Auditory-Visual Speech Processing
2005 (AVSP'05)
British Columbia, Canada
July 24-27, 2005

Perception of Congruent and Incongruent Audiovisual Speech Stimuli

Jintao Jiang [1], *Lynne E. Bernstein* [1,2], *and Edward T. Auer, Jr.* [3]

[1] Department of Communication Neuroscience, House Ear Institute, Los Angeles, CA 90057
[2] National Science Foundation, Social, Behavioral, and Economics Directorate, Arlington, VA 22230
[3] Department of Speech-Language-Hearing: Sciences & Disorders, Univ. of Kansas, Lawrence, KS 66045

ABSTRACT

Previous studies of audiovisual (AV) speech integration have used behavioral methods to examine perception of congruent and incongruent AV speech stimuli. Such studies have investigated responses to a relatively limited set of the possible incongruent combinations of AV speech stimuli. A central issue for examining a wider range of incongruent AV speech stimuli is developing a systematic method for alignment that will work with a wide variety of segments. In the present study, we investigated the use of three different landmarks (consonant-onset, vowel-onset, and minimum distance) for aligning incongruent AV stimuli. Acoustic /ba/ or /la/ syllables were dubbed onto eight visual Consonant-/a/ syllables that spanned different places and manners of articulation. The AV stimuli were presented to ten participants. Results indicated that the effect of alignment landmark was not significant. The distance measures were found to be related to visual influence. Acoustic /ba/ tokens were more influenced by visual stimuli than acoustic /la/ tokens. Visual influence on the acoustic /ba/ tokens was mainly of the McGurk-type and/or of voicing confusion; while visual influence on the acoustic /la/ tokens was mainly of the combination type (/ba/ + /la/ = /bla/).

1. INTRODUCTION

Humans typically perceive and integrate information from multiple sensory channels [1; 2; 6; 10]. One of the most significant examples of multisensory integration is audiovisual (AV) speech perception (e.g., the McGurk effect [8] and AV speech perception in noisy acoustic conditions [11]).

Congruent and incongruent AV speech stimuli have been used widely both in behavioral studies [3; 7; 9] and fMRI (or electrophysiological) studies [12; 13] with relatively little investigation into the nature of the physical stimuli being combined. These congruent and incongruent stimuli have elicited various behavioral and brain activation patterns, but interpretation of these results is limited by our understanding of the physical stimuli.

Typical incongruent AV stimuli were AV speech signals with different temporal alignments [3; 7; 9], McGurk-style stimuli [12; 13], or an /iri/-/ili/ acoustic continuum plus visual /ibi/ [4]. However, these studies did not quantify the degree of incongruity between auditory and visual speech signals. Although the mismatched AV stimuli based on the different levels of synchrony resulted in graded levels of perceptual responses [3; 7; 9], the synchrony is only one of the factors that contribute to incongruity, and the acoustic and optical stimulus attributes should be taken into account. Given that quantitative results can be obtained from behavioral, neuroanatomical, and electrophysiological studies [12; 13], it is desirable to use mismatched AV stimuli with different quantified levels of incongruity to compare with the dependent measures in experiments.

The quantitative examination of AV speech stimulus incongruity is a difficult task: It is not known yet which parts of the signals perceivers are sensitive to in response to AV stimulus incongruity, and at which cortical level the AV speech signals are bound. Currently there is no consensus in the literature regarding how to quantify the perceptual incongruity between auditory and visual speech signals. In the present study, we proposed a novel method for aligning congruent and incongruent AV stimuli with Consonant-/a/ (/Ca/) syllables and for quantifying incongruity.

Incongruent AV speech stimuli were achieved through the mismatch between an acoustic consonant (/b/ or /l/) and visual consonants that spanned different places and manners of articulation (/b, d, g, v, z, l, w, ð/). In the literature, synchronous AV speech signals are typically aligned based on consonant onsets (acoustic bursts) [3; 7; 9]. In the present study, consonants of different durations were used (e.g., auditory /ba/ versus visual /va/). To investigate effects of alignment, (1) consonant-onset-based (***C***-), (2) vowel-onset-based (***V***-), and (3) minimum-AV-distance-based (***M***-) alignments were used to examine the effect of crossmodal alignment.

For the current study, AV incongruity was modeled as the Euclidean distance between the acoustic signals, and what will be referred to here as *phantom* acoustic signals (i.e., the original acoustic signal from the visual speech token). Acoustic features were Line Spectral Pair parameters (LSPs) [see 5]. Because the vowel was held constant as /a/, the Euclidean distances were focused on consonants (including portions of the coarticulation; see **Figure 1**). Given an auditory /C$_1$a/ (A$_{/C1a/}$) and a visual /C$_2$a/ (V$_{/C2a/}$), the distance between the auditory and visual (i.e., phantom auditory) speech stimuli was computed as:

$$d_{AV}(T_d) = \left\| LSP_A^{S_1,L_A} - LSP_V^{S_2+T_d,L_A} \right\|_2, \qquad (1)$$

where LSP_A and LSP_V were from A$_{/C1a/}$ and V$_{/C2a/}$, respectively. S_1 and S_2 represent the consonant onset points for A$_{/C1a/}$ and V$_{/C2a/}$, respectively. L_{A} approximated the consonant duration in A$_{/C1a/}$. T_d represents temporal shifting of A$_{/C1a/}$ across V$_{/C2a/}$. For the ***C***-alignment, the *phantom* acoustic segment began at the consonant onset point, and thus the T_d value was 0. For the ***V***-alignment, the *phantom* acoustic segment ended at the vowel onset point, and thus the T_d value was $L_V - L_A$. The ***M***-alignment was obtained by sliding acoustic signals relative to the video and finding a minimal distance point (see **Figure 1**). An implication from Equation 1 is that L_{A} has an effect on the distances: Acoustic /ba, da, ga/ tokens in general yield small intra-token distances. This method was our initial effort toward quantifying incongruity between auditory and visual speech signals, without implying any particular perceptual or neural representations.

To assess perceptual consequences of the three alignment methods, normal hearing perceivers identified the stimuli in an open-set identification task. To examine the possibility that the large difference between A$_{/ba/}$ and A$_{/la/}$ might draw attention to the audio and defeat visual influence effects, half of the participants viewed A$_{/ba/}$V and A$_{/la/}$V blocked presentations (blocked design), and another half viewed mixed A$_{/ba/}$V and A$_{/la/}$V presentations (mixed design). Behavioral results were examined in terms of response types, auditory-based accuracy, visual-based accuracy, and some other contributing factors (e.g., talker differences).

2. METHOD

2.1 Talkers

The talkers (with American English as a native language) were selected from a larger pool that had been initially screened for their visual intelligibility, as judged by deaf adults. Subsequent extensive additional visual-only speech perception testing, with 16 normal-hearing human subjects, of 320 sentences produced by each of these four talkers showed that F2 was the most intelligible, then M1, followed by M2 and F1. These results were replicated with eight deaf lipreaders, except that M2 was more intelligible than M1 [5].

2.2 Participants

Participants were ten adults (age 19-29 years, mean age 22 years; five females) with normal hearing, American English as a native language, and normal or corrected-to-normal vision. All were screened for their lipreading ability, but their scores were not used to bar entrance to the experiment, only to provide insight into the results. Testing was approved by an Institutional Review Board. Participants gave informed consent and were paid $10 per hour.

2.3 Speech Materials

The corpus was part of a larger database [5]. The original database obtained with the four selected talkers consisted of 69 consonant-vowel syllables. Each syllable was produced at least two times in a pseudo-randomly ordered list. For the present study, eight syllables /ba, da, ga, va, za, la, wa, ða/ were included. The voiced consonants were chosen, because they vary along places and manners of articulation. Two tokens (labeled with '1' and '2') for each syllable were selected for the present study [see 5].

2.4 Data Recording

The recording system comprised a production quality SONY video camera, a SONY recorder, a Qualisys 3-D three-camera motion capture system, a DAT recorder, and a directional Sennheiser microphone [5]. Lighting and positioning were carefully adjusted to obtain clear realistic recordings. The talkers looked directly into the camera, and their faces filled the monitor. The microphone was positioned to be out of the way of video. All of the recorded data streams were synchronized [5]. The audio sampling frequency was 48 kHz for the video recordings. The optical data and the audio from the DAT recordings were not used for the present study.

2.5 Consonant and Vowel Onsets Labeling

The ***C***- and ***V***-alignments of AV speech signals were achieved by hand labeling the consonant and vowel onsets. The acoustic features used for the labeling were transient, frication, aspiration, voicing, and high-frequency attenuation. Five consonants (/v, z, l, w, ð/) did not have bursts. So, consonant onset was instead defined as the consonant release point, which was either the beginning of vocal fold vibration or fricative noise.

Given the coarticulation effect, the vowel onset in /Ca/ was manually determined based on its spectrum and waveform. The consonants /b, d, g, v, z, ð/ had easily defined "boundaries" between consonant and vowel, which was the first vocal fold vibration after the aspiration noise. However, for /w, l/, a boundary was more difficult to define, and the high-frequency spectrum and waveform properties were combined to locate the vowel onsets. Two examples of consonant and vowel onset labeling are given in **Figure 2**. Two utterances (/da_1/ and /ga_2/ from Talker M2) had a voicebar that was a low-frequency hum. After labeling, non-speech sounds such as lip smacking and preceding voicebars were deleted (set to zero).

Figure 3 displays the consonant and vowel durations for /Ca/ syllables. The mean consonant and vowel durations, respectively, were 102 ms and 439 ms for Talker M1, 76 ms and 319 ms for Taker F1, 80 ms and 388 ms for Talker M2, and 79 ms and 374 ms for Talker F2. Talker F1 produced shorter vowels than other talkers. /ba, da, ga/ syllables had short consonant durations. Consonant durations in /wa, ða/ were different across talkers.

2.6 Generating AV Speech Stimuli

2.6.1 Digitizing Video Tapes

The video recordings on BETACAM tapes were digitized using an ACCOM real-time digital disk recorder. Uncompressed video images (740x486) were transferred to a PC as individual frame files. The corresponding acoustic tokens (48 kHz) were also transferred to the PC as individual files. These sounds were normalized (based on average RMS levels derived from A-weighted spectra).

2.6.2 AV Pairing, Synchrony, and Distance

For each talker, AV stimuli were generated by dubbing $V_{/C2a/}$ to $A_{/b1a/}$ and $A_{/l1a/}$, and by dubbing $V_{/C1a/}$ tokens to $A_{/b2a/}$ and $A_{/l2a/}$. Therefore, every stimulus involved dubbing (e.g., $V_{/b1a/}$ and $A_{/b2a/}$).

In addition, each dubbing was achieved with ***C***-, ***V***-, and ***M***-alignments that were derived using Equation 1. For this purpose, acoustic signals from video recordings that were originally sampled at 48 kHz were down-sampled to 16 kHz. Speech signals were then divided into frames. The frame length and shift were 24 ms and 2.8 ms, respectively. For each acoustic frame, 16th-order LSPs [including the **log**(*Energy*)] were obtained [see 5]. In total, there were 384 stimuli generated (8 $V_{/Ca/}$ x 2 $A_{/Ca/}$ x 2 tokens x 3 alignments x 4 talkers). **Figure 4** and **Figure 5** display the alignments and the corresponding AV distances. The distances were smaller for $A_{/ba/}$ than for $A_{/la/}$. The ***C***- and ***V***-alignments were different when the auditory and visual consonants had different durations. For example, in **Figure 4**, the ***C***-, ***M***-, and ***V***-alignments of $V_{/za/}$ (the 5th cluster) with A/$_{ba2}$/ of Talker F2 (the 8th row) were different (i.e., having different vertical positions), and they produced different AV distances (i.e., different bar widths).

2.6.3 AV Encoding

For the video images, the top and bottom three lines were cropped, and the sequence of uncompressed frames for each stimulus was built into an AVI file that was compressed using the LIGOS LSX MPEG-Compressor. The resulting video clips had an image size of 720x480, a frame rate of 29.97 Hz, and a constant bitrate of 7700 Kbits/sec. These video clips were concatenated to create a single large video file that was authored to a DVD using the SONIC ReelDVD. As with the video, all of the audio files were concatenated into a single long file for production of the DVD. Audio concatenation was performed using custom software that ensured frame locked audio of 8008 samples per 5 video frames.

2.7 Procedure

AV stimuli were presented using a Pioneer DVD player and were displayed on a 14" SONY Trinitron monitor at a distance of about one meter from the participant. Audio was presented over calibrated TDH-49 headphones at a level of 65 dB SPL that was checked before and after each session.

Participants performed an open-set identification task by typing their responses using a computer keyboard. Participants were instructed to <u>*watch*</u> and <u>*listen to*</u> the talkers, and then identify the consonant or consonants that they <u>*heard*</u>. Participants were directed to guess if necessary. An experimenter monitored the participants during testing.

At the beginning of each session, instructions were displayed on a PC monitor. After acknowledging having read the instructions, a computer program presented each of the stimuli and recorded behavioral responses. Following each stimulus, a black filled frame was displayed on the video monitor, and an input box was displayed on the PC monitor. After typing and double-checking the response, participants pressed the "ENTER" key to switch to the presentation of the next token. Participants were instructed to report any mistyping during breaks. No feedback was given at any time.

The presentation of the 384 tokens in each session was administrated in two experimental designs.

$A_{/ba/}V$ and $A_{/la/}V$ mixed design (mixed design). The 384 tokens were blocked by talkers, and thus there were four blocks. Each block consisted of 96 tokens, which comprised both $A_{/ba/}V$ and $A_{/la/}V$. Each block took about 10 minutes.

$A_{/ba/}V$ and $A_{/la/}V$ blocked design (blocked design). The 384 tokens were first blocked by talkers and then by auditory types ($A_{/ba/}$ or $A_{/la/}$), and thus there were eight blocks. Each block consisted of 48 tokens of $A_{/ba/}V$ only or $A_{/la/}V$ only from one talker. Each block took about 5 minutes.

Half of the participants were tested on the mixed design, and the other half on the blocked design. The talker order was assigned randomly in each session. For the blocked design, the order of the $A_{/ba/}V$ and $A_{/la/}V$ blocks was randomized within each talker. Within each block, the tokens were randomly ordered.

Participants were tested one session on each day and totally ten sessions. Each participant contributed a total of ten responses for each stimulus token. A five-minute break was given between blocks to prevent fatigue. Instruction on phonemic labeling and a practice set of 16 trials were given on Day 1.

3. RESULTS

Figure 6 shows the frequencies of the majority of responses to $A_{/ba/}V$ and $A_{/la/}V$. For $A_{/ba/}V$, the responses were typically individual consonants. Among the 23 consonants /y, w, r, l, m, n, p, t, k, b, d, g, h, θ, ð, s, z, f, v, ʃ, ʒ, tʃ, dʒ/, six consonants /ʃ, ʒ, tʃ, dʒ, y, h/ were not reported (individually or in combination). For $A_{/la/}V$, there were many combination responses (e.g., /bl/). These combination responses were not symmetric. That is, there were no /lb/ responses for $A_{/la/}V_{/ba/}$, although their ***C***-, ***V***-, and ***M***-alignments were different. At the completion of the experiment, some participants reported having noticed mismatches between auditory and visual stimuli [1].

The results were tallied in terms of **a**uditory-based **a**ccuracy (***AA***; e.g., the response to $A_{/ba/}V_{/wa/}$ was "ba"), **v**isual-based **a**ccuracy (***VA***; e.g., the response to $A_{/ba/}V_{/va/}$ was "va"), **v**oice**l**ess responses (***VL***), and **co**mbinations (***CO***; e.g., "bla"). In order to determine which, if any, main factors were significant, each measure was submitted to an omnibus mixed measures analysis of variance, with video (7; excluding $A_{/ba/}V_{/ba/}$ or $A_{/la/}V_{/la/}$), talker (4), pairing (2), and alignment method (3) as the repeated factors. The stimulus presentation design (blocked versus mixed design) was the between-subject factor. $A_{/ba/}V$ and $A_{/la/}V$ were analyzed separately.

F-test values are listed in **Table 1** for all ANOVAs. The results showed that pooling across alignment and presentation design was permissible. In addition, the pairing effect (see Section 2.6.2) was mainly due to an artifact in the /la$_1$/ sound spoken by Talker F2. Therefore, the ***AA***, ***VA***, ***VL***, and ***CO*** measures examined below (see **Figure 7** and **Figure 8**) were pooled across the three alignments, two pairings, and ten participants.

		$A_{/ba/}V$				**$A_{/la/}V$**			
	(N_1,N_2)	***AA***	***VA***	***VL***	***CO***	***AA***	***VA***	***VL***	***CO***
Video	(6, 3)	22.4	23.2	2.8	1.7	5.6	5.2	1.2	2.2
Talker	(3, 6)	10.1	2.7	4.1	2.3	17.4	3.1	1.5	26.3
Pairing	(1, 8)	5.3	.0	1.9	.8	29.2	.0	.7	11.1
Alignment	(2, 7)	.3	1.6	2.2	2.0	1.4	3.2	1.2	.9
Presentation design	(1, 8)	1.1	.3	.0	.2	.3	1.0	.4	.3

Table 1. ANOVA results (*F* values) with different performance measures and different factors. (N_1, N_2) represent degrees of freedom. The shaded areas indicate significant effects ($p < .05$).

Figure 7 and **Figure 8** showed that in general the $A_{/ba/}$ produced more visually influenced responses (i.e., fewer ***AA***, more ***VA***, and more ***VL*** responses) than did $A_{/la/}$. $A_{/ba/}$ produced more visually dominant responses (with $V_{/va/}$, $V_{/ða/}$, and $V_{/da/}$) than did $A_{/la/}$ (with $V_{/va/}$). $A_{/ba/}$ produced more voiceless responses and fewer combination responses than did $A_{/la/}$. The scoring of the responses in terms of ***AA*** versus ***VA*** versus ***CO*** implies that whenever the number of all of these types of responses was low, the number of McGurk responses was high. Across the two types of audio stimuli, there were more McGurk responses for $A_{/ba/}V$ (e.g., $A_{/ba/}V_{/ga/}$, $A_{/ba/}V_{/za/}$, and $A_{/ba/}V_{/la/}$). Most of the $A_{/la/}V$ stimuli showed no visual

[1] The reported mismatches were $A_{/ba/}V_{/wa/}$ (4 participants), $A_{/ba/}V_{/fa/}$ (1), $A_{/la/}V_{/ma/}$ (1), $A_{/ba/}V_{/θ/}$ and $A_{/la/}V_{/wa/}$ (1), $A_{/la/}V_{/na/}$ and $A_{/ma/}V_{/bla/}$ (1). One participant rarely noticed any mismatch. Another participant noticed the mismatches, but could not give an example.

influence. But when there was a visual influence, it was most likely the combination type.

If **Figure 4**, which shows distances, is compared with **Figure 7**, which shows responses for $A_{/ba/}V$, an overall pattern of relationships can be seen. In particular, when visual distance was small, many responses were visually influenced (i.e., ***VA***, ***VL***, ***CO***). If **Figure 5**, which shows distances, is compared with **Figure 8**, which shows responses for $A_{/la/}$, a similar overall pattern can be seen. In general, also, AV distances for $A_{/ba/}V$ (**Figure 4**) were smaller than those for $A_{/la/}V$ (**Figure 5**).

A more detailed examination of distance versus visual influence yielded additional systematic effects. For $A_{/ba/}$, AV distances for Talker F2 were smaller than those for other talkers, and Talker F2's $A_{/ba/}V$ stimuli yielded more visual influence than those of other talkers [**Figure 7**(a)]. For $A_{/ba/}V_{/da/}$, AV distances for Talkers M2 and F2 were smaller than those for other talkers, and their tokens yielded more visual influence [**Figure 7**(b)]. **Figure 8**(a) showed that the perception of $A_{/la/}$ was affected by $V_{/ba/}$, $V_{/va/}$, and $V_{/wa/}$. For $A_{/la/}$, AV distances for Talker F2 were smaller than those for other talkers, and Talker F2's $A_{/la/}V$ stimuli yielded more visual influence than those of other talkers [**Figure 8**(a)]. In addition, the influence of $V_{/va/}$ and $V_{/wa/}$ to $A_{/la/}$ agrees with their smaller AV distances (**Figure 5**).

Figure 7(c) and **Figure 8**(c) show voiceless responses to the AV stimuli. $A_{/ba/}V$ stimuli of Talker F1, who has the lowest visual intelligibility ratings, yielded the largest number of voiceless responses. In general, $A_{/ba/}V$ yielded more voiceless responses than $A_{/la/}V$. $V_{/ba/}$ and $V_{/wa/}$ produced less voiceless responses for $A_{/ba/}$ and $A_{/la/}$ than the other six $V_{/Ca/}$.

Figure 7(d) and **Figure 8**(d) show combination responses to the AV stimuli. In general, $A_{/la/}$ yielded more combinations than $A_{/ba/}$. $A_{/la/}$ stimuli tended to yield combination responses with $V_{/ba/}$, $V_{/va/}$, and $V_{/wa/}$. As mentioned earlier, the artifact in the /la$_1$/ sound spoken by Talker F2 appeared to be responsible for many combination responses.

4. SUMMARY AND DISCUSSION

$A_{/ba/}V$ and $A_{/la/}V$ mixed and blocked designs were not significantly different. Therefore, the attentional effect was not a main effect in the experimental design. Behavioral results indicated that the alignment effect was not significant. However, ***C***-, ***V***-, and ***M***-alignments resulted in large differences in distance measures using Equation 1. Thus, although alignment was not a significant perceptual factor in the current study, it is possible that other response measures might be more sensitive and produce alignment effects. The visual influence on acoustic tokens varied as a function of syllable type. $A_{/ba/}$ was more influenced by visual stimuli than $A_{/la/}$. Visual influence on $A_{/ba/}$ was of the McGurk-type and/or of voicing confusion; while visual influence on $A_{/la/}$ was of the combination type.

5. ACKNOWLEDGMENTS

This research was supported in part by NIH/NIDCD (DC006035) and by NSF. We acknowledge the help of B. Chaney and J. Jordan. The views expressed here are those of the authors and do not necessarily represent those of NSF and NIH.

6. REFERENCES

[1] Bernstein, L. E., Auer, E. T., Jr., & Moore, J. K. (2004). Audiovisual speech binding: Convergence or association? In G. A. Calvert, C. Spence & B. E. Stein (Eds.), The handbook of multisensory processes (pp. 203-223). Cambridge, MA: MIT Press.

[2] Fowler, C. A. (1986). An event approach to the study of speech perception from a direct-realist perspective. J. Phonetics, 14(1), 3-28.

[3] Grant, K. W., Greenberg, S., Poeppel, D., et al. (2004). Effects of spectro-temporal asynchrony in auditory and auditory-visual speech processing. Seminars in Hearing, 25, 241-255.

[4] Green, K. P., & Norrix, L. W. (2001). Perception of /r/ and /l/ in a stop cluster: Evidence of cross-modal context effects. J.E.P.: H.P.P., 27(1), 166-177.

[5] Jiang, J. (2003). Relating optical speech to speech acoustics and visual speech perception. Unpublished Doctoral Dissertation, University of California, Los Angeles, CA.

[6] Massaro, D. W. (1998). Perceiving talking faces: From speech perception to a behavioral principle. Cambridge, MA: MIT Press.

[7] Massaro, D. W., Cohen, M. M., & Smeele, P. M. T. (1996). Perception of asynchronous and conflicting visual and auditory speech. JASA, 100(3), 1777-1786.

[8] McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. Nature, 264(5588), 746-748.

[9] Munhall, K. G., Gribble, P., Sacco, L., et al. (1996). Temporal constraints on the McGurk effect. Percept Psychophys, 58(3), 351-362.

[10] Schwartz, J.-L., Robert-Ribes, J., & Escudier, P. (1998). Ten years after Summerfield: A taxonomy of models for audio-visual fusion in speech perception. In R. Campbell, B. Dodd & D. Burnham (Eds.), Hearing by eye II: Advances in the psychology of speechreading and auditory-visual speech (pp. 85-108). East Sussex, UK: Psychology Press.

[11] Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. JASA, 26, 212-215.

[12] van Wassenhove, V., Grant, K. W., & Poeppel, D. (2003). Electrophysiology of auditory-visual speech integration. Proc. AVSP 2003, St. Jorioz, France.

[13] Wright, T. M., Pelphrey, K. A., Allison, T., et al. (2003). Polysensory interactions along lateral temporal regions evoked by audiovisual speech. Cerebral Cortex, 13(10), 1034-1043.

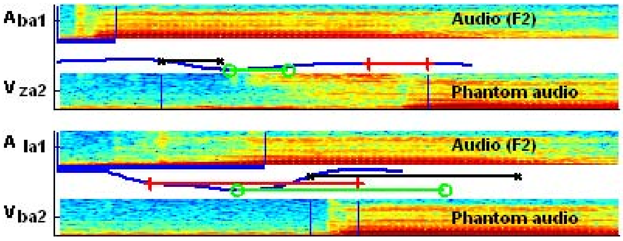

Figure 1. AV alignments and distances for $A_{/ba1/}/V_{/za2/}$ (upper panel) and $A_{/la1/}/V_{/ba2/}$ (lower panel). The blue line between the Audio and the *phantom* Audio represents the distance measure. Lines with 'x' (black line), '+' (red line), and 'o' (green line) represent ***C***-, ***V***-, and ***M***-alignments, respectively.

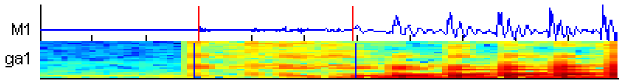

Figure 2. Consonant and vowel onsets for /ga$_1$/ (Talker M1).

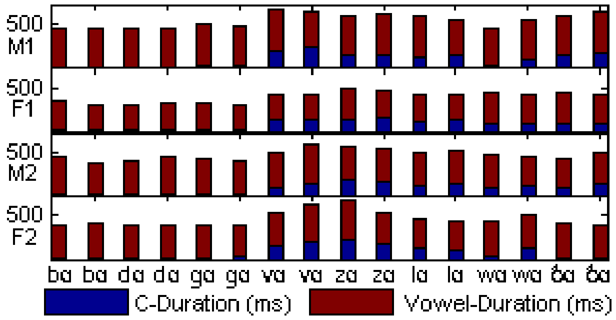

Figure 3. Consonant and vowel durations (two tokens per /Ca/).

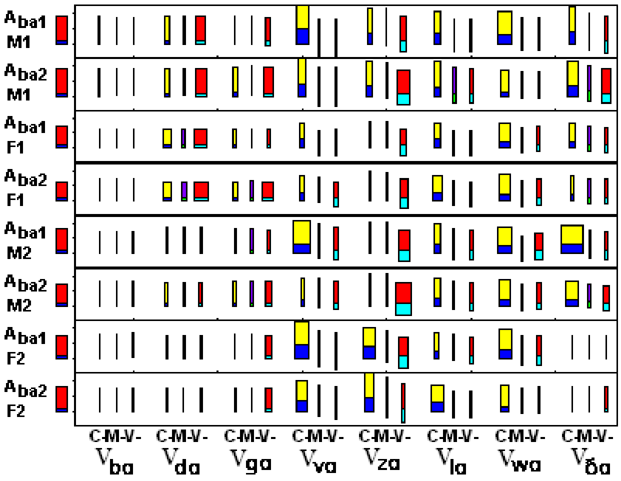

Figure 4. AV alignments and distances for $A_{/ba/}V$. Each row comprises data for one $A_{/ba/}$ token from one talker. Each vertical bar represents the proportions of time measured for the vowel versus the consonant (see text). The lower part represents the consonant segment, and the upper part represents the vowel segment. The **X axis** labeling represents the different alignments (***C***-, ***M***-, and ***V***-alignments) for $V_{/Ca/}$ ($V_{/ba/}$, $V_{/da/}$, $V_{/ga/}$, $V_{/va/}$, $V_{/za/}$, $V_{/la/}$, $V_{/wa/}$, and $V_{/ð/}$). The width in the **X** direction of a bar indicates the magnitude of the distance.

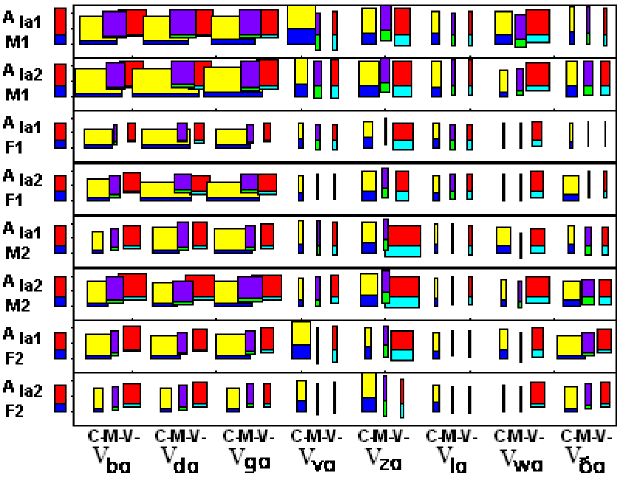

Figure 5. AV alignments and distances for $A_{/la/}V$.

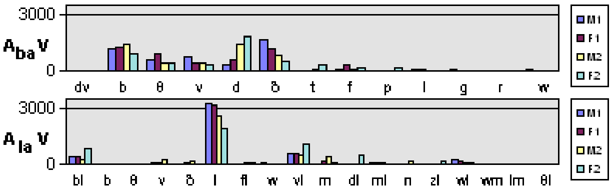

Figure 6. Frequencies (**X axis**) of the various responses (**X axis**) to $A_{/ba/}V$ and $A_{/la/}V$. The infrequent (not more than 10 times for $A_{/ba/}$ or $A_{/la/}$ sound) response types are not listed.

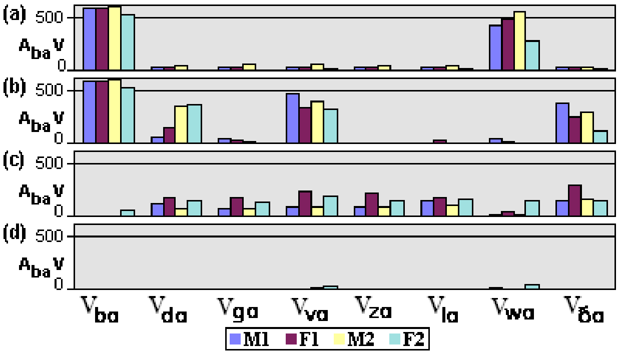

Figure 7. The number (**Y axis**) of **(a)** auditory-based correct responses, **(b)** visual-based correct responses, **(c)** voiceless responses, and **(d)** combination responses with different visual stimuli (**X axis**) for $A_{/ba/}V$.

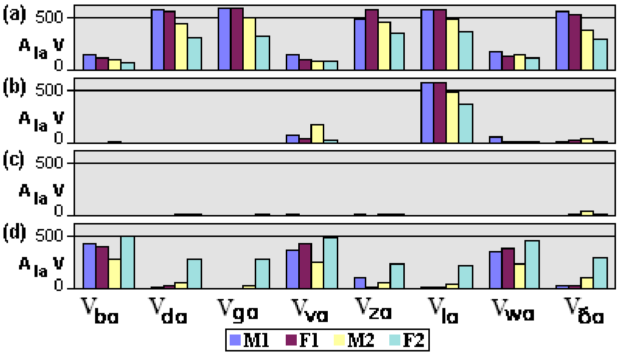

Figure 8. Behavioral responses for $A_{/la/}V$ (refer to **Figure 7**).