SIMILARITY STRUCTURE IN VISUAL PHONETIC PERCEPTION AND OPTICAL PHONETICS

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

[1] House Ear Institute, 2100 W. Third St., Los Angeles, CA 90057. [2] Department of Electrical Engineering, University of California at Los Angeles, CA 90095.

ABSTRACT

This study was undertaken to examine relationships between the similarity structures of optical phonetic measures and visual phonetic perception. For this study, four talkers who varied in visual intelligibility were recorded simultaneously with a 3-dimensional optical recording system and a video camera. Subjects perceptually identified the talkers' consonant-vowel nonsense syllable utterances in a forced-choice identification task. Then, perceptual confusion matrices were analyzed using multidimensional scaling, and Euclidean distances among stimulus phonemes were obtained. Physical Euclidean distances between phonemes were computed on the raw 3-dimensional optical recordings for the phonemes used in the perceptual testing. Multilinear regression was used to generate a transformation vector between physical and perceptual distances. Then, correlations were computed between transformed physical and perceptual distances. These correlations ranged between .77 and .81 (59% and 66% variance accounted for), depending on the vowel context. This study showed that the relatively raw representations of the physical stimuli were effective in accounting for visual speech perception, a result consistent with the hypothesis that perceptual representations and similarity structures for visual speech are modality-specific.

1. INTRODUCTION

A working definition for *speech perception* is that it is a process in which speech signals are transformed into the neural representations that are then projected onto word-form representations in the mental lexicon. *Phonetic perception* is more narrowly defined as the perceptual processing of the linguistically relevant attributes of the physical (measurable) speech signals. Understanding of phonetic perception requires determining the relationship between physical stimulus attributes and perceptual (or neural) consequences. However, very frequently, visual speech stimuli in perception experiments are described only in terms of the gender and language of the talker, how the recordings were made, and the linguistic content of the utterances (phonemes, words, sentences, etc.) [1], not any of the optical phonetic characteristics. The reasons for this might be that until recently speech researchers used primarily acoustic stimuli, and speech perception has been viewed as primarily an auditory function. Explanations for audiovisual and visual-only speech perception have appealed to various theoretical mechanisms such as a common amodal metric [2], a common articulatory representation [3], and abstract features [4] to explain the visual aspects of speech perception, apparently obviating characterization of optical phonetic signals. However, an alternative theory is that visual speech perception relies on modality-specific phonetic processing. If so, the relationship between optical speech signals and visual speech perception needs focused attention. One aspect of this relationship could be due to the perceptually primary processing of overall stimulus similarity [5]. This study investigated the relationship between visual perceptual and physical similarity.

Perceptual similarity. The most frequently noted characteristic of optical phonetic stimuli is that segmental dissimilarity is reduced relative to that obtained under good listening conditions with acoustic phonetic stimuli. Fairly systematic, although far from invariant, clusters of confusions among visual speech segments are regularly observed. For example, [m b p] are highly confused by perceivers. Such groupings of perceptually similar segments have come to be regarded as perceptual categories [e.g., 4], frequently referred to as *visemes*. Visemes have also come to be generally regarded as having no internal perceptual structure.

We have adopted the term *phoneme equivalence class* [PEC] as a generalization of the viseme concept, but one that covers a range of quantitatively defined similarity relationships

among phonemes [6]. In a previous experiment, we showed that subjects could perceive phonetic information within viseme-level PECs and also within PECs comprising yet higher levels of phoneme similarity (based on hierarchical clustering analysis of phoneme confusions) [7,8]. Thus, PECs (or visemes) do have internal perceptual structure related to phoneme categories.

Furthermore, previous results suggest that perceptual structure above the level of the PEC is important. This could be seen recently in a study by Auer [9,10] in which visual spoken word recognition was modeled using the Neighborhood Activation Model [11] and visual phoneme confusion data (phoneme probabilities for all possible phoneme pairs). The model was predictive of performance when the phoneme probabilities were obtained from lipreaders but not when confusion data were substituted from auditory speech-in-noise phoneme identification. This implied that segmental similarity is perceptual modality-specific and not based on an abstract or amodal similarity structure.

2. THE CURRENT STUDY

Perceptual versus physical similarity. Perceptual systems are sensitive to overall similarity [5]. However, few studies have investigated relationships between visual perceptual and physical similarity relationships for speech stimuli. Previously, Montgomery and Jackson [12] examined the relationship between visual vowel perception and physical stimulus characteristics in an experiment with four female talkers, ten viewers, and ten vowels in a format of /h/V/g/. They used a set of static descriptors during a single video frame of the "vowel maximum" to define physical features—lip height, lip width, lip opening area, acoustic duration, and visual duration, and they computed difference scores between measures for pairs of vowels. These measures were entered into multiple regression analyses to predict distances between vowels derived from the perceptual confusions. Multiple correlation coefficients (*R*) across talkers ranged between .49 and .82 (24 to 68% variance accounted for). The large range in multiple *R* values was interpreted as evidence that the measured features were somewhat inadequate, in particular, lacking information about the dynamic properties of the stimuli. However, the approach demonstrated the potential for understanding visual speech perception in terms of the similarity structure derived from measurable features of optical signals.

The current study investigated the relationship between perceptual and physical similarity structure for consonants in nonsense syllables. Perceptual similarity was estimated using multidimensional scaling (MDS) of phoneme identification confusion data. Physical stimulus similarity was measured using recordings from an optical recording system that tracked facial movements in three dimensions. The physical similarity was computed as the Euclidean distances among phonemes, based on the coordinates of the 3-D data. In making use of the raw 3-D data (as opposed to features such as measured lip-spread), we were investigating the hypotheses that perceptual similarity is based on integration across many different potential stimulus properties, and that perceptual representations preserve information about the visible, physical speech movements.

3. METHODS

Stimulus recordings. Talkers were videorecorded using a SONY UVW-1800 video recorder and a SONY DXC-D30 digital video camera. The talker's face filled the screen. Simultaneously, they were recorded using a Qualisys 3-dimensional motion capture system. For this, twenty retroreflector markers were pasted on the face of each talker. Only the 17 that were used in this analysis are shown in Figure 1. Two markers on the eyebrow (used for another study) and one on the nose ridge (reference point) were not used. Of the 17 markers, 6 were on the cheek, 8 were on the lips, and 3 were on the chin. The sampling frequency for the 3-dimensional data was 120 Hz.

Speech material. The speech material comprised two repetitions of 69 consonant-vowel (CV) syllables, where the vowel was one of /a, i, u/ and the consonant was one of the 23 American English consonants, /y, w, r, l, m, n, p, t, k, b, d, g, h, θ, ð, s, z, f, v, ʃ, ʒ, tʃ, dʒ/.

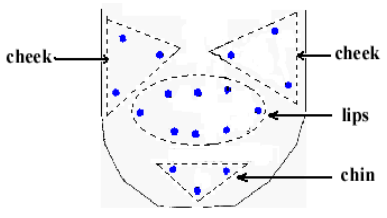


Figure 1. Placement of Qualisys markers.

Talkers. Four native American English talkers (two males and two females) were recorded. In a previous study, their visual intelligibility was judged and ranked against other talkers [13]. These four were selected to represent a range from relatively poor to quite good.

Perceivers. Adults with normal or corrected-to-normal vision were screened for English as a native language and good lipreading ability. The results reported below are for one male and one female with average or above average lipreading ability. Additional subjects are currently being tested.

Procedure for perceptual testing. The 3-dimensional movement recordings were used to quantify phonetic information potentially afforded by the optical signals and were not presented for visual perceptual judgments (i.e., as point-light stimuli). Instead, the simultaneously recorded video (with markers on the face and without sound) was presented for perceptual identification. Subjects were tested in a sound booth. A simulated keyboard with 23 consonants and corresponding sample words was displayed on the monitor. Viewers responded by selecting a consonant using the computer mouse. Stimuli were presented on a 19'' high-resolution SONY Trinitron color monitor placed next to the PC monitor at a distance of about 1 m from the subject. A SONY UVW-1800 videotape player was controlled by the same computer that was used to record the viewer's responses. The audio signal was turned off during the presentation.

For every subject, a practice set of 10 trials was given on Day 1. On each day, subjects were tested with four 138-item lists, one for each talker. Each list comprised two repetitions of the 69 CV tokens. There was one list for each of the four talkers. To counterbalance the effects of token order and talker order, two presentation tapes were made for each. On the first tape, the list items were randomized and the talker order was Male1, Female2, M2, F1. On the second tape, the list items were also randomized and the talker order was F2, M1, F1, M2. No feedback was given. Each list required approximately 16 minutes to finish, and there was a 5-minute break between lists. Testing occurred across 3 weeks.

Physical measurement analysis. The perceptual and physical measures were initially processed separately. The physical measures were used to compute Euclidean distances between every pair of consonants on a channel-by-channel basis, where channels were data streams for the three dimensions for each individual retroreflector marker. Only the initial part of the CV syllable was used for the physical analysis. The initial point for the optical data was based on the onset of the audio signal. A segment was defined to begin 30 ms prior to the onset of the acoustic signal (dashed line) and to extend for 280 ms (between the 2 solid lines).

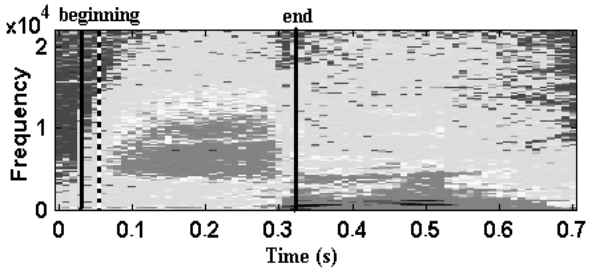


Figure 2. Consonant segment in /sa/ syllable.

The 3-dimensional optical data for each consonant were organized into matrices as follows:

$$O^{T_\alpha,CV,\beta}_{(1:51,1:34)} = \begin{bmatrix} o_{1,1} & \cdots & o_{1,34} \\ \vdots & \vdots & \vdots \\ o_{51,1} & \cdots & o_{51,34} \end{bmatrix}, \quad (1)$$

where α, CV, β stand for the talker number, CV syllable, and repetition number, respectively. For example, $O^{T_1,ba,1}_{(1:51,1:34)}$ represents data for the first repetition of syllable /ba/ for Talker 1. Each matrix has 34 columns, which represent 34 frames (=280 ms) and 51 rows, which represent the Qualisys channels (17 markers in a 3-D space). The physical Euclidean distance between a pair of consonants (C_1, C_2) was measured as follows:

$$PO^{C_1-C_2,V}_{(1:51,1)} = \sqrt{\sum_{i=1}^{4}\left(\sum_{j=1}^{2}\left(\sum_{k=1}^{34}\left(O_k^{T_i,C_1V,j} - O_k^{T_i,C_2V,j}\right)^2\right)\right)} \quad (2)$$

where k is the frame number, j is the repetition number, i is the talker number, and V is the vowel context. $PO^{C_1-C_2,V}_{(1:51,1)}$ has a dimension of 51 by 1. If all the Euclidean distances between the 23 consonants in a vowel V context were put together, a 51 by 253 matrix can be obtained as PO^V, where each row represents a different

optical channel. Three subsets can be derived from PO^V according to the marker location. They are PO^V_{lips} (for the lip markers), PO^V_{chk} (for cheeks), and PO^V_{chn} (for chin).

Perceptual identification analysis. Perceptual data consisted of two subjects' identifications of 23 consonants through lipreading each of the four talkers. For some analyses, results were pooled across the four talkers and resulted in three 23 x 23 confusion matrices (one for each vowel context), which were denoted as V-a, V-i, and V-u. There were 160 responses for each syllable in these confusion matrices. Also, an overall matrix, V-all, was obtained by pooling responses. Spatial representations of the perceptual similarity among consonants were obtained using MDS [14]. From the MDS solution, the Euclidean distances between all possible pairs of consonants in a three-dimensional space were calculated (i.e., 253 distances for 23 consonants). Prior to the MDS analysis, the confusion data were transformed using the phi-square statistic, which corrects for response biases and asymmetries in the data [15].

4. RESULTS

Perceptual results. The mean phonemes correct score was 37% (38% for C/a/, 36% for C/i/, and 36% for C/u/ syllables). The talker previously rated most intelligible with sentence stimuli was perceived most accurately in the current study (39% correct), and the talker rated least intelligible with sentences was perceived least accurately (35% correct). The middle two talkers each were perceived correctly on 37% of trials.

The 3-dimensional MDS representation of the confusion matrices (Fig. 3) agreed well with the results in [15] and represented a typical pattern of visual segmental similarities [16]. Fig. 3 demonstrates that clusters have internal structure (e.g., the members the group /s z ʧ ʃ/ do not have identical coordinates) as well as different distances to other clusters (e.g., /s z ʧ ʃ/ is closer to /t d/ than to /f v r/).

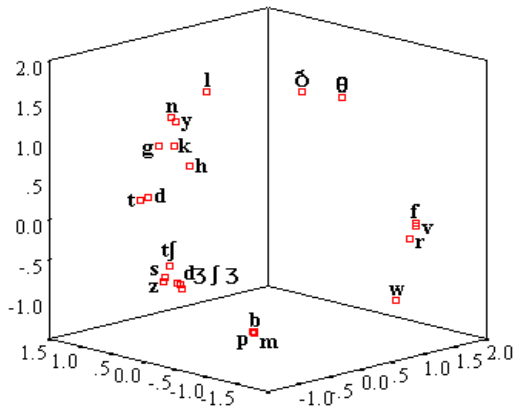


Figure 3. A 3-D MDS analysis of confusion data from the study.

Correlations. Multiple linear regression techniques were used in the evaluation of the relationship between perception and physical measures. The perceptual Euclidean distances were used along with the physical distances to generate a transformation vector. That vector was used to weight the physical distance vectors. Then the Pearson correlation was computed between the physical and perceptual distances. Those correlation coefficients are shown in Table 1. For example, in the vowel /a/ context, these measures are referred to as PO (51 x 253, 17 markers on the face), PO^a_{lip} (24 x 253, 8 markers on the lips), PO^a_{chk} (18 x 253, 6 markers on the cheek), and PO^a_{chn} (9 x 253, 3 markers on the chin).

	PO_{lip}	PO_{chk}	PO_{chn}	PO
V3_a	0.63	0.52	0.44	0.77
V3_I	0.67	0.55	0.61	0.81
V3_u	0.65	0.52	0.50	0.79

Table 1: Pearson correlation coefficients between visual perception and physical measures.

The two types of measures were related to each other using multilinear regression [17]. A transformation vector was computed to transform Euclidean distances from physical measures. In the final step of the study, the perceptual distances were correlated with the transformed physical distances. The last column in Table 1 shows the Pearson correlations using all three types of physical measures ($p < .001$). The table shows that the lips, chin, and cheeks are important for visual perception, and that using all the measures yields high correlations (around 0.8) for the 3-dimensional representations of visual

confusions. When the same procedures were applied on the data for individual talkers, the mean correlations for the two more intelligible talkers were higher (.71 and .72) than for the two less intelligible talkers (.62 and .66).

Figure 4. Scatterplot of Physical vs. Perceptual Distances

C/i/ Syllables

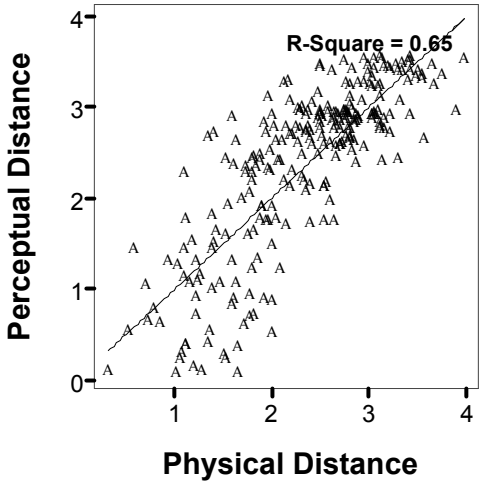


Figure 5. Scatterplot of Physical vs. Perceptual Distances

C/u/ Syllables

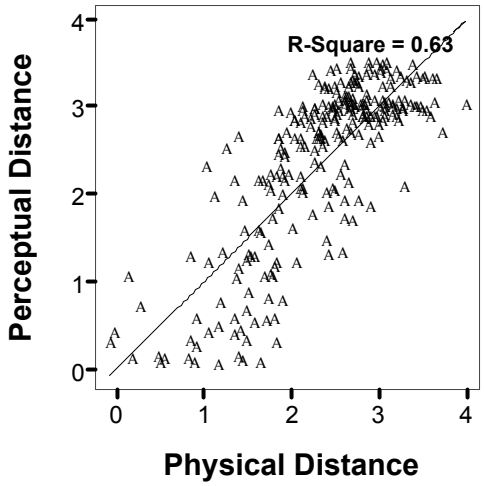
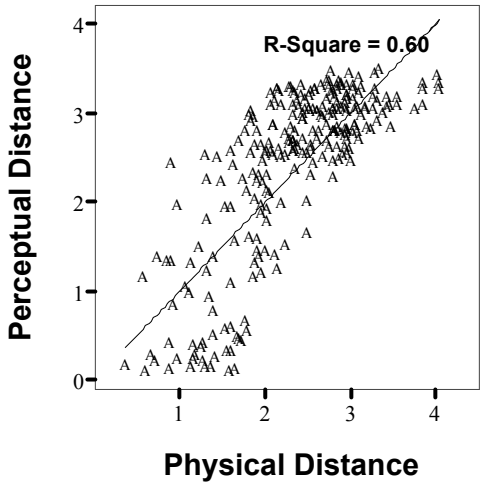


Figure 4 shows a scatterplot for the results for the C/i/ stimuli. Each transformed physical distance between a pair of consonants is plotted against the corresponding perceptual distance for that pair. The figure suggests that although there is a good correlation between the physical and perceptual measures, it is by no means perfect. At the smaller physical distances, the spread among perceptual distances is quite large. Figure 5, which shows the scatterplot for C/u/ stimuli has an opposite appearance at small perceptual values, for which there is a quite wide range of physical values. Figure 6, which shows the scatterplot for

Figure 6. Scatterplot of Physical vs. Perceptual Distances

C/a/ Syllables



C/a/ is similar to Figure 4 in the spread of perceptual distances that correspond with a narrow range of physical distances.

5. DISCUSSION AND CONCLUSIONS

Correlations between perceptual and physical distances using the chin, lips, and cheek markers ranged between .77 and .81 (respectively, between 59 and 66 percent of the variance accounted for). Thus, the physical measures incompletely accounted for perceptual similarity structure. However, several potential sources of visual information were not represented in the measures. For example, the motion of inner lip margins was not obtained, because the retroreflectors must be placed on the lip surface that is not occluded during speech. Perceivers can obtain useful information from inner versus outer lip movement [18]. Also, visible movements of the tongue were not measured. In addition, the physical measures were focused around the acoustic consonant

release. However, optical phonetic information is typically present earlier in the video signal and could have influenced perceptual judgments. Finally, the data in the scatterplots suggests that a non-linear relationship between physical and perceptual similarities might better account for the results. Nevertheless, the magnitudes of the obtained correlations were impressive, given the caveats already suggested.

The fact that the relatively raw measures of the physical stimuli were effective in accounting for visual speech perception is consistent with the hypothesis that visual speech perception is a function of modality-specific perceptual representations and similarity structures. If indeed visual speech stimuli are represented in terms of visual perceptual similarity and not converted to either an amodal or an auditory similarity structure, then audiovisual integration likely also takes place in terms of modality-specific representations.

6. REFERENCES

1. Munhall, K. G., and Vatikiotis-Bateson, E. "The moving face during speech communication," in R. Campbell, B. Dodd, and D. Burnham (Eds.), *Hearing by eye II: Advances in the psychology of speechreading and auditory-visual speech* (pp. 123-139). Psychology Press, East Sussex, UK, 1998.

2. Summerfield, Q. "Some preliminaries to a comprehensive account of audio-visual speech perception," in B. Dodd and R. Campbell (Eds.), *Hearing by eye: The psychology of lip-reading* (pp. 3-51). Psychology Press, East Sussex, UK, 1998.

3. Fowler, C. A. "An event approach to the study of speech perception from a direct-realist perspective," *J. Phonetics*, **14**: 3-28.

4. Massaro, D. W. *Perceiving talking faces*. MIT Press, Cambridge, MA, 1998.

5. Goldstone, R. L. and Barsalou, L. W. "Reuniting perception and conception," *Cogn.*, **65**: 213-262, 1998.

6. Auer, E. T., Jr., and Bernstein, L. E. "Speechreading and the structure of the lexicon: Computationally modeling the effects of reduced phonetic distinctiveness on lexical uniqueness," *J. Acoust. Soc. Am.*, **102**: 3704-3710, 1997.

7. Bernstein, L. E., Iverson, P., and Auer, E. T., Jr. "Elucidating the complex relationships between phonetic perception and word recognition in audiovisual speech perception," *Proceedings of the ESCA/ESCOP Workshop on Audio-Visual Speech Processing*, 1997. ISSN # 1018-4554.

8. Bernstein, L. E. (2001). "Visual speech perception," in E. Vatikiotis-Bateson (Ed.). *Handbook on audiovisual speech perception*, MIT Press, Submitted, 2001.

9. Auer, E. T., Jr., "The influence of the lexicon on speechread word recognition: Contrasting segmental and lexical distinctiveness," Submitted, 2001.

10. Auer, E. T., Jr., Bernstein, L. E., and Mattys, S, "The influence of the lexicon on visual spoken word recognition," AVSP2001. Schmeelsminde, Denmark.

11. Luce, P. A. and Pisoni, D. B. "Recognizing spoken words: The neighborhood activation model," *Ear & Hear.* **19**: 1-36, 1998.

12. Montgomery, A. A., & Jackson, P. L. "Physical characteristics of the lips underlying vowel lipreading performance," *J. Acoust. Soc. Am.* **73**: 2134-2144, 1983

13. Jiang, J., Alwan, A., Bernstein, L.E., Keating, P.A., and Auer, E.T., "On the correlation between facial movements, tongue movements and speech acoustics," *ICSLP 2000*, **Vol. 1**: 42-45, Beijing, P. R. China, 2000.

14. Young, F. W. and Hamer, R. M. *Multidimensional scaling*, Erlbaum, Hillsdale, NJ, 1987.

15. Iverson, P., Bernstein, L.E., and Auer, E.T., "Modeling the interaction of phonemic intelligibility and lexical structure in the audiovisual word recognition," *Sp. Com.* **26**: 45-63, 1988.

16. Owens, E. and Blazek, B., "Visemes observed by hearing-impaired and normal hearing adult viewers," *J. Sp. Hear. Res.* **28**: 381-393, 1985.

17. Kailrath, T., Sayed, A. H., and Hassibi, B. *Linear estimation*. Prentice Hall, 2000.

18. Rosenblum, L. D., Johnson, J. A., and Saldaña, H. M. "Point-light facial displays enhance comprehension of speech in noise," *J. Sp. Hear. Res.* **39**: 1159-1170, 1969.