Abel, J., Barbosa, A. V., Black, A., Mayer, C., & Vatikiotis-Bateson, E. (2011). The labial viseme reconsidered: Evidence from production and perception. In Y. Laprie & I. Steiner (Eds.), 9th International Seminar on Speech Production (ISSP) (pp. 337-344). Montreal, PQ.

The Labial Viseme Reconsidered: Evidence from Production and Perception*

Jennifer Abel, Adriano Vilela Barbosa, Alexis Black, Connor Mayer, Eric Vatikiotis-Bateson

Linguistics – University of British Columbia
2613 West Mall, Vancouver, British Columbia, Canada

adriano.vilela@gmail.com, akblack2g@gmail.com,
connorm@interchange.ubc.ca, evb@interchange.ubc.ca

Abstract. *This study examined the degree to which visual /p,b,m/ are different in production and perception. A simple kinematic measure of orofacial motion derived from optical flow analysis of video demonstrates systematic differences between the three bilabials. Perceptual evaluation of /p,b,m/ presented audiovisually in degraded acoustics shows that identification of /b/ is especially difficult.*

1. Introduction

Beginning with Woodward & Barber (1960), the English labial stops, /p, b, m/, have been known to be difficult, if not impossible, for humans to discriminate visually from one another, although they are easily distinguished from other phonemes. This led to *viseme* classification (Fisher 1968), where phonemes are grouped according to their visually indiscriminablility from one another and collective distinctiveness from other groups (typically at different places of articulation). The validity of viseme classification, particularly for bilabial stops, has been widely accepted in both psychological and engineering approaches to auditory-visual speech processing – e.g., automated visual speech recognition, (Bregler et al. 1993; Potamianos et al. 2003, inter alia). This outcome may be unfortunate on several counts. First, many definitions of viseme are based on a 70-75% within-group confusability of the phonemes in question (e.g., Walden et al. 1977; Owens and Blazek 1985). This cut-off indicates that discrimination within a viseme is difficult, not that it is impossible. Second, psychometric viseme identification has largely been based on video-only evaluation, which we know to be problematic (for overview, see Munhall and Vatikiotis-Bateson 2004). Third, failure to perceive a difference – or more to the point, a failure to demonstrate a perceptual difference – does not mean no systematic difference was produced. To wit, there has been a fair amount of evidence of observable differences in the production of labial stops (Fujimura 1961, Sussman et al. 1973, Vatikiotis-Bateson & Kelso 1984, Summers 1987). At least one implication of this is: if systematic differences in production can be measured, they should be useful in machine-based recognition of audiovisual speech.

* Supported by NSERC, SSHRC, and CFI grants. Thanks to Molly Babel for skilled tutelage in the use of R.

In this paper, we use a simple optical flow analysis (Horn and Schunk 1981) of visible face motion to show that there are indeed systematic differences in the production of the three bilabial stops, /p, b, m/. Surprisingly, these differences are visible from both front and profile views. We also reconfirm that it is difficult, but not impossible, for perceivers to identify bilabial stops produced in sentential contexts and presented audiovisually with a severely degraded audio signal (that was nonetheless recognizable as speech).

2. Study 1: Production

2.1 Methods

Two studies were conducted, a pilot and a larger follow-up. One talker participated in each study.

Talkers were middle-aged females (Study 1 talker (S1): western Canadian, Study 2 talker (S2): midwestern USA). S2 was previously assessed as highly intelligible under similar conditions (Eigsti, Vatikiotis-Bateson, Munhall, and Yano 1995).

Materials. S1 pronounced V_1BV_1 pseudo-words containing the vowels /i, a, u/ and consonants /p, b, m/ or distractors /f, t, k/ in the carrier phrase, "Say ____ again". Five tokens of each bilabial pseudo-word and one token of each non-bilabial pseudo-word were recorded. S2 produced CV_1BV_2 words and pseudo-words in various sentence positions (early, middle, final) , e.g., "The _____ is in the cupboard", "I saw the ____ in the cupboard", and "Look in the cupboard for the ____". The initial C was /t/ or /s/; V_1 was /ae, e, ɛ, i, ɪ, o, u, ʌ/; B was /p, b, m/ or distractor /s, ɾ/; V_2 was /i, o, ə, ɚ/. Five tokens of each word were produced in each of the three sentence positions. Stress was placed on the initial syllable for all tokens in both studies.

Procedure. Broadcast quality video (1280x720 pixels at 60 fps progressive) was recorded for front and profile views of each talker's face and neck. For S1, one camera was used along with a mirror placed at a 45 degree angle at the right side of the talker's face; for S2, two cameras were used, oriented in front of and 90 degrees to the side of the talker. Sound and video were recorded directly to disk using the Apple Pro Res HQ hardware encoding (AJA Ki-Pro).

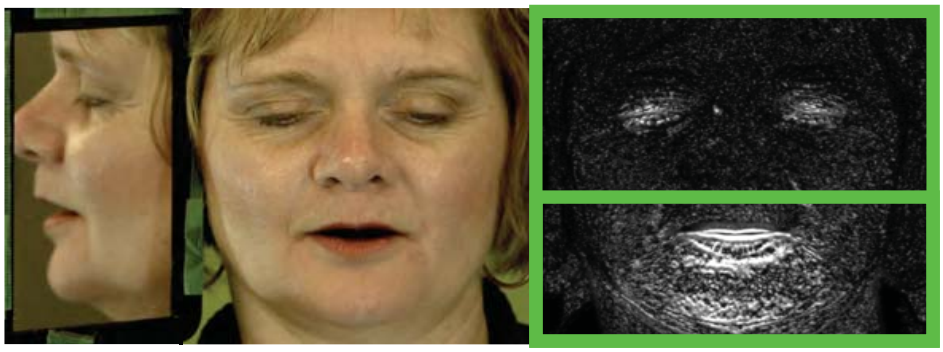

Figure 1. Video frame from Study 1 showing front and profile views; optical flow image showing pixel motion (white for more motion) between video frames within two ROIs – *lower face* and *whole face*.

2.2 Results

Optical flow analysis (Horn and Schunk 1981) was used to estimate pixel motion within regions of interest (ROI) in the video for profile and front views. Since optical flow is the difference in pixel position from one frame to the next, the estimated kinematic measure is change of pixel position or velocity. Two ROIs were defined: lower face (from just below the nose to just below the chin) and whole face (the entire video frame), as shown in Figure 1. Both lower and whole face ROIs were measured and analyzed for S2; only the lower face measures were used for S1. For each ROI, the magnitudes of all velocities were summed independently for the horizontal (x) and vertical dimensions (y) for each frame step in the video sequence.

2.2.1 Peak Velocities

Peaks in the resulting summed-velocity time series for x and y were picked semi-automatically (i.e., corrected by hand). Working from a screen display like that shown in Figure 2 and using the audio signal as a guide, one velocity peak was picked for each segment transition in the CV_1BV_2 (= four peaks total) and the V_1BV_1 (= three peaks total) tokens. For example, for the CV_1BV_2 tokens produced by S2, the peaks correspond to the transitions from C to V_1 (Peak 1), V_1 to B (Peak 2), B to V_2 (Peak 3), and out of V_2 (Peak 4). Horizontal and vertical motion peaks extracted for profile and front views were tested using repeated-measures ANOVA in R (R Core Development Team 2011) for the effects of phoneme (p, b, m, or other) and position in the sentence. Context effects of vowel identity were not tested.

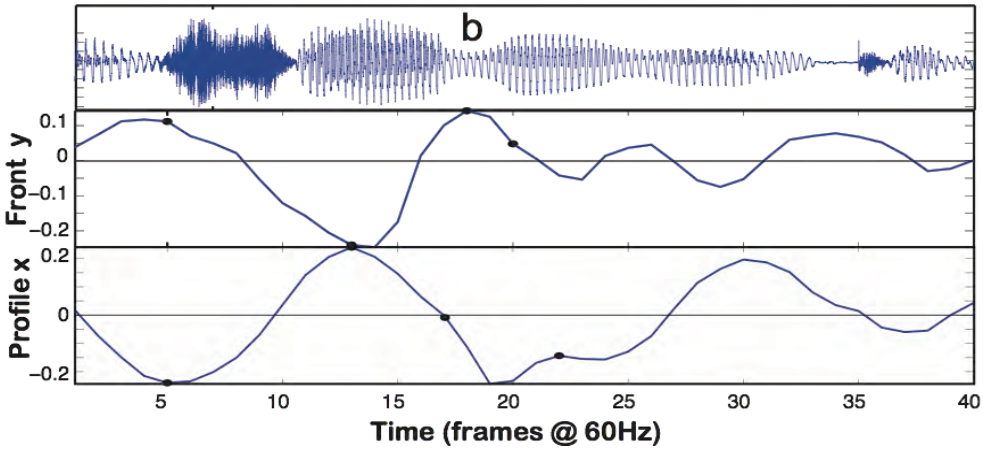

Figure 2. Summed optical flow for *front* y (vertical) and *profile* x (anterior-posterior) during production of /sæbi/ by S2. Four "peaks" are marked in each time-series. /b/ closure is marked in the audio signal.

For both S1 and S2, reliable phoneme effects were obtained only for summed velocity Peak 4 in the lower face ROI. For S2, reliable effects of position were also obtained. There were no interactions between phoneme and position for either subject.

2.2.2 Area Under the Curve

Area under the curve (AUC) provides an estimate of the total motion associated with each transition. AUC) was calculated for each summed-velocity peak by identifying the

value closest to zero on either side of the peak and summing all intervening values. Only periods surrounding Peaks 2-4 were used. These correspond to the CV_1 transition (Peak 2), the V_1 B transition (Peak 3) and the B V_2 transition (Peak 4).

Repeated measures ANOVAs evaluated the effects of *phoneme* (*p, b, m,* and *other*) and/or *position* (*early, middle, late)* for the three AUCs of the three time periods surrounding Peaks 2-4. For S1, sentence position was not included as a factor in the analysis, and only lower face values were measured. Reliable effects of *phoneme* on AUC were found in both front and profile views. Figure 3 shows results for vertical motion in the profile view.

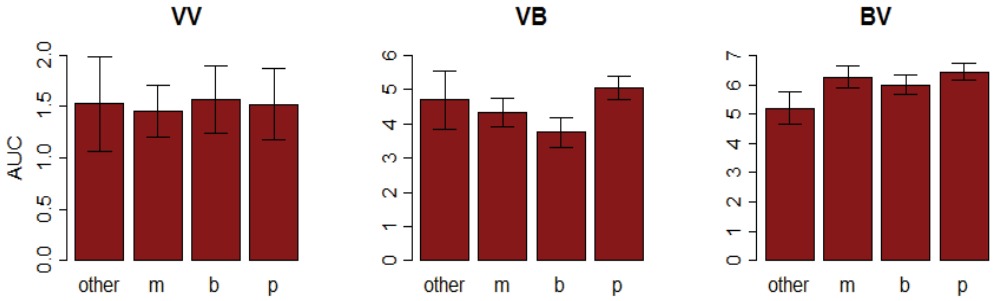

Figure 3. AUC results for S1's vertical motion in profile view are plotted with standard error by *phoneme* for the periods around Peaks 2-4.

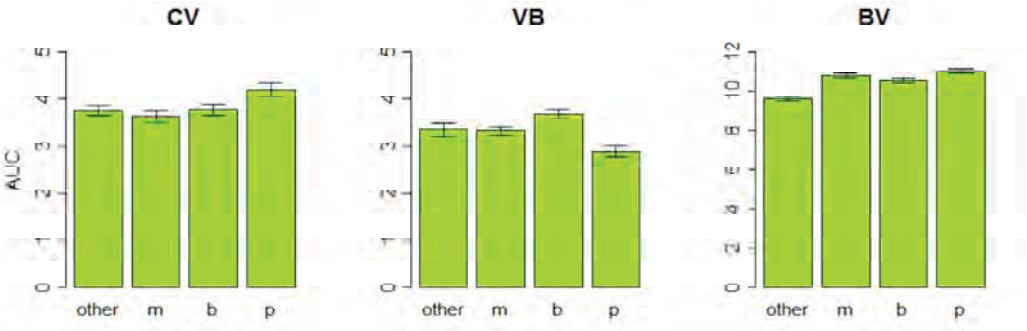

Figure 4. AUC results for S2's vertical motion in front view are plotted with standard error by *phoneme* for the periods around Peaks 2-4.

For S2, AUCs were analyzed for both ROIs. For the front view, only vertical motion was analyzed. There were *phoneme* effects on all three AUCs for both lower and whole face regions. Position in the sentence was reliable for the transitions going into the vowels (CV_1 and BV_2), but not going into the bilabial (V_1 B). In the profile view, both vertical and horizontal motions were analyzed. The horizontal motion analysis produced reliable effects of both *phoneme* and *position* on all three AUCs for both ROIs. Figure 4 provides an example of the phoneme effects.

2.3 Discussion of production results

A single time-varying measure of motion extracted from video demonstrates that there **are** visible differences due to bilabial identity regardless of the position in the sentence where the word containing the bilabial occurs, and regardless (to a surprising degree) of the angle of view. Note also that reliable differences in AUC are distriubted throughout

the target word containing the bilabial segment.

3 Study 2: Perception

The perception study consisted of two conditions, in which participants were exposed to a talker's bilabial productions from either the front or profile views. Initial results based on 20 participants (10 per condition) indicated a significant bias against /b/ identification; the study was therefore replicated with an additional 17 participants (9 in the front view condition) using a different arrangement of the button box keys to control for this possible artefact.

3.1 Methods

Forty participants were recruited from the University of British Columbia. Participants were native speakers of English with no reported history of speech or hearing deficits, and were paid for their participation. Three participants were excluded from the analysis due to mechanical and experimental errors.

In both conditions (front and profile), participants were shown excised portions of the same video clips as used in the production analysis (Study 2, n=397). Clips were 40 frames (.667 seconds) long, and included a small part of the preceding and following words (or sentence-final closure). The audio of each video was masked by pink noise (from Petersen 2004) for use with Praat. Participants were prompted at the end of each trial to identify whether the middle consonant of the video clip was a P, B, M or "other" (i.e., S or T). Their response (via button box) initiated the next trial.

The relative signal-to-noise ratio was set according to each participant's performance on an audio-only pre-experiment task. Participants heard the talker's production of an entire sentence masked by a predetermined amount of noise. Participants then chose, from a list of three words, the word they thought matched the presentation. Signal levels were adjusted depending on participants' performance, until they plateaued at a 30% accuracy rate of response.

3.2 Results

Figure 5 displays mean accuracy rates by condition and experiment type. Since the first experiment revealed a significant bias against responding "B," it was hypothesized that this might have due to the arrangement of the answer keys on the button box: the "B" button was the leftmost of the possible button options. To control for this factor, the experiment was replicated with the keys rearranged so that "P" would sit at the left-hand edge. This rearrangement greatly reduced, but did eliminate, the bias to avoid "B" ("B" vs "P": $t(6603)=-3.54$, $p<.001$; "B" vs "M": $t(6603)=-2.05$, $p=.041$). No such avoidance of "P" was observed in the second experiment, as would be expected if key arrangement alone is the source of perceiver bias.

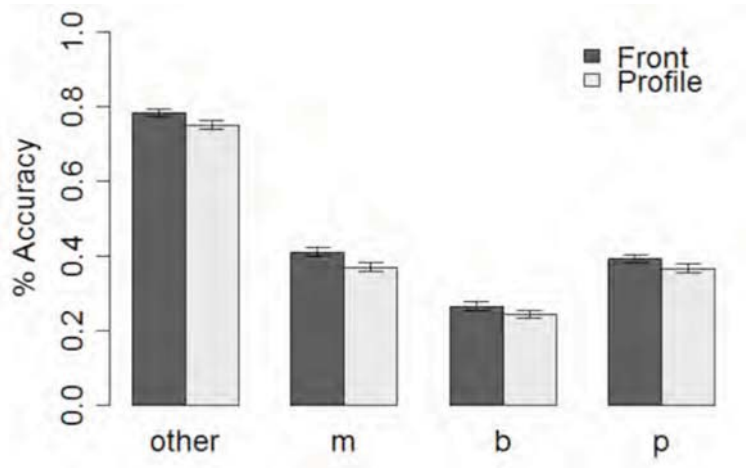

Figure 5. Mean accuracy rates by phoneme condition for two views.

To test if perceiver identifications were better than chance, *chi*-square and *phi* scores were calculated for each condition and for each participant. Both targets and response types of the category "other" were removed from the tests as it was expected that perceiver identification of non-bilabial consonants would be significantly higher than bilabial identification, and would thus skew results. Collapsing across participants, we find significant chi-square models in the first experiment for both Front-view ($\chi(4,2709)=80.16$, $p<.001$) and Profile-view ($\chi(4,2488)=24.48$, $p<.001$ conditions. In the second experiment, the Profile-view yielded a significant chi-square model ($\chi(4,2082)=12.78$, $p=.012$), while the Front-view reached marginal significance ($\chi(4,2498)=8.70$, $p=.069$). These results suggest that perceivers were successful at identifying some of the bilabial consonants at levels better than chance. To determine which of the consonants were being successfully identified, we ran two-by-two chi-square tests for each condition. For the front view conditions in both experiments, and the profile view condition in experiment two, only those chi-square models that involved discrimination of "P" and "M" were significant. In other words, in both front and profile conditions, irrespective of button-box key order, perceivers failed to identify "B," but performed at higher than chance levels of identifying "P" and/or "M." For the profile view of the first experiment, chi-square models examining discrimination of all three consonants were significant.

Individual performance was investigated using the same methods described above. Discrimination between P and B caused the greatest amount of confusion across all participants (i.e. only 3 participants achieved significant identification of P and/or B as shown by individual two-by-two chi-square tests). Cross-comparisons of individual confusion matrices suggest that perceivers employ different strategies. Some are highly biased to respond with a single phoneme, usually either P or M (see Figure 6, left). Others correctly identify two of the three phonemes (Figure 6, right). Finally, some perceivers are not able to perceive bilabial differences, and respond with no bias (not shown).

An important difference between this experimental design and previous studies that have investigated bilabial discrimination/identification is the presence of the real-time audio signal accompanying the video. To determine the influence of this audio signal on participant performance, a Pearson's correlation was calculated between individual phi scores (derived from the cross-tabulation tables used in the chi-square

analyses) and the level of audio each participant received. Two participants with phi scores more than three standard deviations above the mean were excluded from this analysis. The results show a small, positive relation between participant performance and audio level ($r(34)=.50$, $p=.005$).

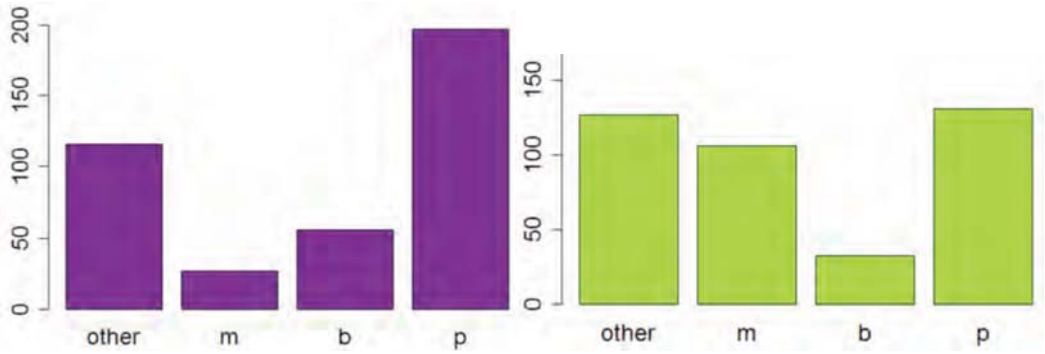

Figure 6. Examples of individual response strategies: left – /p/ bias; right – /p/ and /m/ bias.

3.3 Discussion

The perception study shows, unsurprisingly, that accurate identification of the bilabial consonants is a difficult task. While participants identified "other" consonants between 60% and 80% of the time, correct bilabial identifications hovered between 20% and 30%. Despite this difficulty, however, perceivers are more likely to correctly perceive /p/ and/or /m/ than they are /b/, even under severely degraded audio conditions. We note, however, that performance was negatively correlated with the severity of this audio degradation. This may reflect participants' use of audio cues when given more audio information (e.g. resonance or voiceless bursts may have been more salient through the pink noise than any acoustic cues associated with /b/). Given the low signal-to-noise ratio each participant experienced, however, it is also possible that the poorer performance associated with lower signal levels reflects a divide between speech-oriented processing and visual-only processing. In other words, despite our pre-experiment task, it is possible that participants were not able to perform at a 30% lexical recovery rate when exposed to the shorter clips.

Summary

Human perceivers may have trouble seeing what our measures see, but a machine system is surely capable of capturing the same visible differences in bilabial stops that we observed using optical flow analysis of crudely summed orofacial motion.

References

Bregler, C., Hild, H., Manke, S. and Waibel, A. Improving Connected Letter Recognition by Lipreading. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing* I557-I560, 1993.

Cohen J.D., MacWhinney, B., Flatt, M. and Provost, J. PsyScope: A new graphic interactive environment for designing psychology experiments. *Behavioral Research Methods, Instruments, and Computers* 25(2), 257-271, 1993.

Eigsti, I.-M., Vatikiotis-Bateson, E., Yano, S., & Munhall, K. G. (1995). Effects of listener expectation on eye movement behavior during audiovisual perception. *JASA, 97*, 3286.

Fisher, C.G. Confusion among Visually Perceived Consonants. *Journal of Speech and Hearing Research* 11.796-804, 1968.

Fujimura, O. Bilabial Stop and Nasal Consonants: a Motion Picture Study and its Acoustical Implications. *Journal of Speech and Hearing Research* 4.233-247, 1961.

Horn, B. K. P., and Schunck, B. G. Determining Optical Flow. *Artificial Intelligence*, 17.185-203, 1981.

Munhall, K. G., & Vatikiotis-Bateson, E. (2004). Spatial and temporal constraints on audiovisual speech perception. In G. Calvert, C. Spence & B. Stein (Eds.), *The handbook of multisensory processes* (pp. 177-188). Cambridge, MA: MIT Press.

Owens, E., and Blazek, B. Visemes observed by hearing-impaired and normal-hearing adult viewers. *Journal of Speech and Hearing Research* 28.381-393, 1985.

Petersen, N.R. Create-waveforms [Praat script], retrieved 15 February 2011 from http://www.linguistics.ucla.edu/faciliti/facilities/acoustic/praat.html. 2004.

Potamianos, G., Neti, C., Gravier, G., Garg, A., and Senior, A.W. Recent Advances in the Automatic Recognition of Audiovisual Speech. *Proceedings of the IEEE* 91(9).1306-1326, 2003.

Summers, W.V. Effects of stress and final-consonant voicing on vowel production: Articulatory and acoustic analyses. *Journal of the Acoustical Society of America* 82.847-863, 1987.

Sussman, H.M., MacNeilage, P.F., and Hanson, R.J. Labial and Mandibular Dynamics during the Production of Bilabial Consonants: Preliminary Observations. *Journal of Speech and Hearing Research* 16.397-420, 1973.

Vatikiotis-Bateson, E., & Kelso, J. A. S. (1984). Remote and autogenic articulatory adaptation to jaw perturbations during speech. *Jour. Acous. Soc. Am., 75*, S23-24.

Walden, B.E., Prosek, R.A., Montgomery, A.A., Scherr, C.K., and Jones, C.J. Effects of training on the visual recognition of consonants. *Journal of Speech and Hearing Research* 20.130-145, 1977.

Woodward, M.F, and Barber, C.G. Phoneme Perception in Lipreading. *Journal of Speech and Hearing Research* 3.212-222, 1960.