connected by dashed lines, and consist of the two consonants with their respective order and color. The four response categories (/b/, /g/, /bg/ and /gb/) are shown as a function of SOA. Error bars represent the standard error of the mean. **B)** Integration index (blue) calculated for VC_V (left panel) and V_CV (right panel) stimuli. Concentric circles in the figure denote the individual mean indeces that resulted in the same value, and the dashed lines indicate the mean index across subjects. The indices are statistically different for the asynchronies sharing at least one uppercase letter of the same color. **C)** BG-index (red). Conventions as in **B**.

the stimuli with the consonants in the onset (V_CV), a visual-lead advantage was found. Significant effects of SOA on integration were found for the paired comparisons between –200 ms and all other SOAs ($p < .0001$, Tukey's HSD test). All statistically significant differences across SOAs are illustrated in Figure 3.2B with blue uppercase letters.

The fact that we did not find significant differences on integration across SOA for audiovisual stimuli with the consonant in the offset is somewhat surprising in light of previous research (Massaro and Cohen, 1993; Soto-Faraco and Alsius, 2009). However, these studies did not test stimuli with the consonant in the offset, which have been shown to require an extended range of asynchronies to show substantial differences on integration (Hampson et al., 2003). This suggests that the range of asynchronies tested, between –200 and 200 ms, was not enough to observe the audiovisual window of integration for VC_V combination stimuli.

The effect of syllabic context and audiovisual SOA on the perceived consonant order was studied by estimating the BG-index (Figure 3.2C). Due to the lack of cluster responses, the index was undefined for one subject across all asynchronies in the VC_V context, and for two other subjects at –200 ms. A linear-mixed effects model ANOVA with audiovisual SOA (five levels) and syllabic context (two levels) as fixed factors, and subjects as a random factor was performed on the BG-index. The results showed significant effects of SOA ($F(4{,}155.57) = 12.13$, $p < .0001$), syllabic context ($F(4{,}158.69) = 1319.11$, $p < .0001$), and their interaction ($F(4{,}155.57) = 3.61$, $p = .007$). This indicates that the perceived consonant order was influenced by the syllabic context of the audiovisual stimuli, and varied with the audiovisual asynchrony.

To determine how the interaction between SOA and syllabic context affected the perceived consonant order, we further examined the BG-index for each syllabic context across SOA. When the consonant was in the offset (VC_V), the BG-index was 0.03 at –200 ms, and it increased as SOAs went towards visual-leads, reaching 0.33 at 200 ms. In this context, Tukey's HSD test showed significant differences in the BG-index between –200 ms paired with 100 and 200 ms, and –100 ms contrasted with 200 ms ($p < .0001$). All statistically significant differences found across SOAs are illustrated in Figure 3.2C with red uppercase letters. In contrast, when the stimuli contained the consonants in the onset (V_CV), the index remained close to one across SOAs (the largest difference between –200 and 200 ms with $p = 0.39$, Tukey's HSD test). These findings indicate a differential effect of SOAs on the perceived consonant order, while VC_V audiovisual stimuli seemed prone to consonant order reversals, the V_CV stimuli were not.

3.3.2 Experiment 2: The effect of audiovisual timing on VC and CV combination stimuli

This experiment was designed to assess whether the effects of audiovisual SOA obtained with VCV stimuli were equivalent to those obtained with the VC and CV stimuli, which are governed by different articulatory constraints. In Experiment 1, the unimodal VC_V and V_CV were ambiguously perceived as a cluster response. As this could be due to the intersyllabic stop in the articulation (Bastian, 1962; Dorman et al., 1979), we hypothesized that using VC and CV stimuli, where there is no such stop, could make the unisensory stimuli less ambiguous and remove this confound. Also, since we did not find an effect of SOA on integration between 200 and 200 ms for VC_V stimuli, we extended the range of asynchronies considered, including 400 ms auditory-lead and 400 ms visual-lead.

Figure 3.3A shows the mean response percentages computed for the VC (left panel) and the CV stimuli (right panel). Percentages are shown as a function of SOA for all the same response options than the VCV stimuli. When auditory /ag/ was presented with visual /ab/, the subjects mostly perceived /gb/ or /g/ across all SOAs. From –200 to 0 ms, /gb/ was the most prominent percept with 55% at –100 ms. In this range, the second most frequent response was /g/ that accounted for 38% responses at –200 ms. In contrast, the combination stimuli auditory /ga/ paired with visual /ba/ produced mainly cluster responses /bg/ and auditory /g/ responses, as seen in the case of the V_CV stimuli. In this CV context, the combination /bg/ was the most frequent response with 58% judgments at 100 ms between 0 and 200 ms visual-lead.

As in Experiment 1, the effects of integration and perceived consonant order across SOA were examined using the integration- and BG-indices. Figure 3.3B shows the integration index as a function of SOA for the VC and CV stimuli. A linear mixed-effects model ANOVA with audiovisual SOA (seven levels) and syllabic context (two levels) as factors showed a significant effect of SOA ($F(6,130) = 11.05$, $p < .0001$) and the interaction ($F(6,130) = 5.57$, $p < .0001$), whereas the main effect of syllabic context on the index was not significant ($F(1,130) = 1.66$,

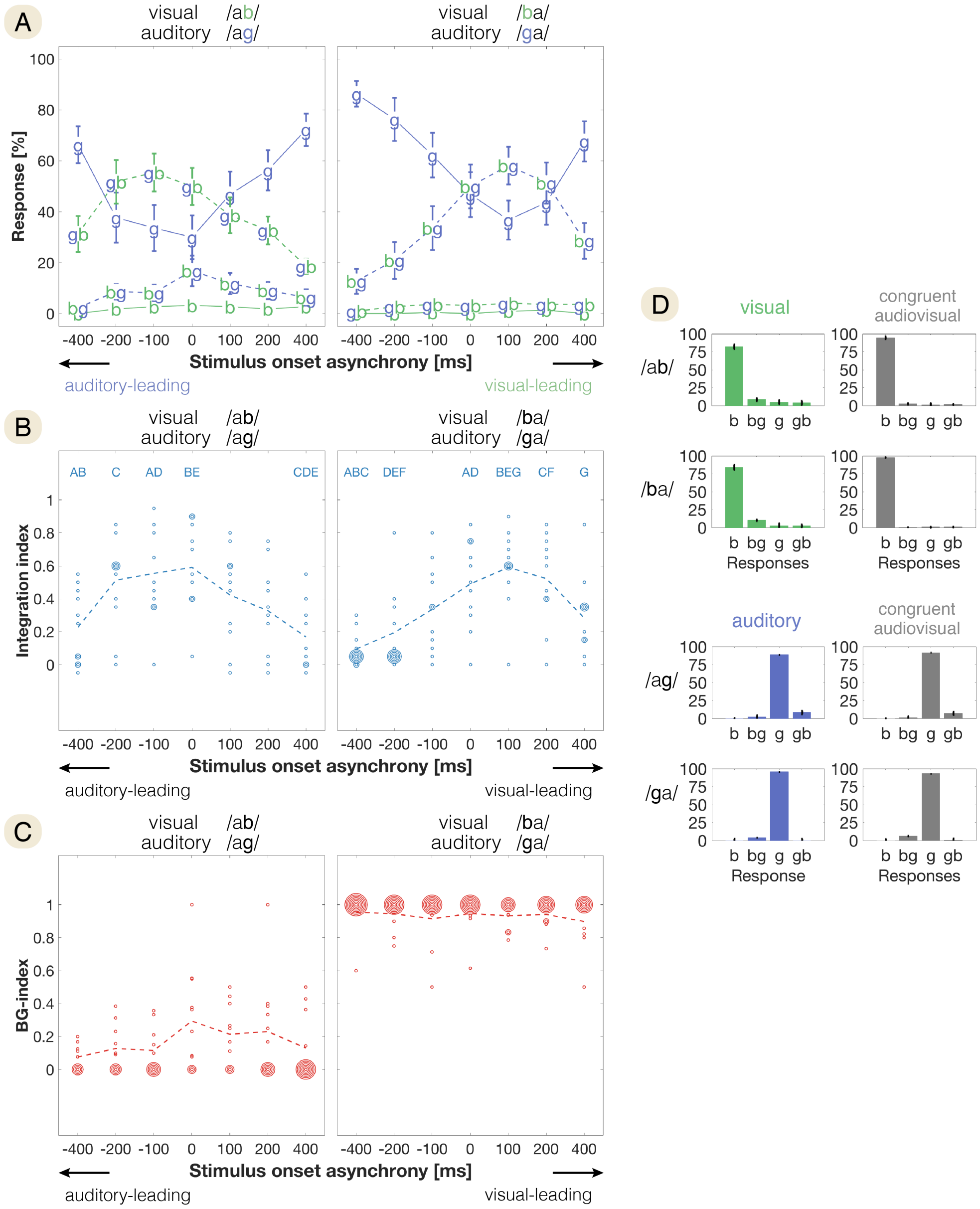


Figure 3.3: Effect of varying audiovisual SOA on the combination illusion for VC and CV stimuli. **A)** Mean response percentages. The responses are presented for the VC (left panel) and CV (right panel) stimuli. Single consonant responses are displayed alone in either green for /b/ or in purple for /g/. These are connected with solid lines of the same color. Cluster responses are connected by dashed lines and consist of the two consonants with their respective order and color. Error bars represent the standard error of the mean. **B)** Integration index (blue) calculated for VC (left panel) and CV (right panel) stimuli. Concentric circles in the figure denote the individual mean indices that resulted in the same value and the dashed lines indicate the mean index across subjects. The indices are statistically different for the asynchronies sharing at least one uppercase letter of the same color. **C)** BG-index (red) calculated for VC (left panel) and CV (right panel) stimuli. Conventions as in **B**. **D)** Auditory (left panel) and visual (right panel) conditions with their corresponding congruent audiovisual conditions tested in the VC and the CV context.

$p = .20$). For VC stimuli, the integration index was larger between –200 and 0 ms, with a maximum mean value of 0.6 at 0 ms. In contrast, the minimum mean value was 0.2 at 400 ms. Interestingly, the opposite asymmetry was found for CV combination stimuli. For this context, the integration index was maximum at 100 ms and minimum at –400 ms with mean values of 0.6 and 0.1, respectively. Blue uppercase letters in Figure 3.3B illustrate all statistically significant differences found with Tukey's HSD test for VC and CV stimuli. These results reflect two different patterns of responses for the VC and CV stimuli. For VC stimuli, the integration was stronger when the auditory stimulus was leading while the reverse was true for CV stimuli.

For each syllabic context, the perceived consonant order was examined comparing the BG-index across SOA (Figure 3.3C). For VC stimuli, the BG-index could not be estimated for one subject at four asynchronies (–400, –200, –100 and 100 ms), and for one subject at –400 ms. In the case of CV stimuli, the index was not calculated for one subject at four asynchronies (–100, 0, 100, and 200 ms), for one subject at two asynchronies (–400, and 400 ms), and for one subject at 200 ms. The mean value of the index varied between 0.29 at 0 ms and 0.07 at –400 ms for VCs, whereas for CV stimuli the mean value of the index was close to one consistently for all SOAs. A linear mixed-effects model ANOVA with audiovisual SOA (seven levels) and syllabic context (two levels) as factors showed a significant main effect of syllabic context ($F(1,118.29) = 751.32$, $p < .0001$), whereas neither SOA ($F(6, 116.94) = 1.18$, $p = .32$) nor the interaction of SOA and syllabic context ($F(6, 116.8) = 1.06$, $p = .39$) were significant. This suggests that altering audiovisual SOAs has no effect on the perceived consonant order for VC and CV stimuli. Also, the results support the findings of Experiment 1 that showed that the syllabic context strongly influences the perceived consonant order.

The congruent audiovisual and unimodal auditory and visual conditions were also tested for VC and CV stimuli. Figure 3.3D shows the mean percent responses. For all conditions, the consonant presented was correctly perceived more than 80% of the time. Compared to VC_V stimuli, when testing VCs the unisensory stimuli were not ambiguous, both auditory /ag/ and visual /ab/ were significantly more correct than the corresponding VC_V unimodal components (one-sided Welch's t-tests, /ag/ vs. /ag_a/: $t(23.96) = 4.86$, $p < .0001$, /ab/ vs. /ab_a/: d = 2.13; $t(29.72) = 4.31$ $p < .0001$, d = 2.28). This indicates that the unimodal confusions between auditory /ag_a/ and the cluster response /agba/

in Experiment 1 were indeed due to the syllabic stop in the VC_V sequence.

3.4 Discussion

Our results provide moderate support for the timing hypothesis. Varying the audiovisual SOA influenced the perceived consonant order only for VC_V but not for V_CV, VC or CV stimuli. In contrast, Hampson and colleagues reported an effect of audiovisual SOA on the perceived consonant order of VC combination stimuli (Hampson et al., 2003). In their results, the visual consonant was perceived to lead the auditory consonant for visual-leads, whereas for visual-lags, the auditory consonant was perceived to lead the visual. The differences between our results and those from Hampson et al. (2003) could indicate that the effect of audiovisual SOA on the perceived consonant order does not depend only on the syllabic context but also on small differences in the articulation of the speech stimuli.

Varying the “internal” timing of consonants by manipulating the syllabic context influenced the perceived consonant order as evidenced by the significant main effects of syllabic context in Experiment 1. These effects could, however, be due to the perceptual similarity of the unisensory stimuli to consonant clusters, since the unisensory VCV stimuli were perceived as ambiguous, particularly the auditory stimulus in the VC_V context. The interaction between syllabic context in the auditory and visual stimuli would indicate an effect of internal timing of speech features, such that the visual component would be more likely to be perceived as leading when it was prominent in the closing phase (VC_V) *and* the auditory component was more prominent in the opening phase (V_CV). This effect could not be explained by perceptual similarity but only trended towards significance.

Experiment 2 provided strong evidence for an effect of syllabic context on perceived consonant order. Across all SOAs, CV syllables produced combination illusions in which the visual component was leading while the perceived consonant order was reversed for VC syllables. These findings were somewhat surprising in light of previous research in which the auditory component was almost always perceived to lead the visual component (MacDonald and McGurk, 1978; Massaro and Cohen, 1993; Walker et al., 1995; Soto-Faraco and Alsius, 2009). The effect is unlikely to be due to the perceptual similarity between the unisensory stimuli and particular consonant clusters as the unisensory stimuli were unambiguous. Articulatory constraints could, however, explain the differences between CV and VC syllables: A visual bilabial CV stimulus imposes a strong constraint on when an additional velar consonant can be produced; it can only be uttered after the opening of the mouth, i.e., following the visual consonant. Conversely, for a VC stimulus, the production of the velar consonant would be ecologically valid only when preceding the bilabial.

As expected, varying audiovisual SOA influenced the strength of the integration, although not for the VC_V context in the range –200 to 200 ms that we studied here. This is consistent with previous findings in which audiovisual SOA affected the strength of the cluster percept, but not the perceived consonant order of CV combination stimuli (Massaro and Cohen, 1993; Soto-Faraco and Alsius, 2009). We also found that the asymmetry of the window of integration was in opposite directions for CV and VC syllables. Integration was stronger when the visual stimulus led for CV syllables, as shown in previous studies (Massaro and Cohen, 1993; Van Wassenhove et al., 2007; Soto-Faraco and Alsius, 2009). Surprisingly, integration was stronger when the visual stimulus lagged for VC syllables. The reason for this could be that if articulatory constraints impose the perceived order, a combination illusion is perhaps a more likely perceptual interpretation if the stimuli actually arrived in that order.

For CV syllables, the visual gestures tend to lead the onset of the voice due to the anticipatory mouth movements (Schwartz et al., 2004; Van Wassenhove et al., 2005; Venezia et al., 2016). This has been suggested to support the common finding of an asymmetric audiovisual window of integration favoring visual-leads (Chandrasekaran et al., 2009). However, Schwartz and Savariaux (2014) more recently demonstrated that for more complex syllabic structures, such as VCV, acoustic and visual speech gestures are reasonably synchronous, spanning a range of SOAs that includes both visual-leads and auditory-leads. The finding that auditory speech information may be available before visible speech has been attributed to the audible speech gestures that could be poorly visible for some articulatory contexts (Troille et al., 2010; Schwartz and Savariaux, 2014). Thus, if the asymmetry of the audiovisual window of integration depends on the natural timing, which could be variable for audiovisual speech, the asymmetry of the window itself could also vary with the syllabic context as we found.

Our results clearly show that the perceived consonant order in combination illusions can be manipulated consistently by varying syllabic context. This finding contrasts with most previous studies in which the cluster response with the labial consonant leading was the most frequent (MacDonald and McGurk, 1978; Massaro and Cohen, 1993; Walker et al., 1995; Soto-Faraco and Alsius, 2009). This could have two main reasons. In the case of monosyllabic stimuli, there seems to be a preference in the literature for using CV combinations, for which the visual component is perceived as leading. The second reason could be that the talkers are usually not given specific instructions on how to articulate. This could be crucial for VCV utterances, in which the talkers could spontaneously tend to intone the consonant in a syllable-initial position, hence producing cluster responses similar to those obtained with V_CV in Experiment 1.

The strength of the McGurk illusion varies greatly across studies (Nath and Beauchamp, 2012; Alsius et al., 2017). By using a metronome to time articulation in our recordings, we limited the variability across stimuli. This could have been especially useful in varying the syllabic context systematically in VCV stimuli. Also, we believe that the inclusion of a metronome has helped us to produce strong combination illusions, as it has been suggested that McGurk stimuli with matched cross-modal speaking rate and clarity of the articulation elicit stronger illusory percepts (Munhall et al., 1996). The use of a metronome during the recordings of McGurk stimuli could be a good step towards more controllable and better-characterized stimuli, which might have been lacking in many studies (Alsius et al., 2017).

It is noteworthy that a full account of the combination illusion might require examining additional stimulus features that were not considered in this study. The burst hypothesis, for instance, could explain why the syllabic context of the auditory stimulus had the most substantial effect on the perceived consonant order as it could be the timing of the burst that determines the cluster response. Further testing is required in order to clarify the specific role of the consonant release burst in the perceived consonant order of the combination. Importantly, the hypotheses described in this study are not mutually exclusive, as the effects of similarity, audiovisual timing, and articulatory constraints all seem to influence our results.

3.5 Conclusion

In conclusion, our findings show that the perceived consonant order depends on the syllabic context of the audiovisual stimulus. Furthermore, we show that the perceived consonant order is informative of the nature of audiovisual integration of speech. Whether a cluster percept takes one order of consonants or another seems to reflect an underlying integration process mediated by audiovisual timing, perceptual similarity, and articulatory constraints. Although audiovisual SOA was found to influence mostly integration, it also affected the order of consonants in the cluster percept to a lesser extent. Constraints imposed by the articulatory gestures seemed to account for the marked difference in the perceived consonant order of VC and CV audiovisual stimuli. Crucially, the asymmetry of the audiovisual window of integration was found to favor either visual-leads or auditory-leads depending on the syllabic context. This result seems to emphasize that in addition to the perceive consonant order, the articulatory constraints could also define the optimal audiovisual timing for integration. Consequently, we argue that future experiments and models of audiovisual integration of speech should include distinct consonant order as response categories.

Acknowledgments

This research was supported by the Oticon Centre of Excellence for Hearing and Speech Sciences (CHeSS) at the Technical University of Denmark (DTU).

4

Feature-based audiovisual speech integration of multiple sequences[a]

Abstract

Speech perception typically involves the integration of auditory and visual information. This is seen in the McGurk combination illusion, in which a visual utterance, i.e., /ipi/, dubbed onto an acoustic utterance, i.e., /iki/, produces a combination percept, i.e., /ipki/. However, it is still unclear how phonetic features are integrated audiovisually. Here, we studied audiovisual speech perception by decomposing the auditory component of McGurk combinations into two streams. We show that auditory /i_i/, where the underscore indicates an intersyllabic silence, dubbed onto visual /ipi/ produces a strong illusion of hearing /ipi/. We also show that adding an acoustic release burst to /i_i/ creates a percept of /iki/. An auditory continuum was created with stepwise temporal alignments of the release burst and /i_i/. When dubbed onto visual /ipi/, this continuum was perceived mostly as /ikpi/ when the burst was closer to the initial vowel, and mostly as /ipki/ when the burst was closer to the final vowel. These results are indicative of feature-based audiovisual integration where the timing between the acoustic burst and the visually perceived articulatory movements determine the acoustically perceived consonant order.

Keywords: Audiovisual speech perception, combination illusion, McGurk effect, cross-modal timing.

---

[a] This chapter is based on Gil-Carvajal et al. (2020c), which was previously presented in preliminary form (Gil-Carvajal et al., 2019).

4.1 Introduction

The visible facial gestures accompanying the voice of the talker in face-to-face conversations facilitate speech perception (Sumby and Pollack, 1954; Arnold and Hill, 2001). This is particularly advantageous in noisy listening situations (Binnie et al., 1974; MacLeod and Summerfield, 1987; Schwartz et al., 2004). Visual speech also influences auditory perception in artificial situations, in which a voice dubbed onto phonetically incongruent facial gestures produces an illusory auditory percept. This remarkable demonstration of phonetic integration for audiovisual speech is known as the McGurk effect (McGurk and MacDonald, 1976; MacDonald and McGurk, 1978; Tiippana, 2014). In the McGurk fusion illusion (McGurk and MacDonald, 1976; Van Wassenhove et al., 2007; Schwartz, 2010), the audiovisual pairing of an auditory consonant (i.e., /ipi/) and a non-labial visual consonant (i.e., /iki/) elicits the perception of a third fused consonant (i.e., /iti/). In the McGurk combination (McGurk and MacDonald, 1976; Green and Norrix, 1997; Soto-Faraco and Alsius, 2009; Baart et al., 2017), the audiovisual pairing of an auditory non-labial consonant (i.e., /iki/) and a visual labial consonant (i.e., /ipi/) leads to a cluster percept in which both consonants are represented (i.e., /ipki/ or /ikpi/). Although the McGurk effect has been extensively investigated to explore the underlying mechanisms of audiovisual speech perception (Alsius et al., 2017), it is still unclear which phonetic cues are integrated to produce fusions or combinations.

While acoustic consonants provide sufficient information about both the manner and place of articulation, the speech cues that humans can reliably extract from the visual articulation of consonants are mostly related to the place of articulation only (Binnie et al., 1974; Dowell et al., 1982; Rouger et al., 2007). Accordingly, place cues seem to play a major role in audiovisual speech perception. According to MacDonald and McGurk's manner-place hypothesis (1978), fusion illusions arise as a compromise between the frontal, or labial, place of articulation in the acoustic speech stream and the velar place of articulation in the visual speech stream. This compromise results in an integrated percept of an alveolar place of articulation. This account does, however, not apply to combination illusions. Instead, MacDonald and McGurk proposed that combination illusions occur because the discrepancy between acoustical and visual place information is too large for the integration mechanism to find a plausible compromise, such that both consonants are perceived. This indicates that combination illusions may not be a result of audiovisual integration. Still, multisensory interactions must take place for the visual consonant to affect auditory perception and for visual information to affect auditory evoked potentials (Baart et al., 2017).

Green and Norrix (1997) presented another account for why some audiovisual phoneme pairings elicit combination illusions while others produce fusion illusions. For stop consonants, such as /k/ and /p/, the most salient perceptual cues are the consonant burst and the formant transitions that precede and/or follow the consonant release (Dorman and Raphael, 1980). The perceived intensity of the consonant burst depends on the place of articulation, and is usually stronger for non-labial consonants (/k/) than for labials (/p/) (Dorman et al., 1977; Colin et al., 2002). Green and Norrix (1997) proposed that this could be the reason for why combination illusions are perceived when the acoustic stimulus contains non-labial consonants but not when it contains labial consonants. In support of this hypothesis, they found that decreasing the sound intensity of consonant bursts decreased the strength of combination illusions. Further support was provided by Colin et al. (2002) showing that combination illusions occur more frequently with unvoiced or aspirated consonants in which the burst is generally stronger than in the case of voiced consonants. Compared to the manner-place hypothesis, these findings indicate that the intensity of the acoustic burst place cue, rather than the incongruence of audiovisual place information, determines whether place cues are heard as a component in combination illusions. These findings also suggest that the acoustic burst is a sufficient acoustic cue for eliciting combination illusions.

A typical finding in many studies of the combination illusion has been that the perception of the labial consonant leads the perception of the non-labial consonant in the combination response (MacDonald and McGurk, 1978; Walker et al., 1995; Massaro and Cohen, 1993; Soto-Faraco and Alsius, 2009; Jiang and Bernstein, 2011), although not always (Hampson et al., 2003). This could be due to some asynchrony inherent in the stimulus (Chandrasekaran et al., 2009; Venezia et al., 2016) or the differences in perceptual processing times (Van Wassenhove et al., 2007; Venezia et al., 2016). A few studies (Massaro and Cohen, 1993; Soto-Faraco and Alsius, 2009) examined if the perceived consonant order varied with the audiovisual stimulus onset asynchrony (SOA), but no substantial effect was found. One problem with this approach is that the cross-modal asynchrony is likely to influence the strength of the integration in addition to the perceived consonant order (Van Wassenhove et al., 2007).

One strategy to minimize the effect of asynchrony on audiovisual integration is to vary only the timing of the consonants while the vowels are kept synchronous across modalities (Gil-Carvajal et al., 2020b). In the present study, we employed a more specific approach by varying the timing of the burst and aspiration cues only. To do so, we isolated the burst and aspiration from a recording of acoustic /iki/. We then added the isolated acoustic cues to a recording that contained the same vowel context but no consonant, and varied the timing between the vowels and the burst across nine SOAs. Our hypothesis was that if the consonant burst and aspiration are necessary cues for eliciting the combination illusion, then varying the timing of these cues when dubbed onto a video of bilabial articulatory movements would be sufficient to vary the perceived consonant order in the combination illusion. To further test the effect of varying the timing between the acoustical burst and the visual articulatory gestures, we paired the auditory continuum with two visual syllabic contexts. This resulted in two vowel-consonant-vowel (VCV) continua in which the visual consonant was pronounced either early, in the offset of the initial syllable, or late, in the onset of the final syllable. If the perceived consonant order depends on the timing between acoustic and visual cues, visual syllabic context should also influence the perceived consonant order.

4.2 Methods

4.2.1 Subjects

Seventeen native French speakers (mean age 25, five female) participated in the study. Fourteen were recruited and tested at the Grenoble Alpes University in France, and three were recruited and tested at the Technical University of Denmark. All subjects reported to have normal hearing and normal or corrected-to-normal vision. Before the testing, all participants provided written consent and all experiments were approved by the Comité d'Ethique pour les Recherches Non Interventionnelles (CERNI) in France, reference IRB00010290, and by the Science-Ethics Committee for the Capital Region of Denmark (reference H-16036391).

4.2.2 Stimuli

The recorded speech material consisted of the disyllables /i_i/, /ip_i/, /i_pi/, /i_ki/, /ikpi/, and /ipki/ articulated by a native French speaker. The underscore represents an intersyllabic silence, which corresponded to 300 ms for /i_i/, while the mouth closure interval for /ip_i/ and /i_pi/ (shown in Figure 1) lasted 120 and 200 ms, respectively. The speaker pronounced each syllable synchronously with two consecutive beats of a metronome, which was delivered through EarPods and set at 130 beats per minute. The purpose of the metronome was to provide an auditory reference that could enable the production of speech utterances with similar speaking rate and duration.

To create stimuli in which only the timing of the burst and aspiration were affected, we used two separate auditory streams. One stream contained only the vowels and consisted of the recorded sound of /i_i/. The second stream contained only the consonantal burst and aspiration with a duration of 100 ms, which were extracted from the recorded articulation of /i_ki/. An auditory continuum was then generated offline by adding the two streams at nine different SOAs with a step size of 50 ms. At one end (–200 ms), the waveform of the initial vowel fully overlapped with the burst, and at the other end (200 ms), the final vowel fully overlapped with the burst, as shown in Figure 4.1. At the center of the continuum (0 ms), the burst was in the middle of the two vowels. Two audiovisual continua were also created by dubbing the auditory continuum onto video recordings of the visual articulatory gestures for /ip_i/ (consonant offset) or /i_pi/ (consonant onset).

Additionally, the recorded cluster articulations /ikpi/, /ipki/, and the articulation of the vowel context /i_i/ were presented in the auditory, visual and congruent audiovisual conditions. Also the articulations /ip_i/ and /i_pi/ were presented only in the visual and the congruent audiovisual conditions. Results for these control conditions are reported in the supplementary material (Figures 4.5, 4.6, and 4.7). In total, 40 stimuli were tested in the experiment (see Table 4.1 in the supplementary material).

The audio of the speech material had a sampling rate of 48 kHz, and a resolution of 24 bits. The video had a resolution of 720 x 576 and frame rate of 25 Hz. The total duration of each stimulus was one second. All visual articulations started and ended with a neutral expression of the speaker with the mouth closed, which lasted at least two video frames.

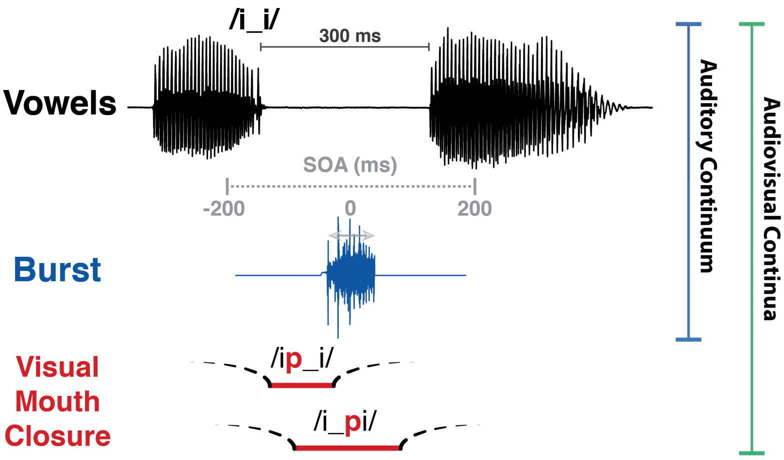


Figure 4.1: Schematic of the tested stimuli. The auditory continuum consisted of the vowels and the burst added at nine different SOAs (from -200 to 200 ms, with a step size of 50 ms). The two audiovisual continua were created by pairing each of the stimuli in the auditory continuum with the visual articulations of /ip_i/ and /i_pi/. The horizontal red solid lines indicate the mouth closure intervals for /ip_i/ and /i_pi/, respectively.

4.2.3 Procedure

The experiments were conducted in sound-proof booths. The subjects were seated 80 cm in front of a 21.5 inch Dell monitor, from which the videos were reproduced. The sound was played back monaurally at 65 dBA from a loudspeaker positioned above the computer monitor. Prior to the experiment, the subjects received written instructions in French and performed a training session that lasted one trial. The experiment was conducted in two separate blocks of ten trials each. Within a trial, all stimuli were presented in random order. The subjects were asked to report what they heard in each trial, or what they saw in the case of the visual trials. The response options were labelled on a computer keyboard and corresponded to /k/, /p/, /kp/, /pk/, and “no consonant”. The subjects took a five-minute break between the experimental blocks. The total duration of the experiment session was 60 minutes.

4.2.4 Data analysis

The analysis of responses was carried out in three steps: First we analysed how observers perceived the consonant burst and how this was influenced by the visual stimuli. To do this, we used responses pooled across response categories that contained the burst, i.e. /k/, /pk/ and /kp/ as our dependant measure. Next we analyzed how observers perceived the phonetic cue from the bilabial articulatory movements both visually and audiovisually. To do this, we used responses pooled across response categories that contained the bilabial closure, i.e. /p/, /pk/ and /kp/ as our dependant measure. Finally, we analyzed the perceived consonant order by defining the PK-index. The index was computed as the proportion of /pk/ responses out of the total number of combination responses, and hence, estimates the conditional probability of obtaining a /pk/ response given the occurrence of a combination response. For all analyses, the effects were evaluated using linear-mixed model analyses of variance (ANOVAs), with subject as a random factor and SOA and stimulus type as fixed factors. Post-hoc multiple comparisons were performed with Tukey's HSD test. Analyses involved the mean proportions of percepts with /k/ (step 1), the mean proportion of percepts with /p/ (step 2) or the mean PK-index (step 3). A significance level of 0.05 was considered in all analyses.

4.3 Results

4.3.1 Perceiving the consonant burst

To study the effect of SOA and articulatory gestures on the perception of the auditory stream, we analyzed the responses containing /k/, (i.e. /k/,/pk/ and /kp/ responses), as these indicate that the burst was indeed perceived as a consonant. Figure 4.2 shows these response percentages obtained for the auditory (panel A) and audiovisual continua (panel B) with the two visual contexts (/ip_i/ and /i_pi/) as a function of SOA. Overall, for the auditory continuum as well as the audiovisual continua, the responses containing /k/ (blue dotted lines) were more frequent than /p/ responses (red solid lines) and "no consonant" responses (black solid lines) across most SOAs. A linear-mixed model ANOVA with fixed factors for visual context (three levels: no visual, /ip_i/, and /i_pi/) and SOA (nine levels), and subject as a random factor, was performed on the sum of /k/, /kp/ and /pk/ response percentages. The outcome of the test revealed a significant main effect of SOA [$F(8,416) = 50.31$, $p < .0001$] and visual context [$F(2,416) = 68.89$, $p < .0001$] on the perception of /k/, whereas their interaction was not significant [$F(16,416) = 1.18$, $p = .278$].

The main effect of the visual context on the perception of the auditory stream reflects the lower percentage of responses containing /k/ in the audiovisual

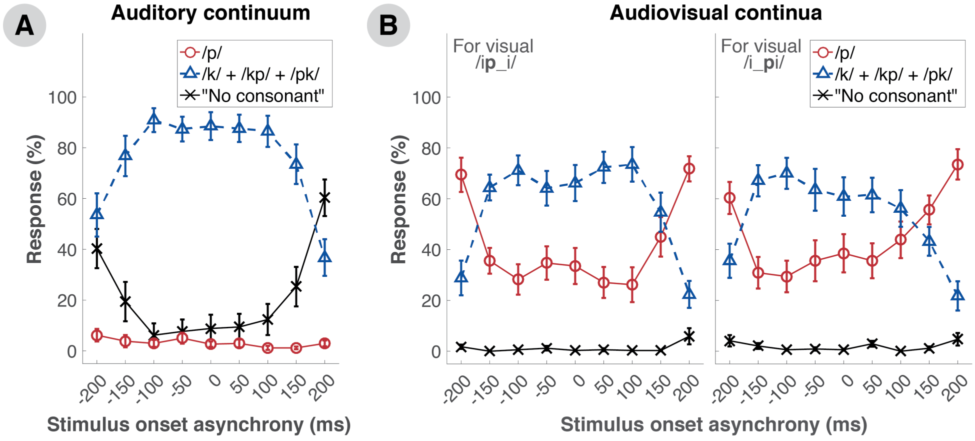


Figure 4.2: Mean response percentages obtained for the auditory continuum **(A)** and the audiovisual continua **(B)** for the two visual articulatory contexts, /ip_i/ and /i_pi/, as a function of SOA. The red solid lines represent /p/ responses, the blue dotted lines represent the responses containing /k/, and the black solid lines indicate the "no consonant" responses. The error bars show the standard error of the mean across subjects.

continua (panel B of Figure 4.2) than in the auditory continuum (panel A of Figure 4.2). Post-hoc multiple comparisons using Tukey's HSD test confirmed the significant differences in the responses containing /k/ between the auditory continuum and each of the two audiovisual continua ($p < .0001$), but not between the two audiovisual continua ($p = .097$). This indicates that both audiovisual continua decreased the perception of /k/ across SOA in a similar way, which could be due to a "visual capture" effect reflected in the larger proportion of /p/ responses obtained in the audiovisual continua than in the auditory continuum.

Varying SOA also affected the perception of /k/ responses. The effect was more pronounced at the extreme SOAs (± 200 ms SOA in Figure 4.2) for which the percentage of responses with /k/ decreased for all continua. The post-hoc multiple comparisons using Tukey's HSD test showed significant differences between –200 ms and all other SOAs ($p < .0001$; except for 200 ms with $p = .015$), and 200 ms contrasted with all other SOAs ($p < .0001$). Interestingly, for all comparisons between –150 ms and 100 ms, no significant differences were found ($p > .37$), revealing a "plateau" region in which /k/ was similarly perceived across SOAs.

The responses obtained for the auditory continuum indicate that the two auditory streams in the continuum (burst and vowels) were perceived as /iki/ across most SOAs despite the lack of formant transitions. Importantly, /k/ was not clearly perceived for all of the SOAs, since at –200 and 200 ms the "no consonant" responses increased at the expense of the responses containing /k/. This suggests that it is not the burst alone that is perceived as a VCV, which is further supported by the fact that the subjects correctly perceived the auditory stimulus /i_i/ (mean response percentage of 97%, Figure 4.7 in the supplementary material).

4.3.2 Perceiving the bilabial phonetic cue

The effect of SOA and visual articulatory gestures on the visual influence was studied by analyzing the sum of /p/, /kp/ and /pk/ responses, in which the visual /p/ influenced speech perception. Figure 4.3 shows the response percentages obtained for the audiovisual continua in the two visual articulatory contexts, /ip_i/ (left panel) and /i_pi/ (right panel), as a function of SOA. For the two audiovisual continua, the responses containing /p/ (red dotted lines) were substantially more frequent than /k/ responses (blue solid lines) as well as "no consonant responses" (black solid lines), which were negligible. A linear-mixed model ANOVA was fitted to the sum of responses containing /p/. The subject was treated as a random factor, whereas the visual context (two levels: /ip_i/ and /i_pi/) and SOA (nine levels) were treated as fixed factors. The test revealed a significant effect of SOA [$F(8,272) = 9.52$, $p < .0001$], while the visual context [$F(1,272) = 2.34$, $p = 0.127$], and the interaction of visual context and SOA [$F(8,272) = 0.81$, $p = .59$] were insignificant.

The main effect of SOA on the perception of the visual stream is reflected by the decreased responses containing /p/ for the median SOAs (between –50 and 50 ms in Figure 4.3). Post-hoc multiple comparisons using Tukey's HSD test confirmed the significant differences between –50 ms contrasted with all other SOAs ($p < .034$), except for 0 and 50 ms ($p > .80$), and for 0 ms compared to 100, 150, and 200 ms ($p < .001$). In contrast, no significant differences were found for all comparisons in the range from –200 to –100 ms ($p = .99$), nor in the range from 100 to 200 ms ($p = .99$). These results suggest that audiovisual integration occurred more frequently when the burst was closer to the vowels.

Across SOAs, the response percentages containing /p/ were independent of the visual context, as both visual /ip_i/ (left panel in Figure 4.3) and /i_pi/ (right panel in Figure 4.3) produced similar responses. This is in agreement with how these two visual stimuli were perceived in the visual-only condition (Figure 4.6, supplementary material). Here the response almost always contained /p/,

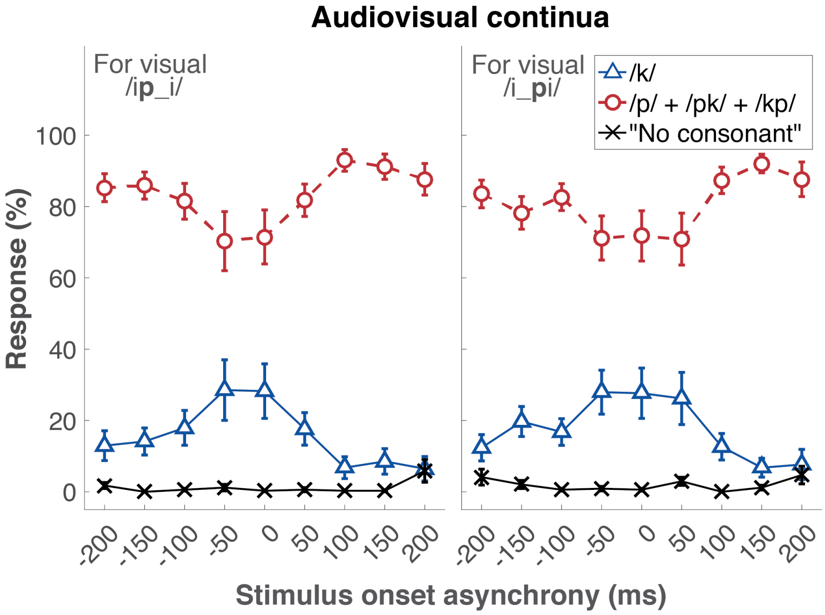


Figure 4.3: Mean response percentages obtained for the audiovisual continua in the two visual articulatory contexts, /ip_i/ (left panel) and /i_pi/ (right panel), as a function of SOA. The blue solid lines represent /k/ responses, the red dotted lines represent the responses containing /p/, and the black solid lines indicate the "no consonant" responses. The error bars show the standard error of the mean across subjects.

as reflected in 94% and 97% of responses for /ip_i/ and /i_pi/, respectively. However, for these visual stimuli, cluster percepts were also frequently obtained due to the difficulty in detecting whether the articulation contained /k/ in addition to /p/. Also, the subjects were remarkably successful in recognizing when the visual articulation did not contain a consonant, as indicated by 99% correct identifications of /i_i/ (Figure 4.6, supplementary material). Thus, since the bilabial /p/ was perceptually strong in the two articulatory visual contexts, it could have influenced the responses to audiovisual continua similarly.

4.3.3 Assessment of the perceived consonant order

The PK-index was used to assess the perceived consonant order of the combination responses. Figure 4.4 shows the PK-index estimated for the two visual contexts as a function of SOA. In general, for both audiovisual continua, /ip_i/ (left panel) and /i_pi/ (right panel), the PK-index was mostly close to zero at negative SOAs (from –200 to –100 ms) and mostly close to one at positive SOAs (from 50 to 200 ms). This is shown by the individual indices (concentric circles) as well as the mean indices across subjects (dotted lines) in Figure 4.4. A linear mixed-effects model was fitted to the mean PK-index. The visual context and SOA were taken as fixed factors and the subjects as a random factor. The outcome of the test revealed a significant main effect of visual context [$F(1{,}229.52) = 18.95$, $p < .0001$] and SOA [$F(8{,}229.46) = 70.21$, $p < .0001$], and an insignificant two-way interaction [$F(8{,}227.58) = 1.31$, $p = .24$].

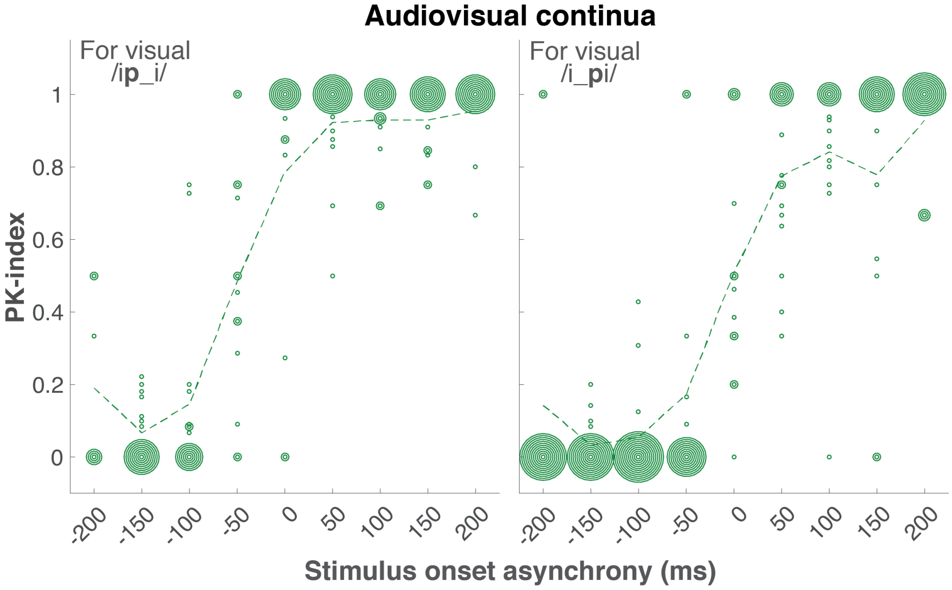


Figure 4.4: PK-index obtained for the audiovisual continua in the two visual articulatory contexts, /ip_i/ (left panel) and /i_pi/ (right panel), as a function of SOA. The concentric circles indicate the individual indices and the dotted lines the estimated mean across subjects.

The effect of SOA on the PK-index for the two audiovisual continua can be seen in the two distinct regions found with different combination responses (Figure 4.4). One region with a small PK-index, from –200 to –100 ms, for which the asynchrony produced mostly /kp/ responses, and another region with a high PK-index, from 50 to 200 ms, where mostly /pk/ responses were perceived. Tukey's HSD test showed significant differences in the PK-index for all possible pairwise comparisons across regions ($p < .0001$). No significant differences were found in the PK-index for any of the pairwise comparisons within the region with small PK-index ($p > .60$), nor within the region with high PK-index ($p > .85$). These results suggest that, for the two audiovisual continua, the subjects perceived one order of consonants when the burst was closer to the initial vowel, and the reverse consonant order when the burst was closer to the final vowel.

For each audiovisual continuum in Figure 4.4, the consonant order reversal occurred at different SOAs. For the continuum paired with visual /ip_i/ (left panel), the reversal occurred (earlier) at approximately –50 ms, and for the continuum paired with visual /i_pi/ (right panel), the reverse percept occurred (later) at about 0 ms. This is consistent with the fact that, in the case of the articulation of /ip_i/, the lips are closed earlier to produce the bilabial consonant than in the case of the articulation of /i_pi/. Post hoc Tukey's HSD test confirmed the significant differences between the two visual contexts on the PK-index ($p <$ .0001). This demonstrates that the perceived consonant order depends on the timing of the acoustic burst and the visual articulatory gestures.

4.4 Discussion

The main result of the current study is that the perceived consonant order in McGurk combinations can be reversed consistently by varying the timing of the burst and aspiration of the auditory component. Importantly, we show that the timing at which the reversal in the perceived consonant order occurs depends on the temporal alignment of the burst relative to the articulatory mouth gestures of the speaker. The results support the hypothesis that the burst and aspiration are important cues for audiovisual speech perception, affecting the perceived consonant order for McGurk combinations, and not only the strength of the integration of the cluster percept as has been reported earlier (Green and Norrix, 1997). The finding that the timing of the acoustic burst can determine the perceived consonant order is novel in light of the current literature on audiovisual speech perception. While Massaro and Cohen (1993) reported that visual lead SOAs increase the strength of the integration for McGurk consonant-vowel (CV) combinations, the study by Hampson et al. (2003) showed an effect of varying SOA on both the strength and the perceived consonant order of McGurk vowel-consonant (VC) combinations. Although these studies only varied the audiovisual SOA, they indicated that the timing of audiovisual speech cues influences the combination illusion. More recently, Gil-Carvajal et al. (2020b) tested VCV combinations by varying either the audiovisual SOA, or the syllabic context, such that the stimulus was either a CV with an added vowel (V_CV) or a VC with an added vowel (VC_V). The study of Gil-Carvajal et al. (2020b) showed that varying the syllabic context produces McGurk combinations with different perceived order of consonants, either with the labial visual consonant leading (for V_CV stimuli) or the non-labial acoustic consonant leading (for VC_V stimuli), but it remained unclear which audiovisual stimulus features influenced the response. On this basis, in the present study, we went a step further by only altering the timing of the burst and aspiration, while the vowels were presented synchronously. This allowed us to isolate the stimulus features, and hence, provide evidence for a feature-based audiovisual integration of speech.

Varying only the timing of the burst and aspiration could be equivalent to changing the syllabic context of natural speech articulations as in Gil-Carvajal et al. (2020b). This idea is consistent with the observation that we mostly found cluster responses with the labial consonant leading when the burst was closer to the onset of the vowel, as would be the case for CV stimuli (Massaro and Cohen, 1993; Soto-Faraco and Alsius, 2009; Gil-Carvajal et al., 2020b). Likewise, the obtained reverse combination responses (with the non-labial consonant leading) when the acoustic burst was closer to the offset of the vowel are in agreement with the previously reported responses for VC combination stimuli (Hampson et al., 2003; Gil-Carvajal et al., 2020b), even though the reverse order of consonants observed in Hampson et al. (2003) was only the most frequent percept when the visual component lagged the auditory component. Taken together, these results suggest that the position of the consonant within the stimulus, either consonant offset or onset, influences the perceived consonant order, and that this effect may be driven by the timing of the burst and aspiration relative to the articulatory gestures of the mouth, as seen in our results.

The strength of the combination illusion can vary substantially across studies (MacDonald and McGurk, 1978; Walker et al., 1995; Green and Norrix, 1997; Jiang and Bernstein, 2011). The cluster percept of the combination illusion also tends to vary sporadically, which could be why some researchers have opted to pool the cluster responses for McGurk combinations into one response category (Green and Norrix, 1997; Colin et al., 2002; Aruffo and Shore, 2012). In most other studies that segregated the cluster responses, the combination with the visual labial consonant leading has been the more frequent percept (MacDonald and McGurk, 1978; Massaro and Cohen, 1993; Soto-Faraco and Alsius, 2009; Jiang and Bernstein, 2011), but not always (Hampson et al., 2003; Gil-Carvajal et al., 2020b). The results presented in this study indicate that such variability in the strength and perceived consonant order, particularly for VCV McGurk combinations, could be related to the fact that for natural speech stimuli, different speakers can naturally produce acoustic bursts with different timing

Our findings also suggest that the cluster percept in McGurk combination provides information about how the individual stimulus features are integrated. While the burst and aspiration seem to be sufficient cues for the perception of the consonant /k/ for most SOAs, visual perception of the bilabial /p/ seems to be a sufficient cue for hearing the consonant /p/. A combination response then arises when both of these cues are strong, whereas the order of the consonants in the cluster percept depends on the temporal organization of these acoustic features (burst and aspiration) and the mouth closing gestures. Finally, the experimental paradigm in this study further revealed the robustness of audiovisual speech perception, as the phonetic features were split into different streams across a range of temporal asynchronies and were yet integrated.

Acknowledgements

The authors would like to thank Christophe Savariaux for his support in the recording of the stimuli and fruitful discussions.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The first, third and fourth authors were supported by the Oticon Centre of Excellence for Hearing and Speech Sciences (CHeSS) at the Technical University of Denmark (DTU). The second author was supported by the European Research Council under the European Community's Seventh Framework Program (FP7/2007-2013 Grant Agreement no. 339152, "Speech Unit(e)s").

Declaration of Conflicting Interests

The authors have declared that no competing interests exist.

Supplementary material

Table 4.1: Type and number of tested stimuli

Tested stimuli		N°
Congruent audiovisual /i_i/, /ikpi/, /ipki/, /ip_i/, and /i_pi/		5
Visual-only /i_i/, /ikpi/, /ipki/, /ip_i/, and /i_pi/		5
Auditory-only /i_i/, /ikpi/, and /ipki/		3
Auditory continuum		
	paired at different SOAs	
/i_i/ + burst	-200, -150, -100, -50, 0, 50, 100, 150, 150 and 200 ms	9
Audiovisual continua		
	paired at different SOAs	
Auditory continuum + visual /ip_i/	-200, -150, -100, -50, 0, 50, 100, 150, 150 and 200 ms	9
Auditory continuum + visual /i_pi/	-200, -150, -100, -50, 0, 50, 100, 150, 150 and 200 ms	9
		Total
		40

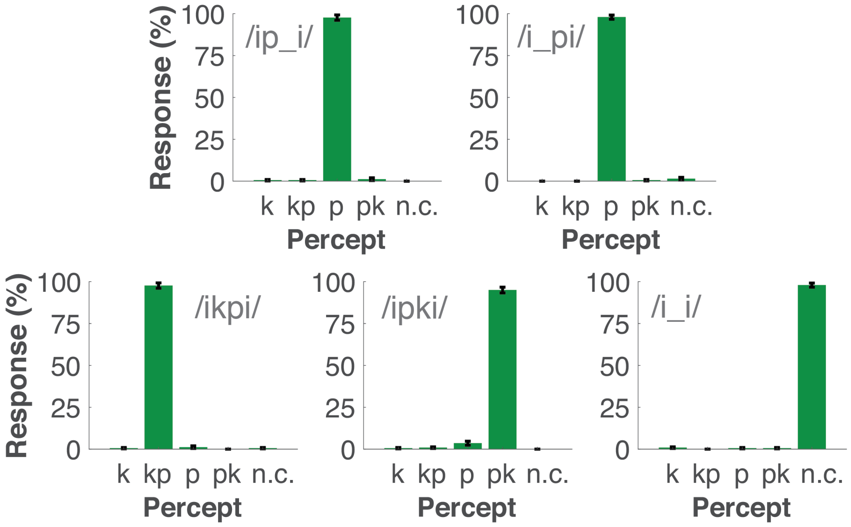


Figure 4.5: Mean response percentages for the congruent audiovisual stimuli. The error bars show the standard error of the mean across subjects.

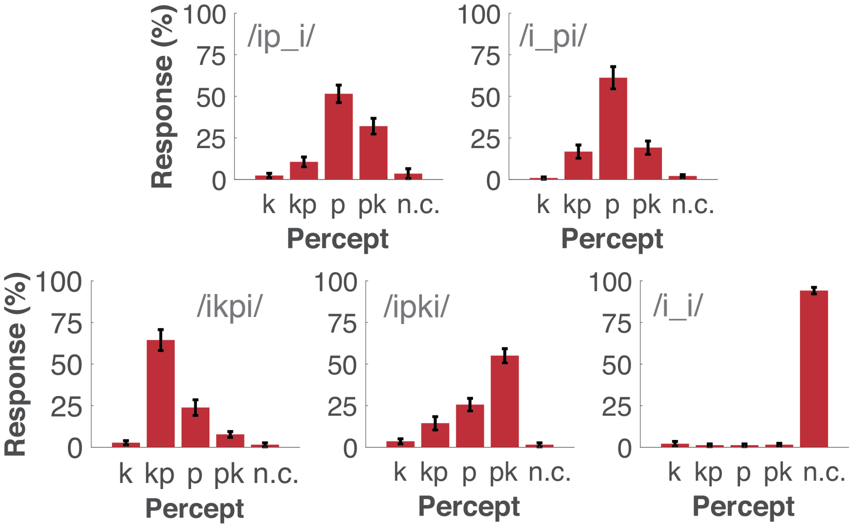


Figure 4.6: Mean response percentages for the visual-only stimuli. The error bars show the standard error of the mean across subjects.

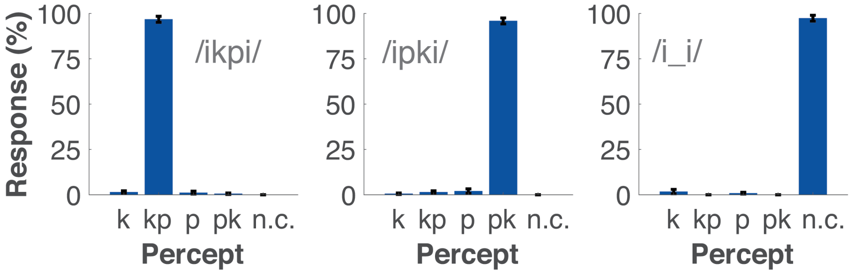


Figure 4.7: Mean response percentages for the auditory-only stimuli. The error bars show the standard error of the mean across subjects.

5

General discussion

This thesis investigated audiovisual speech perception, with a particular focus on providing experimental evidence for a feature-based integration of speech. Three main behavioral studies were carried out. Chapter 2 tested whether the initial and subsequent speech features of consonant segments are integrated separately. Chapter 3 investigated the effect of timing and syllabic context on the McGurk combination illusion. Finally, Chapter 4 studied whether the perceived consonant order in McGurk combinations could be explained in terms of the temporal differences of the audiovisual speech features. In the following, the main results of these studies are summarized and discussed.

5.1 Overall summary

The purpose of the study presented in Chapter 2 was to investigate the integration of audiovisual speech features in consonant segments. The results of this study supported the sequential speech cues hypothesis, which postulates that the initial and subsequent audiovisual speech features of consonant segments are integrated separately. This was demonstrated for the congruent and incongruent audiovisual pairings of the disyllables /aba/, /ada/, /aga/, /abga/, and /adga/. The perception of such audiovisual stimuli led to combination illusions, as well as novel partial fusions and visual dominance illusions. For example, auditory /abga/ dubbed onto visual /aga/ produced mostly /adga/ responses, indicating a partial fusion illusion between the initial auditory consonant /b/ and the initial visual gesture for /g/, while the subsequent auditory consonant was unaffected. Moreover, auditory /aba/ dubbed onto visual /abga/ was mostly perceived as /abda/, reflecting a partial fusion between the subsequent auditory segment for /b/ and the subsequent visual consonant /g/. Thus, the McGurk effect occurred in the initial or subsequent segments of the stimuli. These findings suggested the existence of sequential audiovisual features that are integrated separately, which could underlie the processing of McGurk stimuli. Interestingly, the results also showed an audiovisual enhancement of speech perception, not only for congruent speech stimuli, as has been shown in previous studies (i.e., Sumby and Pollack, 1954; MacLeod and Summerfield, 1987; Arnold and Hill, 2001), but also for partly incongruent audiovisual stimuli. The effect could be attributed to the prominence of the visual closing gesture that enhanced the perception of the initial (bilabial) auditory consonant, whereas the subsequent consonant of the cluster was unaffected.

The study presented in Chapter 3 further investigated the integration of audiovisual speech features in combination illusions. In this variant of the McGurk effect, it was unclear how the auditory and visual consonants are integrated to produce responses that reflect one perceived consonant order or the reverse order. To investigate whether this phenomenon depended on the temporal difference between the auditory and visual components, the effect of the cross-modal timing on the combination illusion was tested in two ways. First, the audiovisual SOA was varied to produce McGurk VCV combination stimuli in which the cross-modal timing of the vowels and consonants was affected, as was done in previous studies (Massaro and Cohen, 1993; Soto-Faraco and Alsius, 2007; Soto-Faraco and Alsius, 2009). Second, only the cross-modal timing of the consonants was varied by pairing auditory and visual speech tokens articulated with different syllabic contexts. Thus, the consonant was articulated in the offset (VC_V) of the initial syllable or in the onset (V_CV) of the subsequent syllable. The latter approach was taken in order to minimize the influence of cross-modal timing on the strength of the integration, which occurs when varying audiovisual SOAs.

The results of the study showed that varying audiovisual SOA influenced the strength of the audiovisual integration, which was in agreement with previous studies (Massaro and Cohen, 1993; Soto-Faraco and Alsius, 2009; Soto-Faraco and Alsius, 2009). In contrast, the perceived consonant order was mostly affected by the syllabic context of the combination stimulus. Auditory /ag_a/ dubbed onto visual /a_ba/ was perceived as /agba/, and conversely, auditory /a_ga/ dubbed onto visual /ab_a/ was perceived as /abga/. The findings provided moderate support of the timing hypothesis, since varying the syllabic context, and hence the cross-modal timing of the consonants, influenced the perceived consonant order. However, the similarity of the unisensory components also influenced the responses, as they were sometimes confused with cluster articulations. These perceptual confusions were likely caused by the intersyllabic silence of the articulation, and therefore, combination stimuli using VC and CV syllables were also tested. The unisensory components of these monosyllabic stimuli were unambiguously perceived, and the obtained combination illusions provided further evidence for the effect of the syllabic context on the perceived consonant order. Irrespective of the audiovisual SOA, CV combination stimuli were most often perceived with the visual consonant leading the auditory, whereas VC combinations were more frequently perceived with the auditory consonant leading the visual. Crucially, the audiovisual window of integration depended on the syllabic context, favouring either visual-leads for CV combination stimuli or visual-lags for VC combinations. Notably, articulatory constraints imposed by the visual mouth gestures accounted for the perceived consonant order in VC and CV combination illusions, as well as for the asymmetry of the audiovisual window of integration.

The results of Chapter 3 suggested a strong effect of the syllabic context on the perceived consonant order. However, it remained unclear whether the phenomenon could be explained by a feature-based approach, and what audiovisual speech features were most important for the effect. An indication for such feature-based processing was found in the study of Green and Norrix (1997), which showed that the proportion of combination illusions substantially decreased when the release burst and aspiration were removed. Furthermore, the study of Colin et al. (2002) showed that the proportion of combination illusions increased when the stimuli were generated with unvoiced consonants (i.e., /k/), which generally contain acoustic release bursts with higher intensity than those found in voiced consonants (i.e., /g/). On this basis, the study presented in Chapter 4 investigated whether the perceived consonant order in McGurk combinations could be varied by manipulating the temporal difference between the release burst and the mouth closing gestures. For this purpose, the auditory stimulus of a McGurk VCV combination was decomposed into two components. One contained the vowels /i_i/ only, and the other contained the acoustic release burst and aspiration for /k/. An auditory continuum was then created by adding the burst and vowels at different temporal alignments. When the burst fell within the vowels, the continuum was perceived as /iki/.

Two additional audiovisual continua were created by dubbing the auditory continuum onto video recordings of the visual articulatory gestures for /ip_i/ or /i_pi/. The continua were perceived mostly as /ikpi/ when the burst was closer to the initial vowel, and as /ipki/ when the burst was closer to the final vowel.

This is consistent with the effect of syllabic context on the perceived consonant order for VC and CV stimuli observed in Chapter 3, which suggests that varying the timing of the burst artificially is equivalent to naturally varying the syllabic context of the articulation. Thus, the results supported the hypothesis that the acoustic burst and aspiration are important for the perception of McGurk combinations, as these stimulus features can affect the strength of the audiovisual integration (Green and Norrix, 1997), but also the perceived consonant order. In this study, seeing the mouth opening before and after the visual consonant may have allowed the burst to be perceived in either order, as the articulatory constraints were weaker than for monosyllabic stimuli. Interestingly, for the audiovisual continuum with /ip_i/ the consonant order reversal occurred earlier than for the continuum with visual /i_pi/, in which the visual bilabial consonant was pronounced later. This indicates that the temporal differences between the articulatory gestures and the acoustic burst influence the perceived order of consonants, providing support for the feature-based integration of audiovisual speech.

5.2 Perceived consonant order in McGurk combinations

At the beginning of this project, it was unknown which audiovisual stimulus features affected the perceived consonant order of the McGurk combination illusion. It was also unclear why the visual bilabial consonant was more frequently perceived to lead the non-labial auditory consonant, although not always (MacDonald and McGurk, 1978; Massaro and Cohen, 1993; Walker et al., 1995; Soto-Faraco and Alsius, 2009; Jiang and Bernstein, 2011). Throughout this thesis, the syllabic context was consistently shown to affect the perceived consonant order of the McGurk combination illusion. In Chapter 2, the McGurk combination stimuli contained a visual bilabial consonant in the offset of the initial syllable and a non-labial auditory consonant in the onset of the subsequent syllable. Accordingly, these stimuli were more frequently perceived with the bilabial consonant leading the non-labial consonant.

Chapter 3 provided a novel finding that supports the strong effect of syllabic context on the perceived consonant order of McGurk combinations. CV syllables were mostly perceived with the labial visual consonant leading the non-labial, which was in agreement with previous findings (Massaro and Cohen, 1993; Soto-Faraco and Alsius, 2007; Soto-Faraco and Alsius, 2009). In contrast, VC syllables were mostly perceived with the reverse consonant order. These percepts could not be explained by the temporal differences between the auditory and visual consonants only, since the consonants were consistently perceived in the same order even after varying audiovisual SOA. Perceptual similarity of unisensory components to cluster articulations did not explain the results either, as the unisensory components were unambiguously perceived. However, articulatory constraints imposed by the bilabial visual gestures on the audiovisual integration provided a likely explanation. For CV syllables there is a strong bilabial constraint at the beginning of the utterance, as the auditory consonant could only be naturally produced after the bilabial opening gesture. Conversely, for VC syllables there is a strong bilabial constraint at the end, which leads to the reverse response. Therefore, in these monosyllabic stimuli, the effect of cross-modal timing on the perceived consonant order could have been overruled by the strong perceptual evidence imposed by the initial or final articulatory constraints.

Notably, the results obtained for VC syllables were partly consistent with the combination illusions reported by Hampson et al. (2003). In their study, the reverse order of consonants was found to be the most frequent percept for VC McGurk combinations, but only when adding a visual lag to the stimulus. The difference between studies could be related to a small variation in the articulation of the consonants, also known as allophonic variation, which is not uncommon in natural speech (Roach, 2000; Pisoni and Remez, 2008). While in the study of Chapter 3 the bilabial consonant was unreleased to provide a strong visual constraint (i.e., the mouth was kept closed in the syllable offset), in the study of Hampson et al. (2003), the consonant might have been produced with a strong release gesture. A prominent visual bilabial gesture could have weakened the articulatory constraints for VC syllables, allowing the auditory consonant to be integrated in either order when the audiovisual SOA provided enough perceptual evidence.

Chapter 4 further investigated the influence of syllabic context on the perceived consonant order, indicating that the effect was driven by the timing of the audiovisual speech features. Varying the timing of the consonant burst relative to the visual gesture had the same effect on the perceived consonant order as changing the syllabic context of VCV McGurk combinations in Chapter 3. This suggests that the reason for the occurrence of consonant order reversals is that naturally changing the articulation affects the timing of the audiovisual speech features. This indicates that the perceived consonant order is indeed informative of the underlying nature of the audiovisual integration process.

5.3 The role of cross-modal timing in audiovisual speech perception

All three studies presented in this thesis showed an effect of cross-modal timing on audiovisual integration of speech. In Chapters 3 and 4, the audiovisual SOA was found to modulate the strength of integration for McGurk stimuli. This is in agreement with previous studies that reported an influence of audiovisual SOA on the rate of McGurk fusions and combinations (Massaro and Cohen, 1993; Van Wassenhove et al., 2007; Soto-Faraco and Alsius, 2009).

Additionally, Chapter 2 showed another effect of timing in audiovisual integration, as the audiovisual speech features of consonant segments were integrated sequentially. This was found for disyllabic audiovisual stimuli consisting of consonant clusters or single consonants in the medial position. Thus, the cross-modal timing seems to determine whether the audiovisual speech features are integrated within the initial or subsequent syllable. Similarly, in Chapter 3, varying the timing of the release burst changed the audiovisual percept from /ikpi/ to /ipki/, hence reflecting the integration of the acoustic burst in the initial or subsequent syllable. These findings are consistent with previous studies that indicated that visual speech is integrated with auditory speech at the average syllabic rate (Arai and Greenberg, 1997; Greenberg, 2006; Venezia et al., 2016), which roughly corresponds to the reported duration of the audiovisual temporal window of integration ($\sim$ 200 ms, Van Wassenhove et al., 2007). Consequently, this suggests that visual speech could indeed be integrated with the auditory speech features at a syllabic rate.

Another important outcome of this thesis is that the asymmetry of the audiovisual temporal window of integration was found in opposite directions for VC and CV McGurk combinations. For CV stimuli, the window favoured visual-lead asynchronies, which is consistent with the reported greater tolerance of audiovisual integration for leads of the visual component (Massaro and Cohen, 1993; Van Wassenhove et al., 2007; Soto-Faraco and Alsius, 2009). In contrast, for VC stimuli, the window favoured lags of the visual stimulus. This novel finding is consistent with a flexible view of the temporal window of integration, evidenced by the fact that the perceptual measure as well as the methodology used to derive the window can influence its width (Soto-Faraco and Alsius, 2009). The opposite asymmetry of the window of integration found here appears consistent with the finding that natural audiovisual speech does not necessarily always exhibit a lead of the visual component (Schwartz and Savariaux, 2014). Instead, it can span a range of temporal asynchronies that includes both leads and lags of the visual component. If the asymmetry of the window of integration depends on the natural timing of audiovisual speech, then it would be expected for some syllabic contexts to show a visual lag advantage for integration. Moreover, the articulatory constraints could also explain the asymmetry of the window, by imposing strong perceptual evidence for the "correct" timing of the audiovisual components, which would be the opposite for VC and CV stimuli.

5.4 Evidence for the feature-based integration of audiovisual speech

Chapter 2 set the initial framework for a feature-based theory of audiovisual integration of speech. The results obtained in this study indicated the existence of audiovisual speech features that are integrated sequentially, which gives rise to the distinct illusory percepts in the McGurk effect. Fusion illusions occur when the audiovisual speech features of the initial and subsequent segments of consonants are integrated into one and the same consonant. In contrast, in combination illusions the initial and subsequent speech features are integrated separately into different consonants, resulting in cluster responses. The same theory would apply to monosyllabic McGurk stimuli shown in previous investigations (Massaro and Cohen, 1993; Van Wassenhove et al., 2007; Soto-Faraco and Alsius, 2009), for which the initial and subsequent features would occur in the same stage (i.e., during the opening or closing mouth gestures).

Chapter 4 provided further evidence for a feature-based integration of audiovisual speech. For McGurk combinations, the visual place information was sufficient for hearing a bilabial consonant /p/, while the release burst and aspiration were sufficient to provide the perception of the velar consonant /k/. The perceived consonant order was then determined by the temporal differences between the release burst and aspiration and the visual bilabial gesture. Such perceptual process could also be mediated by articulatory constraints on audiovisual integration, as well as by the similarity of the unisensory components to possible additional perceptual interpretations, as shown in Chapter 3.

The experimental evidence provided in this thesis establishes the fundament for a quantitative model of audiovisual speech, and constitutes a valuable foundation for modelling challenging aspects of audiovisual speech perception. For example, defining the response categories when modelling McGurk combination stimuli is a difficult task. Here, it has been shown that the initial and subsequent audiovisual speech features are integrated separately. Accordingly, McGurk combinations could be modelled by considering the perceived initial consonant and the subsequent consonant separately. This could limit a potential quantitative model to response boundaries that contain single consonants only, which could decrease the model complexity substantially.

5.5 Consistency of the results

The use of a closed-set response paradigm may have increased the proportion of McGurk responses in all studies presented in this thesis. However, the great consistency of the effects across all three studies support the robustness of the results and the conclusions derived from them. In fact, the effects held across different vocalic contexts (/a/ in Chapters 2 and 3, and /i/ in Chapter 4), across different syllabic contexts (VC and CV in Chapter 3, VCV in Chapters 2, 3 and 4, and VCCV in Chapter 2), across different talkers (one female and two males, one talker for each study), and even after controlling for the native language of the talker and listeners (only native French speakers in Chapter 4). The generally strong McGurk illusion obtained throughout this thesis may have in part been due to the use of a metronome as an auditory reference during the recordings, which could have decreased the variability across utterances. The inclusion of a metronome could contribute to more controlled McGurk stimuli recordings and serve as a tool for standardizing the stimuli descriptions (i.e., by reporting the speaking rate in BPM), which could improve the reproducibility across studies.

5.6 Perspectives

All experiments presented in this thesis were conducted in normal-hearing young adult subjects. Further research would then be needed to establish whether the hypotheses investigated here hold for hearing-impaired listeners. Due to the degraded audition, hearing-impaired listeners tend to weight more strongly the visual place cues than the acoustic place cues (Rouger et al., 2007).

This influences the perception of the McGurk effect, in which visually-driven responses tend to be more frequently reported than for normal-hearing controls (Schorr et al., 2005; Rouger et al., 2007; Desai et al., 2008). In the case of the McGurk combination illusion, the same articulatory constraints would apply for hearing-impaired listeners as these are mostly dependent on the visual perception of the mouth movements. However, a reduced frequency of combination illusions would be expected due to the increased difficulty in perceiving the release burst. The perceived consonant order in McGurk combinations for different types of hearing impairment could then be studied to determine loss of sensitivity to temporal differences between the audiovisual speech features. Moreover, although hearing-impaired listeners could exhibit similar asynchrony detection for speech signals than normal-hearing listeners (Baskent and Bazo, 2011), a broader temporal audiovisual window of integration has been reported for hearing aid users with non-speech signals (Gieseler et al., 2018). It would then be interesting to determine whether the hearing loss, or the use of hearing assistive devices could influence the visual lag advantage for audiovisual integration found for VC McGurk combinations.

Chapter 2 showed an audiovisual enhancement of the bilabial component of acoustic consonant clusters. Further studies could be focused on assessing the audiovisual enhancement of audiovisual speech features in more realistic listening situations. This could be done, for instance, by adding different degrees of reverberation and noise to the acoustic components of the stimuli. Under these challenging listening conditions, the perception of both bilabial and non-labial consonants should be facilitated by seeing the talker, as the acoustic speech signal for both types of consonant would be degraded. However, a greater facilitation would be expected for bilabial consonants, as their place information tends to be acoustically weak and visually strong. Research in this direction could provide a better understanding of the interaction of audiovisual speech features in challenging environments. Eventually, this could lead to applications, such as the optimization of hearing-aid signal processing for restoring the acoustic features that are not compensated for by vision in face-to-face settings. Such implementations, however, should also consider additional higher-order effects (i.e., semantic, linguistic, attention, etc.) that were outside the scope of this thesis.

Bibliography

Alsius, A., M. Paré, and K. G. Munhall (2017). "Forty Years after Hearing Lips and Seeing Voices: The McGurk Effect Revisited". In: *Multisensory Research* 31.1-2, pp. 111–144. DOI: 10.1163/22134808-00002565.

Arai, T. and S. Greenberg (1997). "The temporal properties of spoken Japanese are similar to those of English". In: *Fifth european conference on speech communication and technology*.

Arai, T., E. Iwagami, and E. Yanagisawa (2017). "Seeing closing gesture of articulators affects speech perception of geminate consonants". In: *Journal of the Acoustical Society of America* 141.3, EL319–EL325. DOI: 10.1121/1.4978343.

Arnold, P. and F. Hill (2001). "Bisensory augmentation: A speechreading advantage when speech is clearly audible and intact". In: *British Journal of Psychology* 92.2, pp. 339–355. DOI: 10.1348/000712601162220.

Arnold, P. and A. Köpsel (1996). "Lipreading, reading and memory of hearing and hearing-impaired children". In: *Scandinavian Audiology* 25.1, pp. 13–20. DOI: 10.3109/01050399609047550.

Aruffo, C. and D. I. Shore (2012). "Can you McGurk yourself? Self-face and self-voice in audiovisual speech". In: *Psychonomic Bulletin and Review* 19.1, pp. 66–72. DOI: 10.3758/s13423-011-0176-8.

Baart, M., A. C. Lindborg, and T. S. Andersen (2017). "Electrophysiological evidence for differences between fusion and combination illusions in audiovisual speech perception". In: *European Journal of Neuroscience* 46.10, pp. 2578–2583. DOI: 10.1111/ejn.13734.

Baskent, D. and D. Bazo (2011). "Audiovisual asynchrony detection and speech intelligibility in noise with moderate to severe sensorineural hearing impairment". In: *Ear and Hearing* 32.5, pp. 582–592. DOI: 10.1097/AUD.0b013e31820fca23.

Bastian, J (1962). "Silent intervals as closure cues in the perception of stops". In: *Haskins Laboratories, Speech Research and Instrumentation* 9.

Basu Mallick, D., J. F. Magnotti, and M. S. Beauchamp (2015). “Variability and stability in the McGurk effect: contributions of participants, stimuli, time, and response type”. In: *Psychonomic Bulletin and Review* 22.5, pp. 1299–1307. DOI: `10.3758/s13423-015-0817-4`.

Binnie, C. A., A. A. Montgomery, and P. L. Jackson (1974). “Auditory and visual contributions to the perception of consonants”. In: *Journal of Speech and Hearing Research* 17.4, pp. 619–630. DOI: `10.1044/jshr.1704.619`.

Brancazio, L. (2004). “Lexical influences in audiovisual speech perception”. In: *Journal of Experimental Psychology: Human Perception and Performance* 30.3, pp. 445–463. DOI: `10.1037/0096-1523.30.3.445`.

Brancazio, L., J. L. Miller, and M. A. Paré (2003). “Visual influences on the internal structure of phonetic categories”. In: *Perception and Psychophysics* 65.4, pp. 591–601. DOI: `10.3758/BF03194585`.

Browman, C. P. and L. Goldstein (1990). “Tiers in articulatory phonology, with some implications for casual speech”. In: *Papers in Laboratory Phonology* 1.3, 341–376. DOI: `10.1017/CBO9780511627736.019`.

Byrd, D. (1992). “Perception of assimilation in consonant clusters - a gestural model”. In: *Phonetica* 49.1, pp. 1–24.

Byrd, D. (1996). “Influences on articulatory timing in consonant sequences”. In: *Journal of Phonetics* 24.2, pp. 209–244. DOI: `10.1006/jpho.1996.0012`.

Chandrasekaran, C., A. Trubanova, S. Stillittano, A. Caplier, and A. A. Ghazanfar (2009). “The natural statistics of audiovisual speech”. In: *Plos Computational Biology* 5.7. Ed. by K. J. Friston, e1000436. DOI: `10.1371/journal.pcbi.1000436`.

Colin, C., M. Radeau, P. Deltenre, D. Demolin, and A. Soquet (2002). “The role of sound intensity and stop-consonant voicing on McGurk fusions and combinations”. In: *European Journal of Cognitive Psychology* 14.4, pp. 475–491. DOI: `10.1080/09541440143000203`.

Desai, S., G. Stickney, and F. G. Zeng (2008). “Auditory-visual speech perception in normal-hearing and cochlear-implant listeners”. eng. In: *Journal of the Acoustical Society of America* 123.1, pp. 428–440. DOI: `10.1121/1.2816573`.

Dodd, B (1977). “The role of vision in the perception of speech”. In: *Perception* 6.1, pp. 31–40, 31–40.

Dorman, M. F., M Studdert-Kennedy, and L. J. Raphael (1977). “Stop-consonant recognition: Release bursts and formant transitions as functionally equivalent, context-dependent cues”. In: *Perception and Psychophysics* 22.2, pp. 109–122. DOI: `10.3758/bf03198744`.

Dorman, M. and L. Raphael (1980). “Distribution of acoustic cues for stop consonant place of articulation in VCV syllables”. In: *Journal of the Acoustical Society of America* 67.4, pp. 1333–1335. DOI: `10.1121/1.384186`.

Dorman, M., L. Raphael, and A. Liberman (1979). “Some experiments on the sound of silence in phonetic perception”. In: *Journal of the Acoustical Society of America* 65.6, pp. 1518–1532. DOI: `10.1121/1.382916`.

Dowell, R. C., L. F. A. Martin, Y. C. Tong, G. M. Clark, P. M. Seligman, and J. F. Patrick (1982). “A 12-consonant confusion study on a multiple-channel cochlear implant patient”. In: *Journal of Speech, Language, and Hearing Research* 25.4, pp. 509–516. DOI: `10.1044/jshr.2504.509`.

Erber, N. P. (1972). “Auditory, visual, and auditory-visual recognition of consonants by children with normal and impaired hearing”. eng. In: *Journal of Speech and Hearing Research* 15.2, pp. 413–422. DOI: `10.1044/jshr.1502.413`.

Erber, N. P. (1975). “Auditory-visual perception of speech”. In: *Journal of Speech and Hearing Disorders* 40.4, pp. 481–492. DOI: `10.1044/jshd.4004.481`.

Gieseler, A., M. A. S. Tahden, C. M. Thiel, and H. Colonius (2018). “Does hearing aid use affect audiovisual integration in mild hearing impairment?” In: *Experimental Brain Research* 236.4, pp. 1161–1179, 1161–1179. DOI: `10.1007/s00221-018-5206-6`.

Gil-Carvajal, J. C. and T. S. Andersen (2020a). “Audiovisual integration of consonant clusters”. In: *Manuscript in preparation.*

Gil-Carvajal, J. C., T. Dau, and T. S. Andersen (2020b). “Order matters: timing and syllabic context influence perceived consonant order in audiovisual speech”. In: *Submitted to Journal of Experimental Psychology: Human Perception and Performance.*

Gil-Carvajal, J. C., J.-L. Schwartz, T. Dau, and T. S. Andersen (2019). “Feature-based audiovisual speech integration of multiple streams”. In: *Proceedings of the International Symposium on Auditory and Audiological Research.* Vol. 7, pp. 333–340.

Gil-Carvajal Juan C., S. J.-L., T. Dau, and T. S. Andersen (2020c). “Feature-based audiovisual speech integration of multiple sequences”. In: *Trends in Hearing (under review).*

Goldstein, L. and C. A. Fowler (2003). “Articulatory phonology: A phonology for public language use”. In: *Phonetics and phonology in language comprehension and production: Differences and similarities*, pp. 159–207.

Grant, K. W. and P. F. Seitz (2000). “The use of visible speech cues for improving auditory detection of spoken sentences”. In: *Journal of the Acoustical Society of America* 108.3, pp. 1197–1208. DOI: `10.1121/1.1288668`.

Grant, K. W., B. E. Walden, and P. F. Seitz (1998). “Auditory-visual speech recognition by hearing-impaired subjects: Consonant recognition, sentence recognition, and auditory-visual integration”. In: *Journal of the Acoustical Society of America* 103.5 I, pp. 2677–2690. DOI: `10.1121/1.422788`.

Green, K. P. and P. K. Kuhl (1991). “Integral processing of visual place and auditory voicing information during phonetic perception”. In: *Journal of Experimental Psychology: Human Perception and Performance* 17.1, pp. 278–288. DOI: `10.1037//0096-1523.17.1.278`.

Green, K. P., P. K. Kuhl, A. N. Meltzoff, and E. B. Stevens (1991). “Integrating speech information across talkers, gender, and sensory modality - female faces and male voices in the McGurk effect”. In: *Perception and Psychophysics* 50.6, pp. 524–536. DOI: `10.3758/bf03207536`.

Green, K. P. and L. W. Norrix (1997). “Acoustic cues to place of articulation and the McGurk effect”. In: *Journal of Speech, Language, and Hearing Research* 40.3, pp. 646–665. DOI: `10.1044/jslhr.4003.646`.

Greenberg, J. H. (1965). “Some generalizations concerning initial and final consonant sequences”. In: *Linguistics* 3.18, pp. 5–34.

Greenberg, S. (2006). “A multi-tier framework for understanding spoken language”. In: *Listening to speech: An auditory perspective*, pp. 411–433.

Halle, M, G. Hughes, and J. Radley (1957). “Acoustic properties of stop consonants”. eng. In: *Journal of the Acoustical Society of America* 29.1, pp. 107–116. DOI: `10.1121/1.1908634`.

Hamilton, R. H., J. T. Shenton, and H. B. Coslett (2006). “An acquired deficit of audiovisual speech processing”. In: *Brain and Language* 98.1, pp. 66–73. DOI: `10.1016/j.bandl.2006.02.001`.

Hampson, M., F. H. Guenther, M. A. Cohen, and A. Nieto-Castanon (2003). *Changes in the McGurk Effect across phonetic contexts.* Tech. rep. Boston University Center for Adaptive Systems, Department of Cognitive, and Neural Systems.

Hardcastle, W. J. and P. Roach (1979). “An instrumental investigation of coarticulation in stop consonant sequences”. In: *Current issues in the phonetic sciences* 9, pp. 531–540.

Helfer, K. S. (1997). “Auditory and auditory-visual perception of clear and conversational speech”. In: *Journal of Speech, Language, and Hearing Research* 40.2, pp. 432–443. DOI: `10.1044/jslhr.4002.432`.

Ijsseldijk, F. J. (1992). “Speechreading performance under different conditions of video image repetition and speech rate”. In: *Journal of Speech and Hearing Research* 35.2, pp. 466–471, 466–471.

Jiang, J. and L. E. Bernstein (2011). “Psychophysics of the McGurk and other audiovisual speech integration effects”. In: *Journal of Experimental Psychology: Human Perception and Performance* 37.4, pp. 1193–1209. DOI: `10.1037/a0023100`.

Jordan, T. R. and S. M. Thomas (2001). “Effects of horizontal viewing angle on visual and audiovisual speech recognition”. In: *Journal of Experimental Psychology: Human Perception and Performance* 27.6, pp. 1386–1403. DOI: `10.1037//0096-1523.27.6.1386`.

Kaiser, A. R., K. I. Kirk, L. Lachs, and D. B. Pisoni (2012). “Talker and lexical effects on audiovisual word recognition by adults with cochlear implants”. In: *Journal of Speech, Language, and Hearing Research : Jslhr* 46.2, pp. 390–404.

Kerswill, P. E. (1985). “A sociophonetic study of connected speech processes in Cambridge English: an outline and some results”. In: *Cambridge papers in phonetics and experimental linguistics* 4.1987, pp. 25–49.

Kessler, B. and R. Treiman (1997). “Syllable structure and the distribution of phonemes in English syllables”. eng. In: *Journal of Memory and Language* 37.3, pp. 295–311. DOI: `10.1006/jmla.1997.2522`.

Kuhl, P. K. and A. N. Meltzoff (1982). “The bimodal perception of speech in infancy”. In: *Science* 218.4577, pp. 1138–1141.

Lahiri, A. and J. Hankamer (1988). “The timing of geminate consonants”. In: *Journal of Phonetics* 16.3, pp. 327–338. DOI: `10.1016/S0095-4470(19)30506-6`.

Legerstee, M (1990). “Infants use multimodal information to imitate speech sounds”. In: *Infant Behavior and Development* 13.3, pp. 343–354. DOI: `10.1016/0163-6383(90)90039-B`.

Li, F., A. Menon, and J. B. Allen (2010). “A psychoacoustic method to find the perceptual cues of stop consonants in natural speech”. In: *Journal of the Acoustical Society of America* 127.4, pp. 2599–2610. DOI: `10.1121/1.3295689`.

MacDonald, J. and H. McGurk (1978). “Visual influences on speech perception processes”. In: *Perception and Psychophysics* 24.3, pp. 253–257. DOI: `10.3758/bf03206096`.

MacDonald, J. (2018). “Hearing lips and seeing voices: The origins and development of the ’McGurk effect’ and reflections on audio-visual speech perception over the last 40 years”. In: *Multisensory Research* 31.1-2, pp. 7–18. DOI: `10.1163/22134808-00002548`.

MacDonald, J., S. Andersen, and T. Bachmann (2000). “Hearing by eye: How much spatial degradation can be tolerated?” In: *Perception* 29.10, pp. 1155–1168. DOI: `10.1068/p3020`.

MacLeod, A. and Q. Summerfield (1987). “Quantifying the contribution of vision to speech perception in noise”. In: *British Journal of Audiology* 21.2, pp. 131–141. DOI: `10.3109/03005368709077786`.

Magnotti, J. F., D. Basu Mallick, G. Feng, B. Zhou, W. Zhou, and M. S. Beauchamp (2015). “Similar frequency of the McGurk effect in large samples of native Mandarin Chinese and American English speakers”. eng. In: *Experimental Brain Research* 233.9, pp. 2581–2586. DOI: `10.1007/s00221-015-4324-7`.

Marassa, L. K. and C. R. Lansing (1995). “Visual word recognition in 2 facial motion conditions - full face versus lips-plus-mandible”. In: *Journal of Speech and Hearing Research* 38.6, pp. 1387–1394.

Massaro, D. W. and M. M. Cohen (1993). “Perceiving asynchronous bimodal speech in consonant-vowel and vowel syllables”. In: *Speech Communication* 13.1-2, pp. 127–134. DOI: `10.1016/0167-6393(93)90064-r`.

Massaro, D. W., M. M. Cohen, and P. M. Smeele (1996). “Perception of asynchronous and conflicting visual and auditory speech”. In: *Journal of the Acoustical Society of America* 100.3, pp. 1777–1786. DOI: `10.1121/1.417342`.

Massaro, D. W. and S. E. Palmer Jr (1998). *Perceiving talking faces: From speech perception to a behavioral principle.* Mit Press.

McGurk, H. and J. MacDonald (1976). “Hearing Lips and seeing voices”. In: *Nature* 264.5588, pp. 746–748. DOI: `10.1038/264746a0`.

McLeod, S., J. Van Doorn, and V. A. Reed (2001). “Normal acquisition of consonant clusters”. In: *American Journal of Speech-Language Pathology.*

Menon, K. M. N., P. V. S. Rao, and R. B. Thosar (1974). "Formant transitions and stop consonant perception in syllables". In: *Language and Speech* 17.JAN-M, pp. 27–46.

Miller, G. A. and P. E. Nicely (1955). "An analysis of perceptual confusions among some English consonants". In: *Journal of the Acoustical Society of America* 27.2, pp. 338–352. DOI: 10.1121/1.1907526.

Munhall, K. G., P. Gribble, L. Sacco, and M. Ward (1996). "Temporal constraints on the McGurk effect". In: *Perception and Psychophysics* 58.3, pp. 351–362. DOI: 10.3758/BF03206811.

Munhall, K. G. and E. Vatikiotis-Bateson (2013). "The moving face during speech communication". In: *Hearing Eye II: Advances in the Psychology of Speechreading and Auditory-visual Speech*, pp. 123–139. DOI: 10.4324/9780203098752-14.

Nath, A. R. and M. S. Beauchamp (2012). "A neural basis for interindividual differences in the McGurk effect, a multisensory speech illusion". In: *Neuroimage* 59.1, pp. 781–787. DOI: 10.1016/j.neuroimage.2011.07.024.

Nolan, F. (1992). "The descriptive role of segments: Evidence from assimilation". In: *Papers in laboratory phonology II: Gesture, segment, prosody*, pp. 261–280.

O'Neill, J. J. (1954). "Contributions of the visual components of oral symbols to speech comprehension". In: *Journal of Speech and Hearing Disorders* 19.4, pp. 429–439. DOI: 10.1044/jshd.1904.429.

Peelle, J. E. and M. S. Sommers (2015). "Prediction and constraint in audiovisual speech perception". In: *Cortex* 68.Sp. Iss. SI, pp. 169–181. DOI: 10.1016/j.cortex.2015.03.006.

Pickett, E. R., S. E. Blumstein, and M. W. Burton (1999). "Effects of speaking rate on the singleton/geminate consonant contrast in Italian". In: *Phonetica* 56.3–4, pp. 135–157.

Pisoni, D. B. and R. E. Remez (2008). *The handbook of speech perception.* Ed. by D. B. Pisoni. John Wiley and Sons, pp. 1–708. DOI: 10.1002/9780470757024.

Preminger, J. E., H. B. Lin, M. Payen, and H. Levitt (1998). "Selective visual masking in speechreading". In: *Journal of Speech, Language, and Hearing Research* 41.3, pp. 564–575. DOI: 10.1044/jslhr.4103.564.

Reisberg, D. (2010). *Cognition: Exploring the science of the mind.* WW Norton & Company Incorporated.

Reisberg, D., J. Mclean, and A. Goldfield (1987). “Easy to hear but hard to understand: A lip-reading advantage with intact auditory stimuli.” In: *Hearing by eye: The psychology of lip-reading*, pp. 97–113.

Roach, P. (2000). *English Phonetics and phonology: A practical course Cambridge.*

Rosenblum, L. D. and H. M. Saldaña (1992). “Discrimination tests of visually influence syllables”. In: *Perception and Psychophysics* 52.4, pp. 461–473. DOI: 10.3758/bf03206706.

Rosenblum, L. D. and H. M. Saldaña (1996). “An audiovisual test of kinematic primitives for visual speech perception”. In: *Journal of Experimental Psychology: Human Perception and Performance* 22.2, pp. 318–331. DOI: 10.1037/0096-1523.22.2.318.

Ross, L. A., D. Saint-Amour, V. M. Leavitt, D. C. Javitt, and J. J. Foxe (2007). “Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments”. In: *Cerebral Cortex* 17.5, pp. 1147–1153. DOI: 10.1093/cercor/bhl024.

Rouger, J., S. Lagleyre, B. Fraysse, S. Deneve, O. Deguine, and P. Barone (2007). “Evidence that cochlear-implanted deaf patients are better multisensory integrators”. In: *Proceedings of the National Academy of Sciences of the United States of America* 104.17, pp. 7295–7300. DOI: 10.1073/pnas.0609419104.

Sams, M et al. (1991). “Seeing speech: visual information from lip movements modifies activity in the human auditory cortex”. In: *Neuroscience Letters* 127.1, pp. 141–5, 141–145. DOI: 10.1016/0304-3940(91)90914-f.

Saporta, S. (1955). “Frequency of consonant clusters”. In: *Language* 31.1, pp. 25–30. DOI: 10.2307/410889.

Scheinberg, J. S. (1980). “Analysis of speechreading cues using an interleaved technique”. In: *Journal of Communication Disorders* 13.6, pp. 489–492. DOI: 10.1016/0021-9924(80)90048-9.

Schorr, E. A., N. A. Fox, V. Van Wassenhove, and E. I. Knudsen (2005). “Auditory-visual fusion in speech perception in children with cochlear implants”. In: *Proceedings of the National Academy of Sciences of the United States of America* 102.51, pp. 18748–18750. DOI: 10.1073/pnas.0508862102,10.1073/pnas.0508862102.

Schwartz, J. L. (2010). “A reanalysis of McGurk data suggests that audiovisual fusion in speech perception is subject-dependent”. In: *Journal of the Acoustical Society of America* 127.3, pp. 1584–1594. DOI: 10.1121/1.3293001.

Schwartz, J. L., F. Berthommier, and C. Savariaux (2004). “Seeing to hear better: Evidence for early audio-visual interactions in speech identification”. In: *Cognition* 93.2, B69–B78, B69–B78. DOI: `10.1016/j.cognition.2004.01.006`.

Schwartz, J. L. and C. Savariaux (2014). “No, there is no 150 ms lead of visual speech on auditory speech, but a range of audiovisual asynchronies varying from small audio lead to large audio lag”. In: *Plos Computational Biology* 10.7. Ed. by A. Vatakis, e1003743. DOI: `10.1371/journal.pcbi.1003743`.

Scott, M. and A. Idrissi (2018). “Audiovisual perception of gemination and pharyngealization in Arabic”. In: *Speech Communication* 98, pp. 17–27. DOI: `10.1016/j.specom.2018.01.009`.

Sekiyama K. Tohkura, Y. (1993). “Inter-language differences in the influence of visual cues in speech-perception”. In: *Journal of Phonetics* 21.4, pp. 427–444. DOI: `10.1016/S0095-4470(19)30229-3`.

Smeele, P. M., D. W. Massaro, M. M. Cohen, and A. C. Sittig (1998). “Laterality in visual speech perception”. In: *Journal of Experimental Psychology: Human Perception and Performance* 24.4, pp. 1232–1242. DOI: `10.1037/0096-1523.24.4.1232`.

Soto-Faraco, S. and A. Alsius (2007). “Conscious access to the unisensory components of a cross-modal illusion”. eng. In: *Neuroreport* 18.4, pp. 347–350. DOI: `10.1097/WNR.0b013e32801776f9`.

Soto-Faraco, S. and A. Alsius (2009). “Deconstructing the McGurk-MacDonald illusion”. In: *Journal of Experimental Psychology: Human Perception and Performance* 35.2, pp. 580–587. DOI: `10.1037/a0013483`.

Stevens, K. N. and S. Blumstein (1978). “Invariant cues for place of articulation in stop consonants”. In: *Journal of the Acoustical Society of America* 64.5, pp. 1358–1368. DOI: `10.1121/1.382102`.

Sumby and Pollack (1954). “Visual contribution to speech intelligibility in noise”. In: *Journal of the Acoustical Society of America* 26, pp. 212–215. DOI: `10.1121/1.1907309`.

Summerfield Quentin, M. A. M. M. B. M. (1989). “Lips, teeth, and the benefits of lipreading”. In: *Handbook of research on face processing*, pp. 223–233.

Summerfield, Q (1979). “Use of visual information for phonetic perception”. In: *Phonetica* 36.4-5, pp. 314–331. DOI: `10.1159/000259969`.

Summerfield, Q (1992). “Lipreading and audio-visual speech perception”. In: *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 335.1273, pp. 71–8, 71–78. DOI: 10.1098/rstb.1992.0009.

Thomas, S. M. and T. R. Jordan (2004). “Contributions of oral and extraoral facial movement to visual and audiovisual speech perception”. In: *Journal of Experimental Psychology: Human Perception and Performance* 30.5, pp. 873–888. DOI: 10.1037/0096-1523.30.5.873.

Tiippana, K. (2014). “What is the McGurk effect?” In: *Frontiers in Psychology* 5, p. 725. DOI: 10.3389/fpsyg.2014.00725.

Troille, E., M. A. Cathiard, and C. Abry (2010). “Speech face perception is locked to anticipation in speech production”. In: *Speech Communication* 52.6, pp. 513–524. DOI: 10.1016/j.specom.2009.12.005.

Tye-Murray, N., M. S. Sommers, and B. Spehar (2007). “Audiovisual integration and lipreading abilities of older adults with normal and impaired hearing”. In: *Ear and Hearing* 28.5, pp. 656–668. DOI: 10.1097/AUD.0b013e31812f7185.

Van Wassenhove, V., K. W. Grant, and D. Poeppel (2005). “Visual speech speeds up the neural processing of auditory speech”. In: *Proceedings of the National Academy of Sciences of the United States of America* 102.4, pp. 1181–6, 1181–1186. DOI: 10.1073/pnas.0408949102.

Van Wassenhove, V. (2013). “Speech through ears and eyes: Interfacing the senses with the supramodal brain”. In: *Frontiers in Psychology* 4, Article 388. DOI: 10.3389/fpsyg.2013.00388.

Van Wassenhove, V., K. W. Grant, and D. Poeppel (2007). “Temporal window of integration in auditory-visual speech perception”. In: *Neuropsychologia* 45.3. Ed. by F. P. Micah Murray, pp. 598–607. DOI: 10.1016/j.neuropsychologia.2006.01.001.

Venezia, J. H., S. M. Thurman, W. Matchin, S. E. George, and G. Hickok (2016). “Timing in audiovisual speech perception: A mini review and new psychophysical data”. In: *Attention, Perception, and Psychophysics* 78.2, pp. 583–601. DOI: 10.3758/s13414-015-1026-y.

Walden, B. E., R. A. Prosek, A. A. Montgomery, C. K. Scherr, and C. J. Jones (1977). “Effects of training on the visual recognition of consonants”. In: *Journal of Speech and Hearing Research* 20.1, pp. 130–45, 130–145. DOI: 10.1044/jshr.2001.130.

Walker, S., V. Bruce, and C. Omalley (1995). “Facial identity and facial speech processing: Familiar faces and voices in the McGurk effect”. In: *Perception and Psychophysics* 57.8, pp. 1124–1133. DOI: `10.3758/bf03208369`.

Winn, M. B., A. E. Rhone, M. Chatterjee, and W. J. Idsardi (2013). “The use of auditory and visual context in speech perception by listeners with normal hearing and listeners with cochlear implants”. In: *Frontiers in Psychology* 4, Article 824. DOI: `10.3389/fpsyg.2013.00824`.

Contributions to Hearing Research

Vol. 1: *Gilles Pigasse*, Deriving cochlear delays in humans using otoacoustic emissions and auditory evoked potentials, 2008.
External examiners: Mark Lutman, Stefan Stenfeld

Vol. 2: *Olaf Strelcyk*, Peripheral auditory processing and speech reception in impaired hearing, 2009.
External examiners: Brian Moore, Kathrin Krumbholz

Vol. 3: *Eric R. Thompson*, Characterizing binaural processing of amplitude-modulated sounds, 2009.
External examiners: Michael Akeroyd, Armin Kohlrausch

Vol. 4: *Tobias Piechowiak*, Spectro-temporal analysis of complex sounds in the human auditory system, 2009.
External examiners: Jesko Verhey, Steven van de Par

Vol. 5: *Jens Bo Nielsen*, Assessment of speech intelligibility in background noise and reverberation, 2009.
External examiners: Björn Hagerman, Ejnar Laukli

Vol. 6: *Helen Connor*, Hearing aid amplification at soft input levels, 2010.
External examiners: Inga Holube, Birgitta Larsby

Vol. 7: *Morten Løve Jepsen*, Modeling auditory processing and speech perception in hearing-impaired listeners, 2010.
External examiners: Birger Kollmeier, Ray Meddis

Vol. 8: *Sarah Verhulst*, Characterizing and modeling dynamic processes in the cochlea using otoacoustic emissions, 2010.
External examiners: David Kemp, Stephen Neely

Vol. 9: *Sylvain Favrot*, A loudspeaker-based room auralization system for auditory research, 2010.
External examiners: Bernhard Seeber, Michael Vorländer

Vol. 10: *Sébastien Santurette*, Neural coding and perception of pitch in the normal and impaired human auditory system, 2011.
External examiners: Christopher Plack, Christian Lorenzi

Vol. 11: *Iris Arweiler*, Processing of spatial sounds in the impaired auditory system, 2011.
External examiners: Joost Festen, Jürgen Tchorz

Vol. 12: *Filip Munch Rønne*, Modeling auditory evoked potentials to complex stimuli, 2012.
External examiners: Bob Burkard, Stephen Neely

Vol. 13: *Claus Forup Corlin Jespersgaard*, Listening in adverse conditions: Masking release and effects of hearing loss, 2012.
External examiners: Stuart Rosen, Christian Lorenzi

Vol. 14: *Rémi Decorsière*, Spectrogram inversion and potential applications for hearing research, 2013.
External examiners: Michael Stone, Oded Ghitza

Vol. 15: *Søren Jørgensen*, Modeling speech intelligibility based on the signal-to-noise envelope power ration, 2014.
External examiners: John Culling, Martin Cooke

Vol. 16: *Kasper Eskelund*, Electrophysiological assessment of audiovisual integration in speech perception, 2014.
External examiners: Lawrence Rosenblum, Matthias Gondan

Vol. 17: *Simon Krogholt Christiansen*, The role of temporal coherence in auditory stream segregation, 2014.
External examiners: Shihab Shamma, Guy Brown

Vol. 18: *Márton Marschall*, Capturing and reproducing realistic acoustic scenes for hearing research, 2014.
External examiners: Sascha Spors, Ville Pulkki

Vol. 19: *Jasmina Catic*, Human sound externalization in reverberant environments, 2014.
External examiners: Bernhard Seeber, Steven van de Par

Vol. 20: *Michał Feręczkowski*, Design and evaluation of individualized hearing-aid signal processing and fitting, 2015.
External examiners: Christopher Plack, Enrique Lopez-Poveda

Vol. 21: *Alexandre Chabot-Leclerc*, Computational modeling of speech intelligibility in adverse conditions, 2015.
External examiners: Steven van de Par, John Culling

Vol. 22: *Federica Bianchi*, Pitch representations in the impaired auditory system and implications for music perception, 2016.
External examiners: Ingrid Johnsrude, Christian Lorenzi

Vol. 23: *Johannes Zaar*, Measures and computational models of microscopic speech perception, 2016.
External examiners: Judy Dubno, Martin Cooke

Vol. 24: *Johannes Käsbach*, Characterizing apparent source width perception, 2016.
External examiners: William Whitmer, Jürgen Tchorz

Vol. 25: *Gusztáv Löcsei*, Lateralized speech perception with normal and impaired hearing, 2016.
External examiners: Thomas Brand, Armin Kohlrausch

Vol. 26: *Suyash Narendra Joshi*, Modelling auditory nerve responses to electrical stimulation, 2017.
External examiners: Laurel Carney, Bob Carlyon

Vol. 27: *Henrik Gerd Hassager*, Characterizing perceptual externalization in listeners with normal, impaired and aided-impaired hearing, 2017.
External examiners: Volker Hohmann, Piotr Majdak

Vol. 28: *Richard Ian McWalter*, Analysis of the auditory system via synthesis of natural sounds, speech and music, 2017.
External examiners: Maria Chait, Christian Lorenzi

Vol. 29: *Jens Cubick*, Characterizing the auditory cues for the processing and perception of spatial sounds, 2017.
External examiners: Ville Pulkki, Pavel Zahorik

Vol. 30: *Gerard Encina-Llamas*, Characterizing cochlear hearing impairment using advanced electrophysiological methods, 2017.
External examiners: Roland Schaette, Ian Bruce

Vol. 31: *Christoph Scheidiger*, Assessing speech intelligibility in hearing-impaired listeners, 2018.
External examiners: Enrique Lopez-Poveda, Tim Jürgens

Vol. 32: *Alan Wiinberg*, Perceptual effects of non-linear hearing aid amplification strategies, 2018.
External examiners: Armin Kohlrausch, James Kates

Vol. 33: *Thomas Bentsen*, Computational speech segregation inspired by principles of auditory processing, 2018.
External examiners: Stefan Bleeck, Jürgen Tchorz

Vol. 34: *François Guérit*, Temporal charge interactions in cochlear implant listeners, 2018.
External examiners: Julie Arenberg, Olivier Macherey

Vol. 35: *Andreu Paredes Gallardo*, Behavioral and objective measures of stream segregation in cochlear implant users, 2018.
External examiners: Christophe Micheyl, Monita Chatterjee

Vol. 36: *Søren Fuglsang*, Characterizing neural mechanisms of attention-driven speech processing, 2019.
External examiners: Shihab Shamma, Maarten de Vos

Vol. 37: *Borys Kowalewski*, Assessing the effects of hearing-aid dynamic-range compression on auditory signal processing and perception, 2019.
External examiners: Brian Moore, Graham Naylor

Vol. 38: *Helia Relaño Iborra*, Predicting speech perception of normal-hearing and hearing-impaired listeners, 2019.
External examiners: Ian Bruce, Armin Kohlrausch

Vol. 39: *Axel Ahrens*, Characterizing auditory and audio-visual perception in virtual environments, 2019.
External examiners: Pavel Zahorik, Piotr Majdak

Vol. 40: *Niclas A. Janssen*, Binaural streaming in cochlear implant patients, 2019.
External examiners: Tim Jürgens, Hamish Innes-Brown

Vol. 41: *Wiebke Lamping*, Improving cochlear implant performance through psychophysical measures, 2019.
External examiners: David Landsberger, Dan Gnasia

Vol. 42: *Antoine Favre-Félix*, Controlling a hearing aid by electrically assessed eye-gazed, 2020.
External examiners: Graham Naylor, Jürgen Tchorz

Vol. 43: *Raul Sanchez-Lopez*, Clinical auditory profiling and profile-based hearing-aid fitting, 2020.
External examiners: Judy R. Dubno, Pamela Souza

Vol. 44: *Juan Camilo Gil-Carvajal*, Towards a feature-based theory of audiovisual integration of speech, 2020.
External examiners: Kaisa Tiipana, Salvador Soto-Faraco

The end.

To be continued…

Speech perception is facilitated by seeing the mouth movements of the talker. The visible mouth movements of the talker can also modify the auditory phonetic perception, leading to audiovisual illusions. This is shown in the McGurk effect, which occurs when a speech sound is presented simultaneously with incongruent visual mouth movements corresponding to another speech token. While the McGurk fusion illusion leads to the perception of a consonant that is different from the presented auditory and visual consonants, in the McGurk combination illusion, the two consonants are perceived. Despite decades of research in audiovisual speech, it has remained unclear why some audiovisual stimuli elicit McGurk fusions while others produce McGurk combinations, and which are the stimulus features that affect the perceived consonant order in the latter case.

This PhD project investigated audiovisual speech perception, with a particular focus on providing experimental evidence for a featured-based model of audiovisual integration of speech. The results revealed novel McGurk illusions, which suggest the existence of sequential audiovisual features in consonant segments that are integrated separately. The findings also showed that the perceived consonant order in combination illusions depends on the timing between the auditory phonetic features and the mouth movements of the talker, which provides further support for a feature-based model of audiovisual integration of speech. Overall, this thesis provided experimental evidence that constitutes a valuable foundation for the development of a feature-based model of audiovisual speech perception.

DTU Health Tech
Department of Health Technology

Ørsteds Plads
Building 352
DK-2800 Kgs. Lyngby
Denmark
Tel: (+45) 45 25 39 50
www.dtu.dk