ISSN: 2083-389X
eISSN: 2084-3127

© J Hear Sci, 2022; 12(2): 20–35
DOI: 10.17430/JHS.2022.12.2.2

AUDIOVISUAL SPEECH PERCEPTION IN DIOTIC AND DICHOTIC LISTENING CONDITIONS

Contributions:
A Study design/planning
B Data collection/entry
C Data analysis/statistics
D Data interpretation
E Preparation of manuscript
F Literature analysis/search
G Funds collection

Sandhya[A-F], Vinay[CE-F]

Audiology Group, Institute of Neuromedicine and Neurosciences, Norwegian University of Science and Technology, Trondheim, Norway

Corresponding author: Sandhya, Audiology group, Institute of Neuromedicine and Neurosciences, Norwegian University of Science and Technology, Tungasletta 2, 7491, Trondheim, Norway; email: sandhya.vinay@ntnu.no

Abstract

Background: Speech perception is multisensory, relying on auditory as well as visual information from the articulators. Watching articulatory gestures which are either congruent or incongruent with the speech audio can change the auditory percept, indicating that there is a complex integration of auditory and visual stimuli. A speech segment is comprised of distinctive features, notably voice onset time (VOT) and place of articulation (POA). Understanding the importance of each of these features for audiovisual (AV) speech perception is critical. The present study investigated the perception of AV consonant-vowel (CV) syllables with various VOTs and POAs under two conditions: diotic incongruent and dichotic congruent.

Material and methods: AV stimuli comprised diotic and dichotic CV syllables with stop consonants (bilabial /pa/ and /ba/; alveolar /ta/ and /da/; and velar /ka/ and /ɡa/) presented with congruent and incongruent video CV syllables with stop consonants. There were 40 right-handed normal hearing young adults (20 females, mean age 23 years, *SD* = 2.4 years) and 20 males (mean age 24 years, *SD* = 2.1 years) who participated in the experiment.

Results: In the diotic incongruent AV condition, short VOT (voiced CV syllables) of the visual segments were identified when auditory segments had a CV syllable with long VOT (unvoiced CV syllables). In the dichotic congruent AV condition, there was an increase in identification of the audio segment when the subject was presented with a video segment congruent to either ear, in this way overriding the otherwise presented ear advantage in dichotic listening. Distinct visual salience of bilabial stop syllables had greater visual influence (observed as greater identification scores) than velar stop syllables and thus overrode the acoustic dominance of velar syllables.

Conclusions: The findings of the present study have important implications for understanding the perception of diotic incongruent and dichotic congruent audiovisual CV syllables in which the stop consonants have different VOT and POA combinations. Earlier findings on the effect of VOT on dichotic listening can be extended to AV speech having dichotic auditory segments.

Key words: normal hearing • auditory perception • place of articulation • voice onset time • consonant-vowel syllables

AUDIOWIZUALNA PERCEPCJA MOWY W OBUUSZNYCH I ROZDZIELNOUSZNYCH WARUNKACH SŁUCHOWYCH

Streszczenie

Wprowadzenie: Percepcja mowy jest wielozmysłowa, opiera się zarówno na informacji słuchowej, jak i wzrokowej. Obserwacja narządów artykulacyjnych i gestykulacji, zgodnych bądź niezgodnych z informacją słuchową, może zmieniać percepcję słuchową, co wskazuje na złożoną integrację bodźców słuchowych i wzrokowych. Segment mowy zawiera cechy rozróżniające, takie jak czas rozpoczęcia dźwięczności (*voice onset time*, VOT) i miejsce artykulacji (*place of articulation*, POA). Kluczowe dla audiowizualnej percepcji mowy (*audiovisual*, AV) jest zrozumienie znaczenia tych dwóch cech rozróżniających. W obecnej pracy zbadaliśmy percepcję AV obuuszną niezgodną i rozdzielnouszną zgodną sylab złożonych ze spółgłoski i samogłoski (*consonant vowel*, CV) z użyciem różnych VOT i POAw.

Materiał i metody: Bodźce AV obejmowały obuuszne i rozdzielnouszne sylaby typu CV ze spółgłoskami zwartymi: bilabialne /pa/ i /ba/; dziąsłowe /ta/ i /da/; tylnopodniebienne /ka/ i /ga/ (POA). Prezentowane były ze zgodnym i niezgodnym nagraniem sylab typu CV ze spółgłoskami zwartymi. W eksperymencie wzięło udział 40 praworęcznych młodych osób dorosłych o normalnym słuchu: 20 kobiet (średni wiek 23 lata, *SD* = 2,4 roku) oraz 20 mężczyzn (średni wiek 24 lata, *SD* = 2,1 roku).

Wyniki: W warunkach obuusznej niezgodnej AV krótki VOT (wypowiadanych sylab typu CV) w segmencie wizualnym był rozpoznawany, gdy segment słuchowy zawierał sylabę typu CV z długim VOT (niewypowiadane sylaby typu CV). W warunkach rozdzielnousznej zgodnej AV zaobserwowano wzrost identyfikacji segmentu słuchowego, jeżeli był prezentowany do dowolnego ucha ze zgodnym segmentem wzrokowym, co unieważniało przewagę ucha dominującego w słyszeniu rozdzielnousznym. Wyrazistość wzrokowa sylab zawierających spółgłoskę zwartą bilabialną miała większy wpływ na odbiór wzrokowy (obserwowany jako lepsze wyniki identyfikacji) w porównaniu do sylab zawierających spółgłoskę zwartą tylnopodniebienną i dlatego przeważała nad akustyczną dominacją sylab tylnopodniebiennych.

Wnioski: Wyniki tego badania są ważne dla zrozumienia obuusznej niezgodnej i rozdzielnousznej zgodnej percepcji sylab typu CV zawierających spółgłoski zwarte z różnymi kombinacjami VOT i POA. Wcześniejsze ustalenia dotyczące wpływu VOT na słyszenie rozdzielouszne mogą zostać rozszerzone na audiowizualną percepcję mowy z rozdzielnousznymi segmentami słuchowymi.

Słowa kluczowe: norma słuchu • percepcja słuchowa • miejsce artykulacji • czas rozpoczęcia dźwięczności • sylaby spółgłoska-samogłoska

Key to abbreviations

VOT of diotic incongruent AV stimuli	
SS	Syllables with short VOT in left and right audio segments
LL	Syllables with long VOT in left and right audio segments
POA of diotic incongruent AV stimuli	
AAB	Syllables with alveolar POA in left and right audio segments and syllable with bilabial POA in video segment
VVB	Syllables with velar POA in left and right audio segments and syllable with bilabial POA in video segment
BBA	Syllables with bilabial POA in left and right audio segments and syllable with alveolar POA in video segment
VVA	Syllables with velar POA in left and right audio segments and syllable with alveolar POA in video segment
BBV	Syllables with bilabial POA in left and right audio segments and syllable with velar POA in video segment
AAV	Syllables with alveolar POA in left and right audio segments and syllable with velar POA in video segment
VOT of dichotic congruent left AV stimuli	
SSS	Syllable with short VOT in left audio segment, short VOT in right audio segment, short VOT in video segment
SLS	Syllable with short VOT in left audio segment, long VOT in right audio segment, short VOT in video segment
LSL	Syllable with long VOT in left audio segment, short VOT in right audio segment, long VOT in video segment
LLL	Syllable with long VOT in left audio segment, long VOT in right audio segment, long VOT in video segment
POA of dichotic congruent left AV stimuli	
BAB	Syllables with bilabial POA in left audio segment, alveolar POA in right audio segment, bilabial POA in video segment
BVB	Syllables with bilabial POA in left audio segment, velar POA in right audio segment, bilabial POA in video segment
ABA	Syllables with alveolar POA in left audio segment, bilabial POA in right audio segment, alveolar POA in video segment
AVA	Syllables with alveolar POA in left audio segment, velar POA in right audio segment, alveolar POA in video segment
VOT of dichotic congruent right AV stimuli	
SSS	Syllable with short VOT in left audio segment, short VOT in right audio segment, short VOT in video segment
SLL	Syllable with short VOT in left audio segment, short VOT in right audio segment, long VOT in video segment
LSS	Syllable with long VOT in left audio segment, short VOT in right audio segment, short VOT in video segment
LLL	Syllable with long VOT in left audio segment, long VOT in right audio segment, long VOT in video segment
POA of dichotic congruent right AV stimuli	
ABB	Syllables with alveolar POA in left audio segment, bilabial POA in right audio segment, bilabial POA in video segment
VBB	Syllables with velar POA in left audio segment, bilabial POA in right audio segment, bilabial POA in video segment
BAA	Syllables with bilabial POA in left audio segment, alveolar POA in right audio segment, alveolar POA in video segment
VAA	Syllables with velar POA in left audio segment, alveolar POA in right audio segment, alveolar POA in video segment

Introduction

Speech is one of the most important forms of human communication, and is often multisensory in nature. To understand speech, one does not just use the auditory modality, but also information presented and available from other senses [1,2]. The integration of auditory and visual cues in speech has, in individuals with normal hearing, facilitating effects for communication. A classic example of auditory speech perception being influenced by visual information from the speaker's articulators is the McGurk effect [3] in which it is observed that adult listeners predominantly perceived /da/ when they were presented with an auditory /ba/ accompanied by a visual (a face) articulating /ga/. Schwartz et al. [4] suggested that in multimodal perception, auditory and visual systems integrate early, i.e., the auditory and visual features in speech are transformed into a common representation before recognition occurs. It is known that the influence of visual information on phonetic percepts occurs at an early level of speech processing [5,6]. The articulatory motor movements that produce visual and/or acoustic speech are the common currency between the seen and heard speech signal. Speech segments are comprised of various distinctive features, such as voicing, place of articulation, manner of articulation, etc. [7,8]. The purpose of the present experiment is to study the effects of two such distinctive features – voice onset time (VOT) and place of articulation (POA) of consonant-vowel (CV) syllables with stop consonants – on perception of audiovisual speech (AV) speech under diotic and dichotic listening conditions. Diotic refers to simultaneous presentation of identical segments to both ears, whereas dichotic refer to simultaneous presentation of two different auditory stimuli, one to each ear.

Dichotic listening tasks have been extensively employed to investigate cerebral organization for speech [9,10]. The right ear advantage in dichotic listening for verbal tasks (CV syllables) supports the earlier claim [11] that contralateral pathways having stronger cortical representation take precedence over ipsilateral pathways. Kinsbourne [12] proposed the attentional model of dichotic listening, which partially attributes right ear advantage to the attention that listeners pay to sounds presented in the right ear. Previous research on specialization of the brain hemispheres for cognitive functions has indicated that the left hemisphere is responsible for verbal speech processing [9] whereas the right hemisphere processes emotion and intonation [13,14]. The left hemisphere is recruited for temporal processing of speech, which contributes to the overall left hemisphere lateralization for speech perception [15–18].

Neurophysiological studies have demonstrated hemispheric-specific processing of rapid temporal variations in speech. In a positron emission tomography study, Zatorre & Berlin [19] demonstrated that faster temporal changes resulted in a greater response from the left Heschl's gyrus (HG) whereas the right anterior superior temporal gyrus (STG) was responsible for processing an increased number of spectral elements. For short and rapid temporal events, studies have indicated left hemisphere activation [20,21], whereas for long duration signals, strong activation of the right superior temporal sulcus was observed [22]. Schwartz & Tallal [16] concluded that the left temporal lobe is specialized for analysis of rapidly changing speech and found greater right ear advantage for short transitions (40 ms) in synthesized consonant-vowel (CV) syllables. For speech, the left hemisphere might then preferentially extract and process information from short temporal integration windows, (e.g., segments with rapid spectral changes such as formant transitions which provide information about the place of articulation), while the energy envelope and spectral and prosodic information (which are long temporal events) might be better processed by the right hemisphere [23].

The consonants used in dichotic listening experiments with speech stimuli can be classified based on their manner of articulation, voice onset time, and place of articulation. Voice onset time (VOT) is the difference in time between the release of complete articulatory constriction of a stop consonant and the onset of quasiperiodic vocal cord vibration [24]. In syllable initial position, unvoiced stops are characterized by long VOTs (e.g., /p/, /t/, /k/) while voiced stops have short VOTs (e.g., /b/, /d/, /g/). Place of articulation (POA) is the location of the blockage (partial or complete) of air in the vocal tract. A stop consonant is articulated by a complete blockage of air flow in the vocal tract by the tongue or lips, followed by a sudden release of air. In dichotic listening, right ear advantage for stop CV syllables reflects left hemisphere dominance [25–28] whereas left ear advantage for voicing contrast reflects predominantly right hemisphere dominance [29]. In dichotic listening tasks with speech stimuli, stop consonants have been shown to have a higher and more reliable right ear advantage than fricatives [26], liquids [27], or vowels [25,28].

Sub-phonemic features (e.g., VOT) of a speech stimulus are factors which affect dichotic listening [30]. In a behavioral study on 89 normal hearing listeners with Norwegian as their native language, Rimol et al. [30] studied the effect of VOT on dichotic listening with stop CV syllables. With three short VOT syllables (/ba/, /da/, /ga/) and 3 long VOT syllables (/pa/, /ta/, /ka/), the four possible combinations of VOTs for the dichotic pairs (left and right ear) were short–short, short–long, long–short, and long–long. The results revealed that short–long syllable pairs produced the largest right ear advantage, whereas long–short stimuli pairs resulted in significant left-ear advantage due to possible right hemisphere dominance for the processing of long acoustic events (as with long VOT of voiceless stops). They also suggested that syllables with long VOT are perceptually stable in competing listening conditions such as dichotic listening, and analysis of long VOT syllables may require lesser temporal precision than short VOT syllables. A shift in ear advantage determined by the VOT of CV syllables suggests VOT to be a more powerful determinant of dichotic listening performance than classic right ear advantage [30].

In terms of POA cues and speech perception, Speaks et al. [31] reported that velar syllables tend to have higher identification over bilabial and alveolar syllables. This might be attributed to more compact distribution of spectral energy in velar syllables compared to bilabial and alveolar syllables. O'Brien [32] hypothesized that owing to

diffuse distribution of spectral energy, bilabial and alveolar stop consonants require greater processing than velar stop consonants before recognition. Voyer & Techentin [33] studied the effect of POA and stimulus dominance on dichotic listening and concluded that spectral and temporal factors make specific stimuli (velars) more salient, irrespective of the ear of presentation. They also suggested that exclusion of dominant syllables from a dichotic task would likely be beneficial in avoiding misinterpretation of auditory asymmetries. In AV speech, fusion perception in incongruent AV conditions is more for voiced stop consonants than for unvoiced stops (in other words, stops with shorter VOT than longer VOT) [5]. Alm & Behne [38] attributed greater identification of POA of voiced stops compared to voiceless stops to the distinct spectral distribution of voiced stops.

Perception of AV speech has been studied under both monaural and diotic listening conditions [34]. Participants in a study by Scott [34] were presented with McGurk-like AV stimuli (with auditory /aba/ and visual /aga/) under monaural and binaural conditions. Scott suggested that the right ear advantage (due to left hemisphere dominance for speech) makes it harder to induce the McGurk effect when the auditory component is presented to the right ear rather than the left. This experimental design is important, as it may serve as a behavioral measure of hemispheric dominance for language.

Although the classic McGurk effect [35] is observed for diotic stimuli, Omata & Mogi [36] demonstrated that the McGurk effect also occurs in dichotic listening condition. Their experiment included dichotic auditory stimuli (/ba/ to the left ear, /ga/ to the right; /ga/ to the left ear, /ba/ to the right) together with visual /ba/ and /ga/, as well as classic McGurk stimuli with diotic auditory /ba/ and visual /ga/. They demonstrated the dominance of visual /ba/ in the AV dichotic condition, and that right ear advantage may be effective when visual /ga/ was paired with dichotic /ba/–/ga/. Audiovisual integration in the perception of Swedish vowels, in which lip rounding is a distinctive feature, has been studied by Öhrström & Traunmüller [37]. They concluded that AV fusion occurs for vowels: auditory /e/ presented with visual /y/ was perceived mostly as an /ø/, while an auditory /y/ combined with a visual /e/ was perceived mostly as an /i/. They suggested that fusion percepts occur for longer segments such as vowels rather than just for shorter segments such as consonants. The role of visual input is dominant for the perception of roundedness of vowels which is a distinctly visible feature, analogous to bilabial consonants. In a later study, Traunmüller & Öhrström [6] used the Swedish vowels /i/, /y/, and /e: / embedded in a CVC syllable – with /g/ in auditory, visual, and incongruent AV conditions –and confirmed that roundedness of the vowels is perceived visually whereas the openness of vowels is perceived auditorily.

Recently, Sandhya et al. [39] investigated the distribution of modality-specific responses to dichotic incongruent AV speech stimuli in normal hearing young adults. The study used incongruent AV stimuli with simultaneous presentation of three different stop CV syllables to the right ear, left ear, and visually. They reported that a salient video segment, such as of a bilabial CV syllable, reduced the overall recognition of the auditorily dominant velar CV syllable presented to the left ear, thus suggesting an effect of POA of the syllables on speech perception. They also demonstrated that right ear advantage for velar CV syllables was observed only for voiced stimuli, inferring that perception of dichotic incongruent AV stimuli depends on POA, VOT, auditory and visual salience, and modality (left, right, or video) of how the syllables are presented.

When identifying consonants, the POA cues are coded acoustically as fine variations in frequencies, whereas VOT provides temporal information. These features of speech segments are independent, and are often described using phonological categories [40]. In the case of multimodal audiovisual speech perception, the perceptual role of temporal structure in audiovisual speech needs further research [39,41]. The existing literature does not provide enough evidence about the role of VOT and POA on perception of speech under diotic and dichotic AV conditions. Thus, the present study aimed to broaden the understanding of AV speech perception by incorporating three AV conditions: *diotic incongruent* (diotic audio stimuli are presented with an incongruent video segment), *dichotic congruent left* (dichotic audio stimuli with visual segment congruent to the left ear), and *dichotic congruent right* (dichotic audio stimuli with visual segment congruent to the right ear). Dichotic congruent left and dichotic congruent right AV conditions were incorporated to obtain ear-specific responses and thus to study how the ear advantage observed in dichotic listening is affected by the simultaneous presentation of a video segment. As a baseline, a dichotic listening experiment was carried out prior to the AV experiment to compare the right and left ear responses in the audio only condition.

We hypothesize that, in the diotic incongruent AV condition, identification of POA as well as AV interaction in terms of fusion would be greater for short VOT CV syllables than long VOT syllables. For dichotic congruent AV presentations, we hypothesize that distinctly visible syllables boost responses of congruent auditory stimuli, overriding the ear advantage observed in dichotic listening. Furthermore, we hypothesize that long VOT would result in greater responses from the ear of presentation.

Material and methods

Participants

We recruited 40 right-handed young adults (age range 20–29 years; 20 females, mean age 23 years, $SD = 2.4$ years and 20 males, mean age 24 years, $SD = 2.1$ years) among students at the Norwegian University of Science and Technology (NTNU). All had Norwegian as their native language. All participants had audiometric thresholds better than 20 dB hearing level in both ears for octave frequencies from 0.25 to 8 kHz [42]. All participants had binocular visual acuity of 20/25 or better as evaluated with the Snellen test [43] adjusted for viewing on a 24-inch monitor at a resolution of 1920 × 1200 pixels, and they had no history of disorders related to visual acuity or color blindness. None of the participants reported neurological, speech or language, attention, or motor disorders. Participation was voluntary and all participants gave written informed

Table 1. The three types of audiovisual stimuli used in the study

	Stimuli type	Stimuli description
1	Diotic incongruent	Diotic auditory stimuli, with video syllable always incongruent with the auditory syllable
2	Dichotic congruent left	Dichotic auditory stimuli, with video syllable congruent with the auditory syllable in the left ear
3	Dichotic congruent right	Dichotic auditory stimuli, with video syllable congruent with the auditory syllable in the right ear

consent to participate in the study. The study was registered with the Norwegian Social Science Data Services.

Audiovisual speech stimuli

The stimuli were the consonant-vowel (CV) syllables /ba/, /da/, /ɡa/, /pa/, /ta/, and /ka/. In Norwegian, /pa/, /ta/, and /ka/ with long VOT are realized as aspirated /pʰɑ/, /tʰɑ/, and /kʰɑ/ respectively in word initial position. However, /ba/, /da/, and /ɡa/ with short VOT are realized as unaspirated /bɑ/, /dɑ/, and /ɡɑ/ respectively in word initial position [44]. In Norwegian, ba, da, ga, and ta are words which translate as ask, then, gave, and take respectively. The auditory syllables were presented either dioticaly (same syllable at the two ears) or dichotically (different syllables at the two ears). The auditory syllables were presented together with video recordings of CV syllables. In the diotic AV stimuli, the video syllable was always incongruent with the auditory syllable, i.e., the video syllable differed from the audio syllable. In the dichotic AV stimuli, the video syllables were congruent with the auditory syllable either in the left ear or the right ear.

In summary, there were three types of AV test stimuli: *diotic incongruent*, *dichotic congruent left*, and *dichotic congruent right* (**Table 1**). Congruent AV stimuli (congruent audio and video segments) were used to confirm that participants made correct phonetic judgments when there was no conflict between the audio and video syllables.

Recording of AV materials

The talker was a young adult female speaker with an urban Eastern Norwegian dialect. Recordings were made in this dialect since it is familiar to most Norwegians [45]. The AV recordings were made in a sound-insulated studio at the Speech Lab, Department of Psychology, NTNU. The AV recordings were made in a sound-insulated studio with a PDW-F800 Sony Professional XDCAM HD422 camcorder (Tokyo, Japan) positioned approximately 2 m in front of the speaker and two Røde NT1-A microphones (Sydney, Australia) placed in front of her at knee height; one of the microphones was connected to the camcorder and the other was fed through an RME Fireface 400 (Haimhausen, Germany) box to an Apple Macintosh G5 computer (Cupertino, CA). The two audio channels were recorded at a sampling rate of 48 kHz using Praat version 5.1 [46]. The speaker repeated each syllable 10 times, and was instructed to look at the camera, keep a relatively flat intonation, have a neutral facial expression, and keep facial gestures such as eye blinks to a minimum. MPEG-4 video files from the camcorder were divided into 1000 ms segments, each containing one syllable, using the AVID Media Composer, version 3.5 (Burlington, MA). The audio files recorded on the G5 computer were also segmented into separate syllables using Praat. A subset of the recordings of the audio on the G5 computer and video files for each of the segmented syllables was selected to have a natural dialect and speaking rate, absence of background noise, absence of rising or falling intonation, an emotionally neutral voice, neutral eyes and mouth, absence of eye blinks, mouth completely closed and neutral before and after articulation, and typical intensity and pitch contour, based on ratings by two native Norwegian speakers [45]. **Figures 1–6** show photographs of the mouth of the speaker at the time of stop release (upper panel) and spectrograms (lower panel) for the CV syllables used in the study. The red vertical lines in the spectrograms indicate the burst. Voice onset times of the CV syllables used in the present study are given in **Table 2**.

Creation of AV stimuli

The audio segments were first edited using Praat [46] to give the same unweighted intensity for all syllables and then synchronized with the audio recorded by the camcorder using Logic Pro (version 8.0.2) (Cupertino, CA), and then substituted for the camcorder audio using AVID media composer.

Diotic and dichotic auditory stimuli were created by temporally aligning the stop consonant onset (release of obstruction) in the two stereo channels using Logic Pro. Stop consonant onset/release was based on inspection of the spectrogram. For diotic stimuli, the audio syllables were identical in the left and right stereo channels (e.g., /pa/ in both channels). For dichotic stimuli, different syllables were used for the left and right channels (e.g., /pa/ in the left and /ka/ in the right). The onset of mouth movements for the AV stimuli was synchronized with the onset of the release burst of the audio syllables. The diotic incongruent AV stimuli were created by combining the audio of one syllable (e.g., /pa/) with the video of a different syllable (e.g., /ka/). The dichotic congruent left and dichotic congruent right AV stimuli were created by combining dichotic audio syllables with a video syllable corresponding to either the left or right ear audio syllable.

Tables 3, 4, and **5** summarize the characteristics of the dichotic audio stimuli, diotic incongruent AV stimuli, and dichotic congruent left and dichotic congruent right AV stimuli, respectively. Of the six CV syllables, /ba/, /da/, and /ɡa/ had a short VOT, denoted S, while /pa/, /ta/, and /ka/ had a long VOT, denoted L. This yielded four possible syllable pairs for the audio stimuli: (1) SS, short VOT in both ears; (2) SL, short VOT in left ear and long VOT

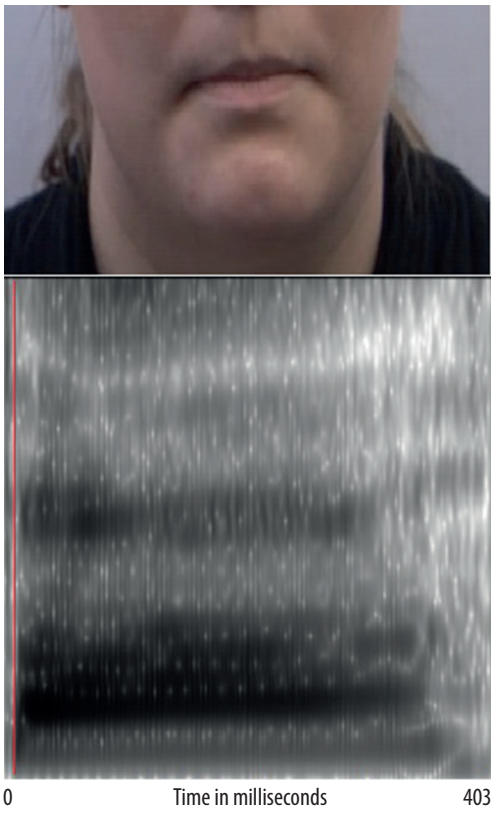

Figure 1. Upper panel is photograph of mouth of speaker at the time of stop release for CV syllable /ba/; lower panel is its spectrogram. The red vertical line marks the burst

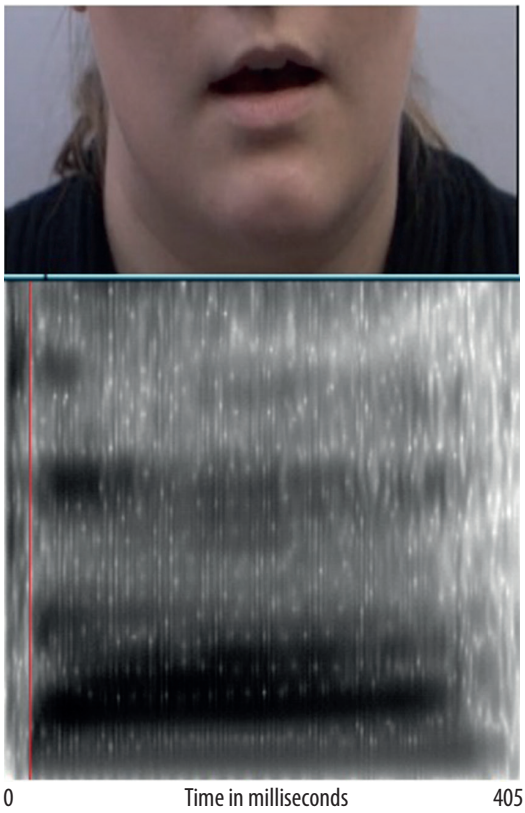

Figure 2. As for Figure 1 but for CV syllable /da/

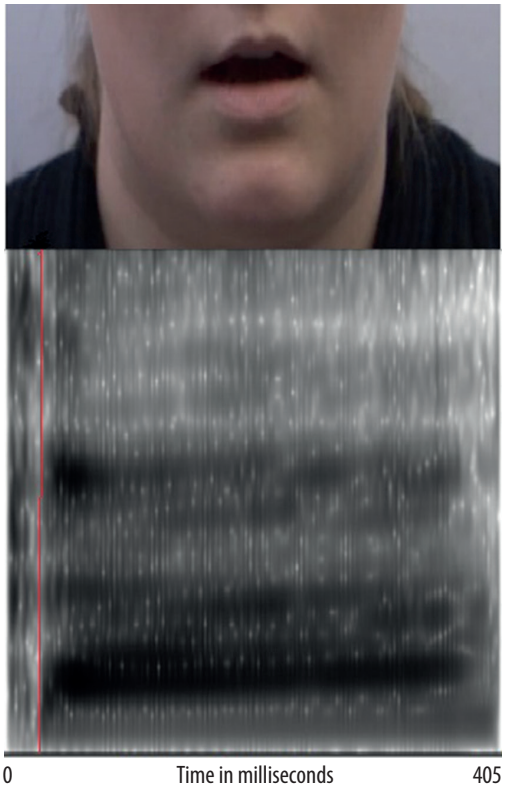

Figure 3. As for Figure 1 but for CV syllable /ga/

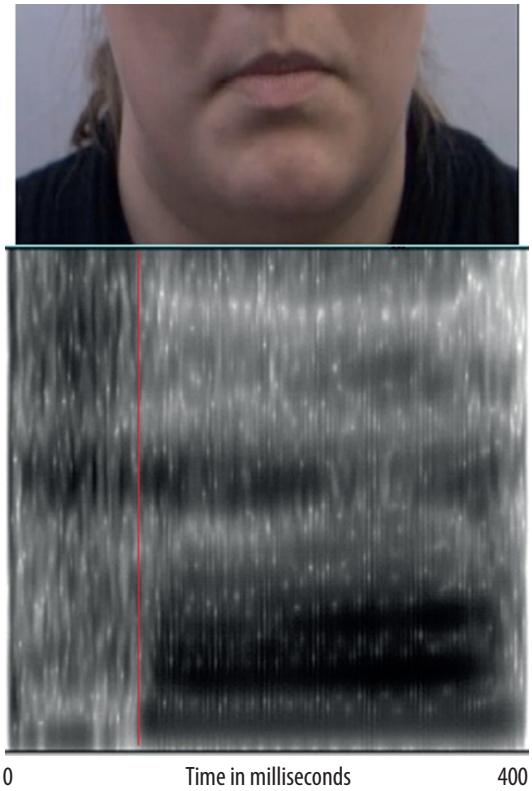

Figure 4. As for Figure 1 but for CV syllable /pa/

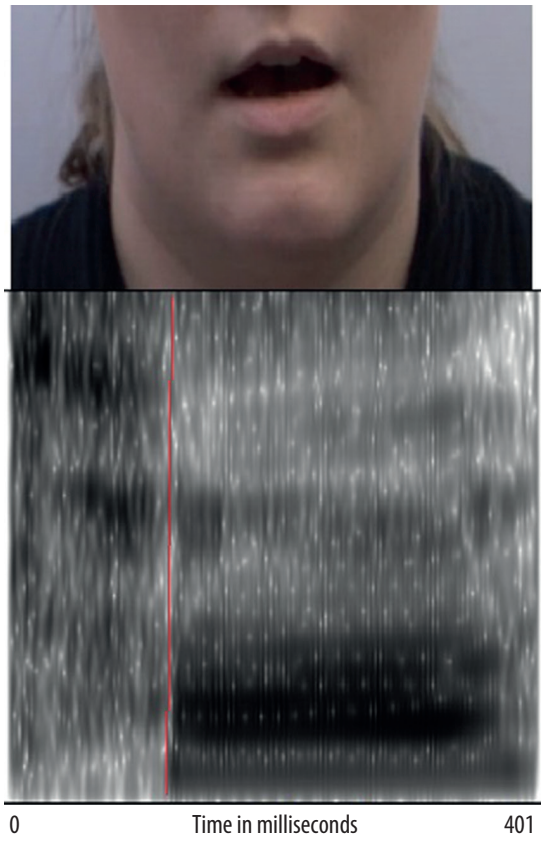

Figure 5. As for Figure 1 but for CV syllable /ta/

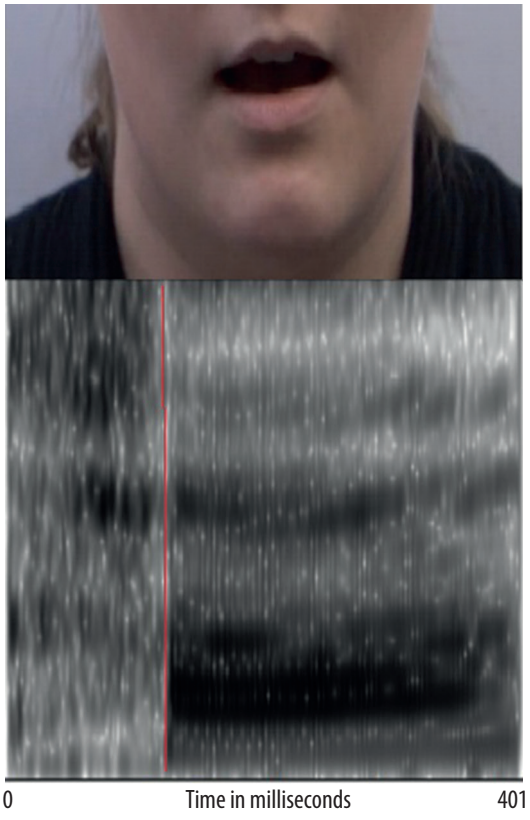

Figure 6. As for Figure 1 but for CV syllable /ka/

in right ear; (3) LS, long VOT in left ear and short VOT in right ear; and (4) LL, long VOT in both ears. The place of articulation (POA) of the syllables was categorized as bilabial (B) with /ba/ and /pa/, alveolar (A) with /da/ and /ta/, and velar (V) with /ga/ and /ka/.

Procedure

The experiment was conducted using Superlab version 4 (San Pedro, CA) run on an iMac 11.3 (Cupertino, CA). The participant sat facing a 24-inch monitor (1920 × 1200 pixels) at approximately 70 cm. The auditory stimuli were presented through AKG K271 studio headphones (Vienna, Austria) at a level of approximately 68 dBA. A Cedrus RB-730 seven-button response box (San Pedro, CA) was placed in front of the participant. The order of the response options /pa/, /ba/, /ta/, /da/, /ka/, and /ga/ on the response box was randomized across participants to avoid bias due to the sequence of buttons.

The dichotic listening experiment was carried out before the AV experiment. The AV experiment commenced with practice trials using 10 randomly selected AV stimuli. The diotic incongruent, dichotic congruent left, and dichotic congruent right AV stimuli (with three repetitions of each stimulus) were randomly divided into 11 blocks, each lasting about 5 minutes. There were 18 congruent AV stimuli (6 congruent AV stimuli presented three times each) which were randomly interspersed with the test stimuli. At the start of each trial, before the presentation of a stimulus, a 'fixation cross' appeared at the center of the monitor for

Table 2. Voice onset time (VOT) of the CV syllables

CV syllables	Voice onset time [ms]	Syllable duration [ms]
/ba/	14	403
/da/	18	405
/ga/	27	405
/pa/	103	400
/ta/	118	401
/ka/	122	401

200 ms, where the video would be presented, to focus the participant's attention. The participant was instructed to watch and listen to each AV stimulus and press a button on the response box as quickly as possible to indicate the perceived syllable. Each subject had the same response order for the entire experiment and the participants were familiarized with the order of the response buttons before the start of the experiment. If two syllables were perceived, the participant was instructed to press two buttons, one for each syllable. The responses were logged in Superlab. The inter-trial interval was 5 s and a break was given after each block. The entire experiment, including the pre-evaluations of vision and hearing, took approximately 1.5 hours to complete.

Table 3. Dichotic audio stimuli comprised of two syllables, categorized in terms of the VOT of the syllables (short, S; long, L). In each column, the first syllable is presented to the left ear and the second to the right

Voice onset time of CV syllables			
SS	**SL**	**LS**	**LL**
bada	bata	pada	pata
baga	baka	paga	paka
daba	dapa	taba	tapa
daga	daka	taga	taka
gaba	gapa	kaba	kapa
gada	gata	kada	kata

Table 4. Diotic incongruent AV stimuli comprised of three syllables (same audio syllables in the two ears), categorized in terms of the VOT of the syllables (short, S; long, L) and their POA (bilabial, B; alveolar, A; velar, V). The first syllable is presented to the left ear, the second syllable to the right, and the third is presented visually

Place of articulation of video syllable	**Voice onset time of CV syllables**			
	SSS	**SSL**	**LLS**	**LLL**
AAB	dada:ba	dada:pa	tata:ba	tata:pa
VVB	gaga:ba	gaga:pa	kaka:ba	kaka:pa
BBA	baba:da	baba:ta	papa:da	papa:ta
VVA	gaga:da	gaga:ta	kaka:da	kaka:ta
BBV	baba:ga	baba:ka	papa:ga	papa:ka
AAV	dada:ga	dada:ka	tata:ga	tata:ka

Table 5. Dichotic congruent left and dichotic congruent right AV stimuli comprised of three syllables, categorized in terms of the VOT of the syllables (short, S; long, L) and the POA of the syllables (bilabial, B; alveolar, A; velar, V). The first syllable was presented to the left ear, the second to the right, and the third visually

Voice onset time of CV syllables									
Place of articulation of the syllables	**SSS**	**SLS**	**LSL**	**LLL**	**Place of articulation of the syllables**	**SSS**	**SLL**	**LSS**	**LLL**
	Dichotic congruent left					**Dichotic congruent right**			
BAB	bada:ba	bata:ba	pada:pa	pata:pa	ABB	daba:ba	dapa:pa	taba:ba	tapa:pa
BVB	baga:ba	baka:ba	paga:pa	paka:pa	VBB	gaba:ba	gapa:pa	kaba:ba	kapa:pa
ABA	daba:da	dapa:da	taba:ta	tapa:ta	BAA	bada:da	bata:ta	pada:da	pata:ta
AVA	daga:da	daka:da	taga:ta	taka:ta	VAA	gada:da	gata:ta	kada:da	kata:ta
VBV	gaba:ga	gapa:ga	kaba:ka	kapa:ka	BVV	baga:ga	baka:ka	taga:ga	paka:ka
VAV	gada:ga	gata:ga	kada:ka	kata:ka	AVV	daga:ga	daka:ka	paga:ga	taka:ka

Analyses

Participants' responses to the AV stimuli were categorized based on whether they matched the VOT (in terms of short and long) and POA of the audio and/or video segment of the stimuli. An alveolar CV syllable (/ta/ or /da/) in response to an AV stimulus with bilabial audio segment and velar video segment was categorized as a fusion response. Responses to the dichotic listening experiment were categorized as to whether they matched the right audio or the left audio segment.

For the diotic incongruent stimuli, response category *Apm* refers to a response by the participant *matching the POA* of the *audio* segment; *Vpm* refers to a response by the participant matching the *POA* of the *video* segment; while response category *Fp* refers to the participant perceiving an AV *fusion* based on *POA*. In a similar way, response category *Avm* refers to a response by the participant *matching the voicing* (VOT) of the *audio* segment, while *Vvm* refers to a response by the participant *matching* the *video* segment based on VOT. For the dichotic congruent left and dichotic congruent right stimuli, response categories *Lpm* and *Rpm* refer to responses by the participant *matching* the *POA* of the *left* audio segment or *right* audio segment respectively. *Fp* again refers to the participants perceiving an AV fusion based on POA. Response categories *Lvm* and *Rvm* refer to a response by the participant *matching* the *voicing* (VOT) of the *left* audio segment and *right* audio segment respectively.

Dependent and independent factors for the repeated-measure ANOVA are described under each condition in the results section. A Greenhouse-Geisser correction procedure was applied on all ANOVAs to correct for violation of the sphericity assumptions.

Results

Across stimuli, the identification of the congruent AV stimuli was consistently near ceiling (mean = 96%, SD = 10%), indicating that participants made correct phonetic judgements for stimuli for VOT (voicing) and POA. The results for dichotic listening are presented first, followed by results for the diotic incongruent AV condition, then dichotic congruent left and dichotic congruent right AV conditions.

Dichotic listening

Pairwise comparisons of the right ear and left ear responses to the dichotic stimuli with various VOT combinations (SS, SL, LS, and LL) revealed significantly higher right ear responses than left ear responses for SS stimuli [$t(665) = -5.680$, $p < 0.001$], SL stimuli [$t(665) = -12.880$, $p < 0.001$], and LL stimuli [$t(665) = -1.000$, $p < 0.001$].

Diotic incongruent AV condition

Responses to diotic incongruent AV stimuli were analyzed in a repeated-measures ANOVA with POA (AAB, VVB, BBA, VVA, BBV, and AAV) and VOT (SSS, SSL, LLS, LLL) of the segments as intra-individual independent factors and matched responses (Apm, Vpm, F, Avm, and Vvm) as dependent factors.

Voice onset time

A significant main effect of VOT of diotic incongruent AV stimuli was observed for Apm responses [$F(3, 117) = 10.845$, $p < 0.001$, $\eta^2_p = 0.218$] but not for Vpm responses [$F(3, 117) = 0.737$, $p = 0.524$]. Pairwise comparisons of VOT combinations revealed that SSS stimuli had the least Apm responses compared to SSL stimuli ($p < 0.001$), LLS stimuli ($p = 0.004$), and LLL stimuli ($p = 0.012$). However, results did not reveal any significant differences in Apm responses among other VOT combinations.

To study how the VOT of diotic incongruent AV stimuli affected voicing match responses, stimuli with VOT combinations SSL and LLS were considered. Results revealed a significant main effect of VOT on Vvm responses [$F(1, 39) = 13.323$, $p = 0.001$, $\eta^2_p = 0.255$] but not on Avm responses [$F(1, 39) = 2.910$, $p = 0.096$]. Pairwise analyses revealed that Vvm responses were significantly higher for LLS stimuli than for SSL stimuli. Video segments with short VOTs were identified correctly when the audio segments had long VOTs.

For potential fusion of diotic incongruent AV stimuli (bilabial audio segment with velar video segment), results indicated a significant effect of VOT on fusion responses [$F(3, 117) = 59.912$, $p < 0.001$, $\eta^2_p = 0.606$]. Fusion responses for SSS stimuli were significantly higher than LLL ($p < 0.001$), SSL ($p < 0.001$), and LLS stimuli ($p < 0.001$). There were no significant differences between fusion responses for SSL, LLS, and LLL stimuli.

Place of articulation

Results revealed a significant main effect of POA of diotic incongruent AV stimuli on Apm responses [$F(5, 195) = 23.058$, $p < 0.001$, $\eta^2_p = 0.372$] and on Vpm responses [$F(5, 195) = 17.900$, $p < 0.001$, $\eta^2_p = 0.315$]. Pairwise comparisons of POA combinations revealed that Apm responses were significantly higher for stimuli with alveolar and velar audio segments than for stimuli with bilabial audio segments. However, irrespective of the video segment, there was no significant difference between Apm responses for stimuli with alveolar and velar audio segments. Pairwise comparison for POA combinations revealed diotic stimuli with bilabial video segments resulted in higher Vpm responses compared to diotic stimuli with alveolar and velar video segments.

Interaction between VOT and POA

Based on Apm responses, the results, shown graphically in **Figure 7A**, reveal a significant interaction between VOT and POA of diotic incongruent AV stimuli [$F(15, 585) = 13{,}167$, $p < 0.001$, $\eta^2_p = 0.252$]. As shown in **Figure 7A**, Apm responses are similar across combinations of POA and VOT stimuli – except for BBV stimuli with SSS VOT, where the short VOT velar video segment led to lowest Apm responses (green bar at left). The short green bar can be attributed to the presence of fusion responses due to possible AV integration, since the BBV stimulus is a classic McGurk stimulus.

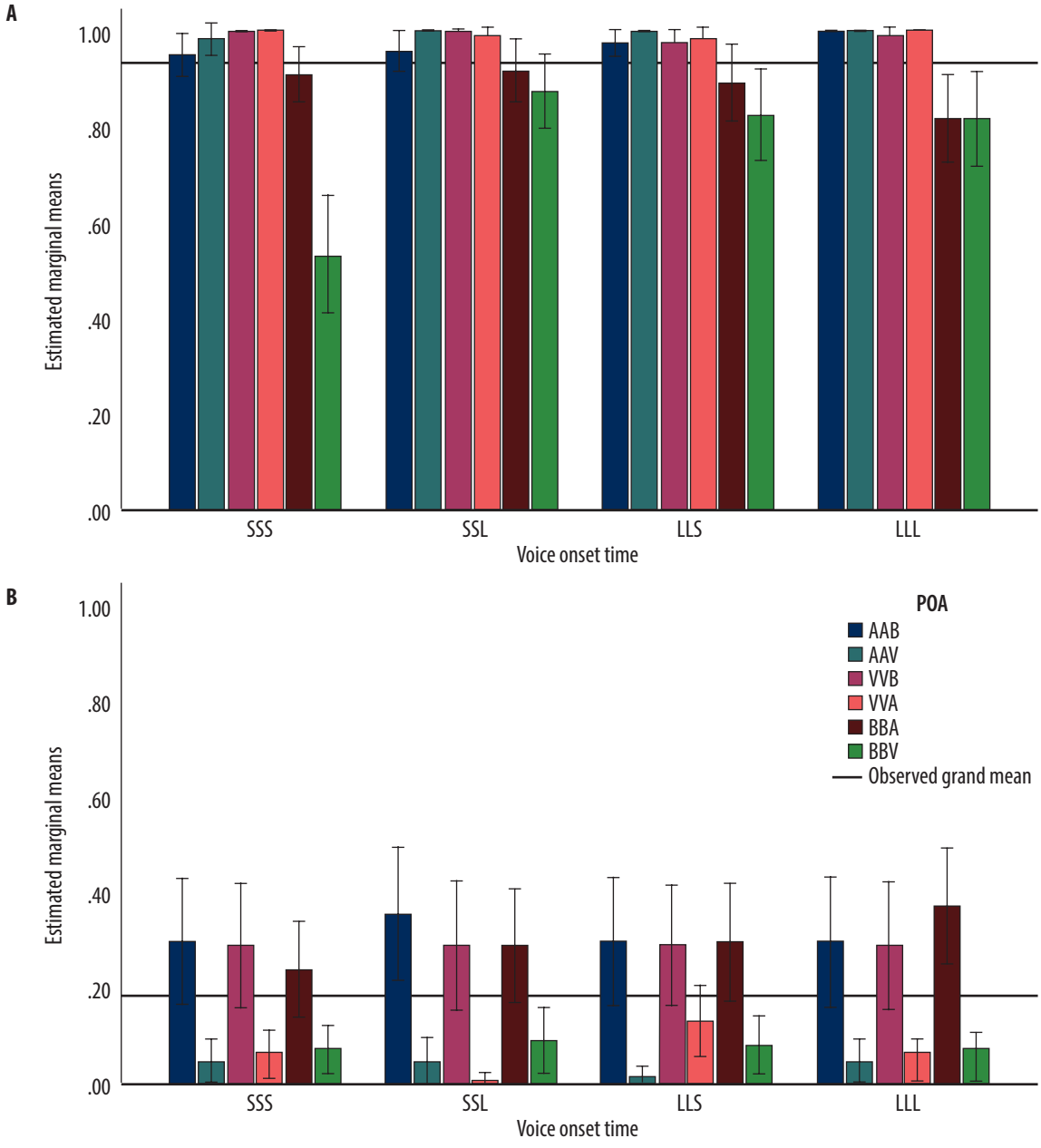

Figure 7. (**A**) Distribution of Apm responses for diotic incongruent stimuli with four VOT and POA combinations. (**B**) Distribution of Vpm responses for diotic incongruent stimuli with four VOT and POA combinations. Error bars are ± 2 standard errors

For Vpm responses, however, the interaction between VOT and POA of diotic incongruent AV stimuli [$F(15, 585) = 2{,}063$, $p < 0.001$, $\eta^2_p = 0.050$] was weak, with very low statistical power and large standard errors, as shown in **Figure 7B**. The results do not reveal any specific pattern for interaction between VOT and POA.

Dichotic congruent left and dichotic congruent right AV condition

Responses to dichotic congruent left and dichotic congruent right AV stimuli were analyzed with dichotic congruence (video congruent with left ear and video congruent with right ear), POA of audio segment (BA, BV, AB, AV, VB, and VA), and VOT of audio segment (SS, SL, LS, and LL) as intra-individual factors, while matched responses (Lpm, Rpm, Fp, Lvm, and Rvm) were dependent factors.

Voice onset time

Results revealed a significant main effect of VOT of dichotic congruent AV stimuli on Lpm responses [$F(3, 117) = 12.625$, $p < 0.001$, $\eta^2_p = 0.245$] and Rpm responses [$F(3, 117) = 13.373$, $p < 0.001$, $\eta^2_p = 0.255$]. Pairwise comparison for VOT combinations revealed that Lpm responses were not significantly different for SS and LL stimuli (when both ears received either short VOT or long VOT segments); however, Rpm responses were higher for SS stimuli than LL stimuli. The auditory segments with long VOT (as in LS stimuli), resulted in significantly higher

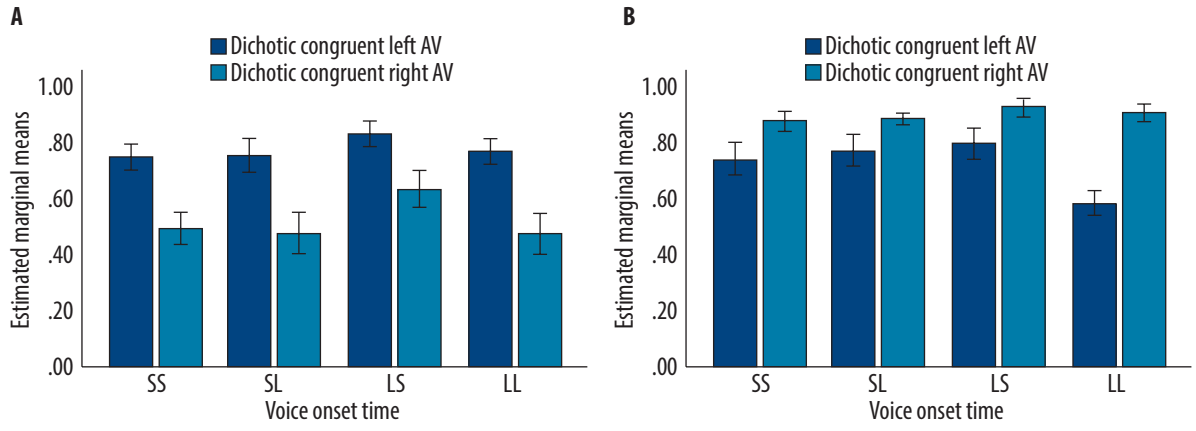

Figure 8. (**A**) Distribution of Lpm responses for dichotic congruent AV stimuli with four VOT combinations. (**B**) Distribution of Rpm responses for dichotic congruent AV stimuli with four possible VOT combinations. Error bars are ± 2 SE

Lpm responses than SS ($p < 0.001$), SL ($p < 0.001$), or LL stimuli ($p < 0.001$). However, no pattern for the effect of VOT on Rpm responses was observed.

However, as **Figure 8A** and **8B** show, there was significant interaction between dichotic congruence and VOT for both Lpm responses [$F(3, 117) = 3{,}760$, $p = 0.013$, $\eta^2_p = 0.088$) and Rpm responses [$F(3, 117) = 39.008$, $p < 0.001$, $\eta^2_p = 0.500$]. Lpm responses were significantly higher for dichotic congruent left than dichotic congruent right irrespective of VOT (**Figure 8A**), but Rpm responses (**Figure 8B**) were significantly higher for dichotic congruent right than dichotic congruent left only for LL stimuli (which had the longest VOT consonants).

Place of articulation

Overall results showed a significant main effect of dichotic congruence on Lpm responses [$F(1, 39) = 134.364$, $p < 0.001$, $\eta^2_p = 0.776$] and Rpm responses [$F(3, 117) = 95.802$, $p < 0.001$, $\eta^2_p = 0.711$], with Lpm responses higher for dichotic congruent left stimuli than dichotic congruent right AV stimuli ($p < 0.001$) and Rpm responses higher for dichotic congruent right stimuli than dichotic congruent left stimuli ($p < 0.001$).

As shown in **Figure 9A**, there was a significant interaction between dichotic congruence and POA of dichotic congruent AV stimuli for Rpm responses [$F(5, 195) = 79.133$, $p < 0.001$, $\eta^2_p = 0.670$] and, as shown in **Figure 9B**, also for Lpm responses [$F(5, 195) = 34.239$, $p < 0.001$, $\eta^2_p = 0.467$]. Lpm responses were similar for VAA and VAV stimuli, irrespective of dichotic congruence. An alveolar video segment congruent to the right audio segment (as in a VAA stimulus) did distract the listener from identification of the velar segment in the left ear. This can be attributed to the acoustic salience of velar consonants.

Dichotic congruent left AV condition

Responses to dichotic congruent left AV stimuli were analyzed in a repeated-measures ANOVA with POA of the stimuli (BAB, BVB, ABA, AVA, VBV, and VAV) and VOT of the stimuli (SSS, SLS, LSL, and LLL) as intra-individual independent factors, with matched responses (Lpm, Rpm, Lvm, and Rvm) as dependent factors.

Voice onset time

VOT of dichotic congruent left AV stimuli had a significant effect on both Lpm responses [$F(3, 117) = 5.267$, $p = 0.002$, $\eta^2_p = 0.119$] and Rpm responses [$F(3, 117) = 29.167$, $p < 0.001$, $\eta^2_p = 0.428$]. Pairwise comparison revealed that Lpm responses were higher for LS stimuli than for SS ($p = 0.004$) and SL stimuli ($p = 0.003$). Pairwise comparisons also revealed that Rpm responses were lower for LL stimuli than SS ($p < 0.001$), SL ($p < 0.001$), and LS stimuli ($p < 0.001$).

Place of articulation

The POA of dichotic congruent left AV stimuli had a significant effect on both Lpm responses [$F(5, 195) = 17.969$, $p < 0.001$, $\eta^2_p = 0.315$] and Rpm responses [$F(5, 195) = 34.643$, $p < 0.001$, $\eta^2_p = 0.470$]. ABA stimuli gave the highest Lpm responses and AVA stimuli gave the lowest Lpm responses, whereas AVA stimuli had the highest Rpm responses and ABA stimuli had the lowest Rpm responses. This can be attributed to the dominant velar syllable presented to the right ear. Fewer Rpm responses were obtained for stimuli with bilabial audio segment in the right ear.

Dichotic congruent right AV condition

Responses to dichotic congruent right AV stimuli were analyzed in a repeated-measures ANOVA with POA of stimuli (BAA, BVV, ABB, AVV, VBB, and VAA) and VOT of stimuli (SSS, SLL, LSS, and LLL) as intra-individual independent factors and matched responses (Lpm, Rpm, Fp, Lvm, and Rvm) as dependent factors.

Voice onset time

There was a significant main effect of the VOT of dichotic congruent right AV stimuli on Lpm responses

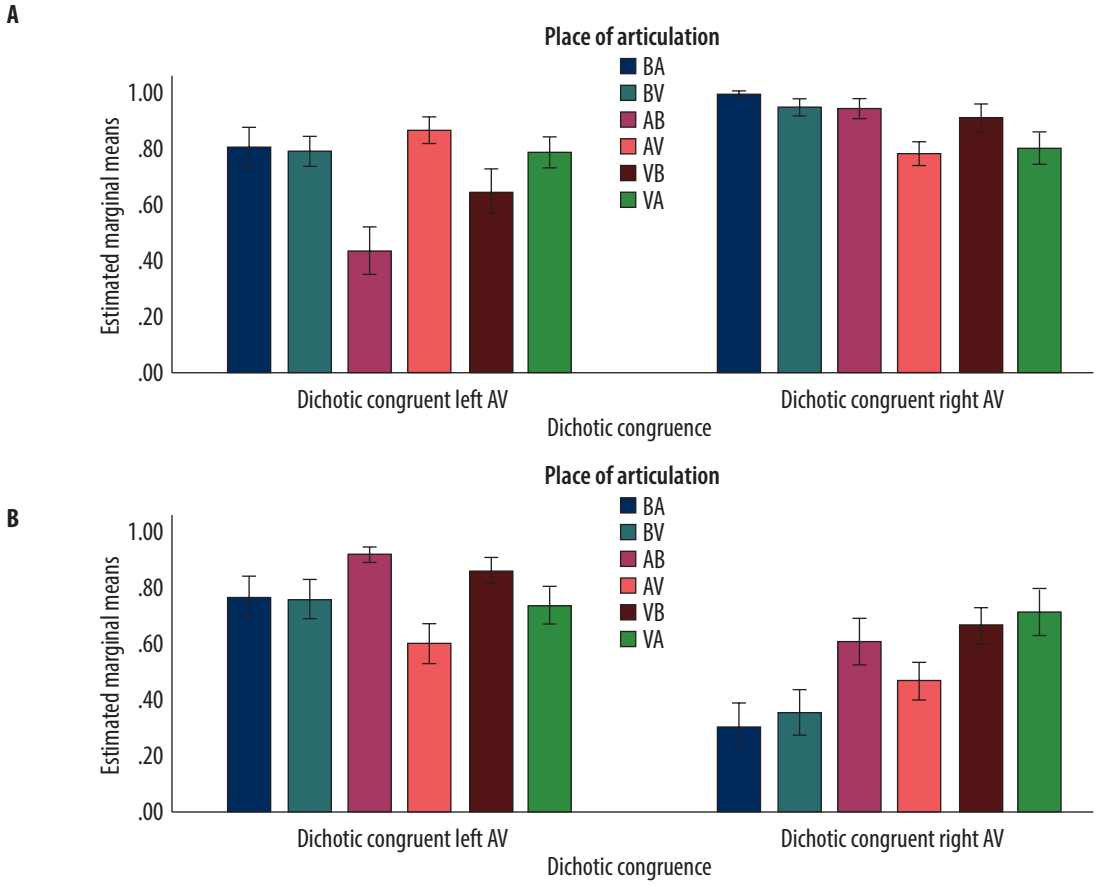

Figure 9. (**A**) Distribution of Rpm responses for dichotic congruent left and dichotic congruent right AV stimuli for six POA combinations. (**B**) Distribution of Lpm responses for dichotic congruent left and dichotic congruent right AV stimuli for six POA combinations. Error bars are ± 2 SE

[$F(3, 117) = 12.527$, $p < 0.001$, $\eta^2_p = 0.119$] but not on Rpm responses [$F(3, 117) = 1.768$, $p = 0.157$]. Pairwise comparison revealed that Lpm responses for LS stimuli were significantly higher than SS ($p < 0.001$), SL ($p < 0.001$), or LL stimuli ($p < 0.001$).

Looking at potential fusion of dichotic congruent AV stimuli, results indicated a significant effect of VOT on Fp responses [$F(3, 114) = 4.574$, $p = 0.007$, $\eta^2_p = 0.107$] but with low statistical power. Fusion responses were highest for LL stimuli and least for SS stimuli compared to other VOT combinations. Fusion responses for LL stimuli were significantly higher than SS stimuli ($p = 0.008$).

Place of articulation

There was a significant main effect of the POA of dichotic congruent right AV stimuli on Lpm responses [$F(5, 195) = 35.335$, $p < 0.001$, $\eta^2_p = 0.482$] and Rpm responses [$F(5, 195) = 17,291$, $p < 0.001$, $\eta^2_p = 0.313$]. VAA stimuli had highest Lpm responses and BAA stimuli had the least Lpm responses. This can possibly be attributed to the dominant velar syllable presented to the right ear. BAA stimuli had the highest Rpm responses and AVV stimuli had the least Rpm responses.

Discussion

The present study aimed at investigating and understanding the perception of diotic incongruent and dichotic congruent audiovisual speech, comprised of CV syllables with stop consonants having various VOTs and POAs. Previous studies have reported that the perception of auditory information is considerably influenced by the presence of visual signals. Many different approaches have been used to study the identification of auditory signals in the presence of visual stimuli. For diotic incongruent AV presentations in the present study, the probability of Vvm responses was higher when the video had short VOT (paired with an auditory segment with long VOT) compared to when video had long VOT (paired with auditory segment with short VOT). This could be due to subtle differences in the articulation of segments with short and long VOT. In syllable initial position, articulators are less tense in the production of syllables with short VOT in comparison to syllables with long VOT which are also aspirated. Studies in audiovisual perception with diotic stimuli have shown higher identification and higher audiovisual integration of voiced syllables than voiceless syllables [5,38].

The present study observed that under dichotic AV listening conditions, a visual segment congruent to the auditory component presented to either ear increases correct

identification of the respective auditory segment, and this influence of the visual segment overrides the otherwise observed right ear advantage. These findings of the present study are consistent with the earlier findings of Sams & Rusanen [47] who used AV stimuli with dichotic auditory presentation and concluded that the effect of visual stimuli concordant with auditory input to one ear (irrespective of the ear) is stronger than the right ear advantage in dichotic listening. Öhrström et al. [48] studied AV integration in speech perception using a dichotic listening task with attention focused to the right ear, and a monaural task, with random presentation in either ear. They showed that visual influence was lower in the monaural and dichotic condition compared to the binaural condition, which they attributed to increased attention to auditory stimuli in random monaural and dichotic presentations. Öhrström et al. [48] concluded that the attentional component plays a role in AV speech perception [49–51]. The previous study by Sandhya et al. [39] studied the distribution of modality-specific responses (for right ear, left ear, and visual segment) to incongruent AV speech. Stimuli comprised presentation of dichotic (audio) CV syllables with video CV syllables incongruent to both ears. Higher scores were reported for audio segments with longer VOT than that with shorter VOT. Their findings suggested that speech perception depends on the VOT and POA of CV syllables presented in each modality. Presentation of an incongruent salient visual segment reduces the identification of an acoustically salient velar syllable, particularly if the velar CV syllable was presented to the left ear. Stimulus dominance for velar CV syllables was, however, observed only for short VOT and not long VOT. They suggested that the perception of complex signals such as dichotic incongruent AV speech is affected by lateral asymmetries, stimulus dominance, and VOT and POA of the CV syllables. Considering dichotic listening as a task in which inputs to the two ears compete, a concordant visual signal benefits the respective auditory signal competing with the discordant auditory signal. This may be regarded as similar to speech perception in the presence of noise [2,52], where supplementary visual information facilitates speech perception.

The results of the present study showed that, in dichotic listening, long VOT syllables had better identification than short VOT syllables. There was significant left ear advantage for long–short (LS) pairs in dichotic congruent left and dichotic congruent right AV conditions, which is consistent with an earlier study in dichotic listening [30]. The sub-phonemic feature of VOT is a stimulus-driven factor that plays a significant role in determining ear advantage in dichotic listening. Rimol et al. [30] suggested that perceptual analyses of long VOT syllables might require less temporal precision compared to short VOT syllables, thus indicating a more stable perceptual trace for long VOT syllables. The results of the dichotic listening experiment in the present study were that there was no significant difference between right ear and left scores, and these findings are consistent with Rimol et al. [30]. In the AV condition for an LS pair, a left ear advantage was observed only during video congruent with the left ear. However, a long VOT in the left ear did not override visual influences when the video was congruent with the right ear. For AV stimuli with similar VOT in both ears, LL stimuli resulted in greater right ear or left ear scores, depending on whether the video was congruent with the respective ear. The present findings are consistent with Sams & Ramusen [47] who found that the presence of a congruent visual signal can override the right ear advantage. However, for SS stimuli, right ear advantage was only observed when the video was congruent with the right ear but not the left. Specialization of the right temporal lobe for processing high temporal precision (short VOT) might increase the correct identification of syllables presented to the left ear (although it does not lead to a left ear advantage).

The identification of video bilabial and alveolar segments might be related to the visual salience of these syllables (with front and middle places of articulation respectively) compared to velars with a posterior place of articulation. The study by Dodd & Hermelin [53] suggested that visually salient front consonants such as bilabials are also better identified by combined auditory and visual input. In the present study, visual input (irrespective of the place of articulation) facilitated identification, which was evident when the video segment was congruent with the left ear audio (directing stimuli primarily to the non-dominant right hemisphere). However, perception of an audio signal in the left ear was most enhanced when presented with a congruent bilabial video segment, followed by alveolar and velar video segments.

The findings of the present study suggest the dominance of velar syllables in dichotic AV perception tasks, with high identification of velars irrespective of the ear of presentation. Regarding stimulus dominance, findings of the present study are consistent with earlier research [31–33] pertaining to the dominance of velar syllables seen in a dichotic listening task. A study by O'Brien [32] attributed this to the compact distribution of spectral energy in velars. In dichotic listening experiments where the primary goal is to assess auditory asymmetry, dominance of velar syllables can be problematic [33]. However, Voyer & Techentin [33], by excluding the dominant stimuli from their data analysis, confirmed that auditory asymmetries are independent of stimulus dominance, but that dominant stimuli can affect the magnitude of cerebral asymmetries. The present study extends these findings to AV perception with dichotic presentations, and thus any conclusion on perceptual asymmetry must consider the sub-phonemic components of speech that lead to stimulus dominance.

In the present experiment, participants were asked, under all listening conditions, not to direct their attention to either auditory or visual signals; hence it resembled an identification task. A working memory component could have unintentionally played a role, as the participants were not instructed to limit the number of responses for each trial. Rimol et al. [30] suggested that having to report both stimuli in a dichotic listening task induces a working memory component (because the subjects must remember the second stimuli while the first is being reported), which might affect the accuracy of the responses. However, in the present work the order of the responses was not considered during the analyses. In addition, the phonemic syllables used as stimuli in the present study represent meaningful words in Norwegian. Previous studies have also involved the same stimuli as used in the present study [45]. For example, the syllables /ba/, da/, /ga/, /ta/ represent meaningful words,

and the possibility that the participants' perception could have been affected by the meaning is remote since the utterances of these syllables by the Norwegian native speaker was carried out with a neutral intonation. Similar stimuli have been used in other studies [54,30]. It may also be noted that the Eastern Norwegian dialect was chosen as it is familiar to most Norwegians [45]. In daily conversation, perception of speech involves a dynamic and complex block of multimodal information. Furthermore, a stimulus through the visual modality can serve as a critical complement to auditory sensory information, with visual stimuli providing additional information such as place of articulation, temporal aspects, and lexical information.

Limitations of the study

The present study explored the influence of stimuli related factors – POA and VOT – on the perception of AV speech, although not specifically on AV integration. The sample in the present study was representative of young, native Norwegian speakers. Typical McGurk stimuli were only a part of the experiment, and the findings are not solely concerned with McGurk stimuli. Rosenblum et al. [55] suggest that sensory integration is a function of modality-neutral information contained in different streams occurring early in the neurophysiological and perceptual process.

Basing the automaticity of AV integration based solely on McGurk stimuli might be erroneous. Incongruent AV presentations may not be often perceived as clearly as AV congruent or auditory alone speech, and the combined streams may be perceived as closer to the auditory stream alone [e.g., 56–58]. Failures of the McGurk effect can be erroneously interpreted as failure of AV integration [e.g., 56,59]. Studies suggest that with McGurk stimuli, AV stimuli will induce auditory cortical activity identical to an audio-only stimulus [5,60,61], suggesting that visual and auditory speech information might be handled similarly by the brain. Audiovisual and visual speech induce activity in motor areas similar to that of auditory speech [e.g., 60,62]. Phonetic distinctions are extracted similarly from both auditory and visual streams, and listeners may use the familiarity of speech (depending on the subject's linguistic and perceptual background) in one modality to facilitate perception of speech in the other [63,64]. Brancazio & Miller [59] showed that even though participants failed to provide a classic McGurk effect response for the combined tokens from an auditory /pi/–/bi/ continuum when the visible tokens of /ti/ were spoken at different rates, the perception of auditory VOT was still influenced by the visible rate of the visual /ti/ segment. Rosenblum [6] suggests that the McGurk effect is the quintessential example of AV speech integration. Though lexical and semantic processing occur later in the linguistic process, top–down influences of lexical and semantic context can affect the ambiguous nature of (audio segment) incongruent segments [56].

Conclusions

In a dichotic listening condition, the presentation of a video segment congruent to either ear increases the perception of the respective auditory segment, irrespective of ear advantage for dichotic listening. Overall, long VOT syllables are identified correctly, both in terms of VOT and POA, compared to short VOT syllables. Phonetic attributes such as POA and voicing might have different influences on perception. Due to their anterior POA, the distinct visual salience of bilabial syllables has a greater visual influence than velar syllables, overriding the acoustic dominance of velar stop syllables. The findings on the effect of VOT in dichotic listening can be extended to audiovisual speech perception. Future research should use more measures of place of articulation and voicing cues to examine the effects of audiovisual speech as a multisensory system.

Conflict of Interest: All authors of this manuscript declare no conflict of interest.

Funding statement: No funding was received for the present work.

References

1. Ross LA, Saint-Amour D, Leavitt VM, et al. Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments. Cereb Cortex, 2007; 17(5): 1147–53.
2. Sumby WH, Pollack I. Visual contribution to speech intelligibility in noise. J Acoust Soc Am, 1954; 26(2): 212–5.
3. McGurk H, MacDonald J. Hearing lips and seeing voices. Nature, 1976; 264(5588): 746.
4. Schwartz J-L, Robert-Ribes J, Escudier P. Ten years after Summerfield: a taxonomy of models for audio-visual fusion in speech perception. In: Hearing by Eye II. Advances in the psychology of speechreading and auditory-visual speech. Psychology Press: Hove, UK, 1998; p. 85–108.
5. Colin C, Radeau M, Soquet A, et al. Mismatch negativity evoked by the McGurk-MacDonald effect: A phonetic representation within short-term memory. Clin Neurosci, 2002; 113(4): 495–506.
6. Traunmüller H, Öhrström N. Audiovisual perception of openness and lip rounding in front vowels. J Phonetics, 2007; 35(2): 244–58.
7. Chomsky N, Halle M. The sound pattern of English. Harper & Row: New York, 1968.
8. Jakobson R, Fant CGM, Halle M. Preliminaries to speech analysis: The distinctive features and their correlates. MIT Press: Cambridge, MA, 1951.
9. Hugdahl K. Lateralization of cognitive processes in the brain. Acta Psychol, 2000; 105(2-3): 211–35.
10. Voyer D. On the magnitude of laterality effects and sex differences in functional lateralities. Laterality, 1996; 1(1): 51–84.
11. Kimura D. Cerebral dominance and the perception of verbal stimuli. Can J Psychol, 1961; 15(3): 166–71.
12. Kinsbourne M. The cerebral basis of lateral asymmetries in attention. Acta Psychol, 1970; 33: 193–201.
13. Bulman-Fleming MB, Bryden MP. Simultaneous verbal and affective laterality effects. Neuropsychologia, 1994; 32(7): 787–97.
14. Jäncke L, Buchanan TW, Lutz K, Shah NJ. Focused and nonfocused attention in verbal and emotional dichotic listening: an FMRI study. Brain Lang, 2001; 78(3): 349–63.

15. Binder, J. The new neuroanatomy of speech perception. Brain, 2000; 123(12): 2371–2.
16. Schwartz J, Tallal P. Rate of acoustic change may underlie hemispheric specialization for speech perception. Science, 1980; 207(4437): 1380–1.
17. Schönwiesner M, Rübsamen R, Von Cramon DY. Hemispheric asymmetry for spectral and temporal processing in the human antero-lateral auditory belt cortex. Eur J Neurosci, 2005; 22(6): 1521–8.
18. Tervaniemi M, Hugdahl K. Lateralization of auditory-cortex functions. Brain Res Rev, 2003; 43(3): 231–46.
19. Zatorre RJ, Belin P. Spectral and temporal processing in human auditory cortex. Cereb Cortex, 2001; 11(10): 946–53.
20. Nicholl ME. Temporal processing asymmetries between the cerebral hemispheres: evidence and implications. Laterality, 1996; 1(2): 97–137.
21. Samson S, Ehrlé N, Baulac M. Cerebral substrates for musical temporal processes. Ann NY Acad Sci, 2001; 930(1): 166–78.
22. Boemio A, Fromm S, Braun A, Poeppel D. Hierarchical and asymmetric temporal sensitivity in human auditory cortices. Nat Neurosci, 2005; 8(3): 389–95.
23. Poeppel D, Guillemin A, Thompson J, Fritz J, Bavelier D, Braun AR. Auditory lexical decision, categorical perception, and FM direction discrimination differentially engage left and right auditory cortex. Neuropsychologia, 2004; 42(2): 183–200.
24. Lisker L, Abramson AS. A cross-language study of voicing in initial stops: acoustical measurements. Word, 1964; 20(3): 384–422.
25. Cutting JE. Two left-hemisphere mechanisms in speech perception. Percept Psychophys, 1974; 16(3): 601–12.
26. Darwin CJ. Dichotic backward masking of complex sounds. Q J Exp Psychol, 1971; 23(4): 386–92.
27. Haggard MP. Encoding and the REA for speech signals. Q J Exp Psychol, 1971; 23(1): 34–45.
28. Hugdahl K, Andersson L. The "forced-attention paradigm" in dichotic listening to CV-syllables: a comparison between adults and children. Cortex, 1986; 22(3): 417–32.
29. Cohen H. Hemispheric contributions to the perceptual representation of speech sounds. Doctoral dissertation, Concordia University, 1981.
30. Rimol LM, Eichele T, Hugdahl K. The effect of voice-onset-time on dichotic listening with consonant–vowel syllables. Neuropsychologia, 2006; 44(2): 191–6.
31. Speaks C, Niccum N, Carney E, Johnson C. Stimulus dominance in dichotic listening. J Speech Lang Hear Res, 1981; 24(3): 430–7.
32. O'Brien SM. Spectral features of plosives in connected-speech signals. Int J Man Mach Stud, 1993; 38(1): 97–127.
33. Voyer D, Techentin C. Dichotic listening with consonant–vowel pairs: the role of place of articulation and stimulus dominance. J Phonetics, 2009; 37(2): 162–72.
34. Scott M. The McGurk effect affected by the right ear advantage. Can Acoust, 2008; 36(3): 156–7.
35. MacDonald J, McGurk H. Visual influences on speech perception processes. Percept Psychophys, 1978; 24(3): 253–7.
36. Omata K, Mogi K. Fusion and combination in audio-visual integration. Proc Roy Soc A, 2007; 464(2090): 319–40.
37. Öhrström N, Traunmüller H. Audiovisual perception of Swedish vowels with and without conflicting cues. In: Proc Fonetik 2004, p. 40–43.
38. Alm M, Behne D. Voicing influences the saliency of place of articulation in audio-visual speech perception in babble. In: Proc Interspeech 2008, p. 2865–8.
39. Sandhya, Vinay, Manchaiah V. Perception of incongruent audiovisual speech: distribution of modality-specific responses. Am J Audiol, 2021; 30: 968–79.
40. Irwin J, DiBlasi L. Audiovisual speech perception: a new approach and implications for clinical populations. Lang Ling Compass, 2017; 11(3): 77–91.
41. Van Tasell DJ, Greenfield DG, Logemann JJ, Nelson DA. Temporal cues for consonant recognition: training, talker generalization, and use in evaluation of cochlear implants. J Acoust Soc Am, 1992; 92(3): 1247–57.
42. British Society of Audiology. Recommended procedure: Pure-tone air-conduction and bone-conduction threshold audiometry with and without masking. Reading, UK, 2011.
43. Strouse Watt W. "How visual acuity is measured," [Internet] Available at https: //lowvision.preventblindness.org/2003/10/06/how-visual-acuity-is-measured/ Viewed 2022/05/28.
44. Kristoffersen G. The Phonology of Norwegian. Oxford: Oxford University Press, 2007.
45. Alm M, Behne D. Audio-visual speech experience with age influences perceived audio-visual asynchrony in speech. J Acoust Soc Am, 2013; 134(4): 3001–10.
46. Boersma P, Weenink D. Praat: doing phonetics by computer (Version 5.1.05; computer program). Retrieved 2009 May 1.
47. Sams M, Rusanen S. Integration of dichotically and visually presented speech stimuli. In: Proc AVSP'98, 1998, p. 89–92.
48. Öhrström N, Arppe H, Eklund L et al. Audiovisual integration in binaural, monaural and dichotic listening. In: Proc Fonetik, 2011, p.29–32.
49. Alsius A, Navarra J, Campbell R, Soto-Faraco S. Audiovisual integration of speech falters under high attention demands. Curr Biol, 2005; 15(9): 839–43.
50. Alsius A, Navarra J, Soto-Faraco S. Attention to touch weakens audiovisual speech integration. Exp Brain Res, 2007; 183(3): 399–404.
51. Tiippana K, Andersen TS, Sams M. Visual attention modulates audiovisual speech perception. Eur J Cogn Psychol, 2004; 16(3): 457–72.
52. Erber NP. Auditory-visual perception of speech. J Speech Hear Dis, 1975; 40(4): 481–92.
53. Dodd B, Hermelin B. Phonological coding by the prelinguistically deaf. Percept Psychophys, 1977; 21(5): 413–7.
54. Alm M, Behne D. Do gender differences in audio-visual benefit and visual influence in audio-visual speech perception emerge with age? Front Psychol, 2015; 6: 1014.
55. Rosenblum LD. Audiovisual speech perception and the McGurk effect. In: Oxford Research Encyclopedia of Linguistics. Oxford University Press, 2019.
56. Brancazio L. Lexical influences in audiovisual speech perception. J Exp Psychol Human, 2004; 30(3): 445–63.
57. Brancazio L, Best CT, Fowler CA. Visual influences on perception of speech and nonspeech vocal-tract events. Lang Speech, 2006; 49(1): 21–53.
58. Jerger S, Damian MF, Tye-Murray N, Abdi H. Children perceive speech onsets by ear and eye. J Child Lang, 2017; 44(1): 185–215.
59. Brancazio L, Miller JL. Use of visual information in speech perception: evidence for a visual rate effect both with and without a McGurk effect. Percept Psychophys, 2005; 67(5): 759–69.
60. Callan DE, Callan AM, Kroos C, Vatikiotis-Bateson E. Multimodal contribution to speech perception revealed by independent component analysis: a single sweep EEG case study. Cogn Brain Res, 2001; 10: 349–53.

61. Mottonen R, Krause CM, Tiippana K, Sams M. Processing of changes in visual speech in the human auditory cortex. Cogn Brain Res, 2002; 13: 417–25.
62. Calvert GA, Campbell R. Reading speech from still and moving faces: the neural substrates of visible speech. J Cogn Neurosci, 2003; 15: 57–70.
63. Rosenblum LD, Miller RM, Sanchez K. Lipread me now, hear me better later: cross modal transfer of talker familiarity effects. Psychol Sci, 2007; 18: 392–6.
64. Sanchez K, Dias JW, Rosenblum LD. Experience with a talker can transfer across modalities to facilitate lipreading. Atten Percept Psychophys, 2013; 75(7): 1359–65.