Predicting Visual Consonant Perception from Physical Measures

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

[1]Department of Electrical Engineering, University of California at Los Angeles, USA
[2]Department of Communication Neuroscience, House Ear Institute, Los Angeles, USA
{jjt,alwan}@icsl.ucla.edu, {eauer,lbernstein}@hei.org

Abstract

The long term goal of our work is to predict visual confusion matrices from physical measurements. In this paper, four talkers were chosen to record 69 American-English Consonant-Vowel syllables with audio, video, and facial movements captured. During the recording, 20 markers were put on the face and an optical Qualisys system was used to track three-dimensional facial movements. The videotapes (with markers on the face and without sound) were presented to normal hearing viewers with average or above average lipreading ability, and visual confusion matrices were obtained. Results showed that the facial measurements were correlated with visual perception data by about 0.79 and account for about 63% of the variance.

1. Introduction

Lipreading is important for the hearing-impaired, yet, what makes a talker intelligible is not well quantified. Although several papers have examined the visual intelligibility of CVs, VCVs, and words [2, 3, 4], only a few have investigated the relationship between visual perception and physical (facial) measurements.

Montgomery and Jackson [1] examined the relationship between visual vowel perception and physical characteristics in an experiment with four female talkers, ten viewers, and fifteen vowels in a format of /hVg/. Since vowel pronunciation is relatively steady and of long duration, the authors used a set of static descriptors to define physical characteristics: lip height, lip width, lip aperture, acoustic duration, and visual duration. Their results indicated that the physical measures were moderately successful as predictors of vowel perception (approximately 50% of the variance was accounted for).

However, little is known about the relationship between consonant lipreading confusion and physical characteristics of the signal. One reason is that consonant pronunciation can be short and the fast transition from the consonant to the vowel involves significant information that might help visual perception. Therefore, it is necessary to capture dynamic characteristics of the face and perhaps the tongue. The goal of this study was to predict visual confusion matrices of consonants from physical measures such as lips, chin, and cheek movements.

2. Method

In this study, an optical Qualisys system was used to track facial movements with small infrared retroreflectors put on the face (Dataset2, see [5, 6]), and three Qualisys cameras were used to reconstruct the motion. The videotapes (with Qualisys markers on the face and without sound) were presented to viewers in the visual perception experiments.

2.1. Talkers and viewers

Four native American English talkers (two males, M1 and M2, and two females, F1 and F2) with different intelligibility ratings were recorded [5]. Normal-hearing people with normal or corrected vision were screened for English as a native language and lipreading ability, and two viewers (one male and one female) with average or above average lipreading ability participated in this study as viewers.

2.2. Material

The speech material consisted of two repetitions of 69 CV syllables where the vowel is one of /a, i, u/ and the consonant is 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/. The data were recorded acoustically and with an optical and Qualisys data stream. The videotapes were presented to each viewer 10 times. Therefore, for each CV syllable, there were 160 responses (2 repetitions, 10 trials, 4 talkers, and 2 viewers).

2.3. Recording channels

There were 20 optical markers put on the face of which only 17 were used in this analysis. The 2 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 placement of these markers is shown in Figure 1. All the data streams were aligned. Please refer to [5, 6] for details.

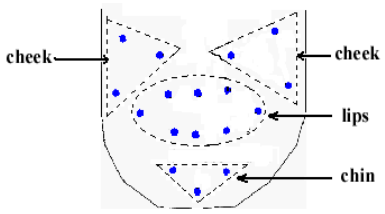

Figure 1. Placement of Qualisys markers.

2.4. Procedure

Each viewer was seated in a sound booth. A Sony BETACAM videotape player was controlled by a personal computer that was placed on the table. The computer was also used to record the viewer's responses. A simulated keyboard with 23 consonants and corresponding sample words were displayed on the monitor. Viewers responded by selecting a consonant using the computer mouse. Stimuli

were presented on a 19'' high-resolution color monitor placed next to the PC monitor at a distance of about 1 m from the viewer. The audio signal was turned off during the presentation. For every viewer, a practice set of 10 trials was given on day 1.

On each day viewers were tested with four 138-items lists, one for each talker. Each list consisted of two repetitions of the 69 CV tokens. There were four talkers recorded so that there were 4 lists. To counterbalance the effects of token order and talker order, two presentation tapes were made. On the first tape, the list items were randomized and the talker order was M1-F2-M2-F1. On the second tape, the list items were also randomized and the talker order was F2-M1-F1-M2. Half of the viewers began with tape 1 and the other half began with tape 2, and then they viewed the other tape. Each list took about 16 minutes to finish and there was a 5-minute break between lists. For each viewer, the experiments lasted for 2-3 weeks and no feedback was given to the viewers.

2.5. Analysis

2.5.1. Physical measures

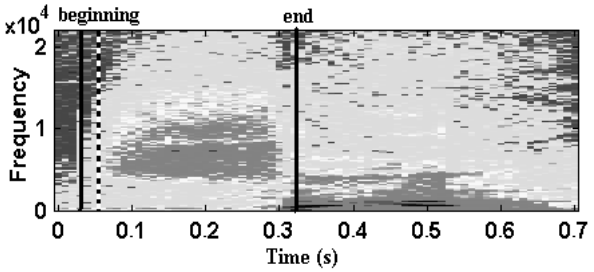

Figure 2. Consonant segment in a /Sa/ syllable.

The sampling frequency for the Qualisys data was 120 Hz. In this study, since we were interested in analyzing the consonants, only the initial part of the CV was used. Figure 2 illustrates how this was done for /Sa/. Endpoint detection was based on the audio signal. A segment was defined 30 ms prior to the beginning (dashed line) of the CV and lasted for 280 ms (between the 2 solid lines). The Qualisys data for each segment were then organized into matrices as follows:

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

where α, CV, β stand for the talker number, CV syllable, and repetition number, respectively. For example, $O_{(1:51,1:34)}^{T_1,ba,1}$ 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) were measured as follows:

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

where k is the frame number, j is the repetition number, i is the talker number, and V is the vowel context. $PO_{(1:51,1)}^{C_1-C_2,V}$ 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 Qualisys channel. Three subsets can be derived from PO^V according to the marker location. They are PO_{lips}^V (for the lip markers), PO_{chk}^V (for cheeks), and PO_{chn}^V (for chin).

2.5.2. Visual perception confusion matrices

Perceptual data consisted of two viewers' identifications of 23 consonants through lipreading each of the four talkers. Results were pooled across the four talkers and resulted in three 23x23 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 adding all responses.

To generate phonemic equivalence classes, it was necessary to transform confusion data into similarity estimates. This was done by applying the phi-square statistic to these confusion matrices. The phi-square measure is

$$\text{dist\_PHISQ}(x,y) = \sqrt{\frac{\sum\frac{(x_i-E(x_i))^2}{E(x_i)} + \sum\frac{(y_i-E(y_i))^2}{E(y_i)}}{N}} \quad (3)$$

where x_i and y_i are the frequencies that phonemes x and y were identified as response category i. The phi-square coefficient for an individual consonant pair (C_1 and C_2) is then independent of other consonants in the experiment. Further, the phi-square coefficient has an advantage when there are response biases and asymmetries [4].

2.5.3. Multidimensional Scaling (MDS) and Hierarchical clustering analysis (HCA)

MDS was used to yield spatial representation of the consonants from which the Euclidean distances between all possible pairs of consonants in a three-dimensional space were calculated (the number of distances is 253 for the 23 consonants). Euclidean distances between consonants in the MDS space represent a mathematically tractable and stable transformation of the confusion into distances. For the zeroth-entry cell in the confusion matrices, this was more important. These distances were used as an additional perceptually based metric of consonant dissimilarity. Given these visual distances and corresponding physical distances, the predictability of visual perception from physical measures can be assessed.

From the visual perception confusion matrices, the HCA method was used to build partitions of phonemes, where recursive clusters are formed to grow a binary tree representing an approximation of similarities between phonemes. In this study, HCA was also used to determine the dimensionality of MDS which was 3.

SPSS was used for MDS and HCA where phi-square was applied, the measure was Euclidean distance, and the clustering method was the average linkage between groups.

Table 1: Overall confusion matrix V_all which was pooled across 4 talkers, 2 viewers, 10 trials, 2 repetitions, and 3 vowels

		RESPONSE																							
		y	**w**	**r**	**l**	**m**	**n**	**p**	**t**	**k**	**b**	**d**	**g**	**h**	**θ**	**ð**	**s**	**z**	**f**	**v**	**ʃ**	**ʒ**	**tʃ**	**dʒ**
	y	100	1		29		39		24	120		35	52	52		1	2	2	13	2			5	3
	w	4	416	50		6		1			3													
	r	2	89	196	1	4	3		4	2	1	1			1		1	2	126	42			1	4
	l	6			327	1	22		41	22	1	26	14	2	4	3	2	5					3	1
	m			1		191		126			161	1												
	n	18			203		40		59	41		41	28	23	1	1	6	3				1	11	4
	p			2		130		204			144													
S	**t**	6			7		20		174	38		96	13	9	3	10	46	30			1	1	14	12
T	**k**	16			6	1	6		19	210		12	66	143					1					
I	**b**			1		114		211			154													
M	**d**	10			49		30		161	12		98	10	2	5		39	32	1		1	3	15	12
U	**g**	36		1	22		10		30	205		19	69	65			7	9			2		1	4
L	**h**	5	2		3	3	3	2	3	68	1	7	21	359			2		1					
I	**θ**	1			29		17		1	3		1	6		239	183								
	ð	3			22		35		26	6		21	8	3	176	179						1		
	s						5		72	1		36	2			4	204	115		1	5	3	12	20
	z						1	3	74	3	1	38	3		1		198	109			9	3	18	19
	f				1		1		1										358	119				
	v		1	4			1		1										354	119				
	ʃ	6					7		19	4		15	19				59	20			74	31	131	95
	ʒ	1			2	1	6		16	1		9	42				45	22			79	21	133	102
	tʃ	7			1		5		47	10	1	23	41	2	1		30	28			58	15	140	71
	dʒ	1					2		25	5		11	44	1			46	37			65	13	143	87

3. Results

3.1. Overall visual perception results

The average recognition accuracy was 36.9% (38.4% for /Ca/, 36.1% for /Ci/, and 36.0% for /Cu/ syllables). The results are somewhat similar to these reported in [2] (40% for /aCa/, 33% for /iCi/, and 24% for /uCu/); but lower than the result reported in [4] (48% for /CVs/). For /Ca/ syllables, the intelligibility of consonants, from most to least, were{w h f l k tʃ s ð r p b t θ m g v d z dʒ y n ʃ ʒ}. For /Ci/ syllables, the intelligibility of consonants, from most to least, were {w f h l r ð s p k θ t v b m z ʃ dʒ d tʃ g n ʒ y}. For /Cu/ syllables, the intelligibility of consonants, from most to least, were {w h f l θ m y t p s k b r d z v tʃ dʒ ð ʃ g n ʒ}. The overall confusion matrix V_all which was pooled across the 4 talkers, 2 viewers, 10 trials, 2 repetitions, and 3 vowels is shown in Table 1.

3.2. Dendrograms and corresponding MDS analysis

Dendrograms were obtained via the HCA method from the confusion matrices. The dendrogram shown in Figure 3 corresponds to the confusion matrix V_all in Table 1. From Figure 3, the following phonemic equivalence classes were obtained: {m, b, p}, {f, v}, {r}, {w}, {θ, ð}, {ʃ, dʒ, ʒ, tʃ}, {t, d, s, z}, {l, n}, {k, g, y h}. These are in agreement with results reported in [4] except that /d/ was in a group with {ʃ, dʒ, ʒ, tʃ}. The dendrograms from visual confusion matrices V_a, V_i, and V_u were in general similar to that from the confusion matrix V_all. The difference was that in V_i and V_u dendrograms, {s, z} and {t, d} were separate classes.

In [1], the authors used a 2-D map to represent visual perception for vowels. We found that a 2-D MDS was not sufficient to represent consonant confusion (a dendrogram could not be generated properly and only 60% of the variance was accounted for). Instead, a 3-D MDS (Figure 4) was applied and 75% variance was accounted for. A 6-D MDS was also used and 96% of the variance was accounted for.

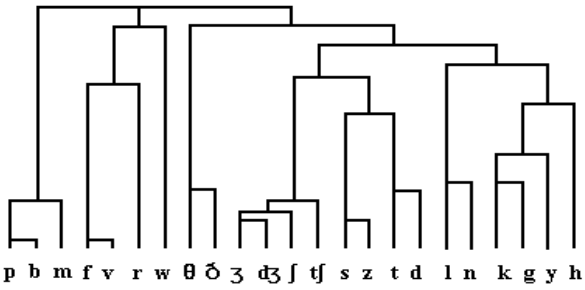

Figure 3. Dendrogram generated from the overall confusion matrix in Table 1.

The 3-D MDS representation of the confusion matrices agrees well with the results in [4]. From Figure 4, the 253 visual distances for confusion matrix V_all were calculated and denoted as V3_all (1x253). In addition, V3_a (1x253), V3_i (1x253), and V3_u (1x253) were computed for confusion matrices V_a, V_i, and Vu, respectively. The distances from the 6-D MDS were used to represent the raw perceptual confusions. They are referred to as V6_a (1x253), V6_i (1x253), and V6_u (1x253). The correlation between V6_a and V6_i is 0.91, the correlation between V6_a and V6_u is 0.83, and the correlation between V6_i and V6_u is 0.85. This implies the visual consonant confusions were relatively similar for /Ca/ and /Ci/ syllables, but different for /Cu/ syllables.

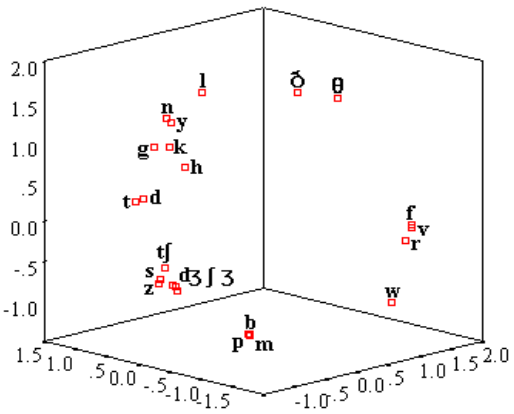

Figure 4. A 3-D MDS analysis of confusion matrix in Table 1.

From the 3-D representation of consonant confusions, one could examine its relationship to place of articulation, manner of articulation, and voicing features. The correlation between the 3-D representation and place of articulation was 0.97 for /Ca/, 0.92 for /Ci/, and 0.94 for /Cu/. This suggests there was one direction in this 3-D space representing place of articulation. For manner of articulation, the correlation was 0.55 for /Ca/, 0.63 for /Ci/, and 0.61 for /Cu/. But for voicing, the correlation was 0.09 for /Ca/, 0.08 for /Ci/, and 0.13 for /Cu/. These correlations emphasize the fact that visual perception is highly related to place, modestly to manner, and unrelated to voicing.

3.3. Predicting visual perception from physical measures

Multiple linear regression techniques were used to assess the relationship between visual consonant perception and physical measures so that the factors contributing to visual intelligibility could be examined. The physical measures employed here included geometry, timing, duration, and dynamic information to some degree. The physical distances between consonants were discussed in Section 2.5.1. For example, in the vowel /a/ context, these measures are referred to as PO^{a} (51x253, 17 markers on the face), PO^{a}_{lip} (24x253, 8 markers on the lips), PO^{a}_{chk} (18x253, 6 markers on the cheek), and PO^{a}_{chn} (9x253, 3 markers on the chin). In Section 3.2, visual distances were obtained (V3_a, V3_i, and V3_u). V6_a, V6_i, and V6_u approximate the raw confusion matrices which were used to examine how good the 3-D MDS representation is.

Table 2: Correlation coefficient between visual perception and physical measures

	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
V6_a	0.58	0.49	0.38	0.70
V6_i	0.60	0.53	0.51	0.73
V6_u	0.66	0.56	0.48	0.76

In Table 2, we show the correlation of the visual distances from 3-D and 6-D MDS to the physical measures from either the markers on the lips, cheeks, or chin. The last column shows correlations with all three physical measures. The table shows that the lips and cheeks are important for visual perception and that using all the measures yields high correlation (around 0.8) for the 3D representations of visual confusions and account for 63% of the variance. Table 2 shows that for /Ca/ and /Ci/ syllables, the 3-D dimensional representation of visual perception was better than the raw confusions (6-D MDS) in the sense that their correlations with physical measures were higher. This confirms the results reported in [1]. But for /Cu/ syllables, raw confusions were slightly better correlated with physical measures.

4. Summary and Conclusions

In this paper, we examined the relationship between visual confusion matrices and physical measures of the signal. We showed a high correlation (around 0.8) between facial measures and visual confusions. The 3-D MDS was sufficient to represent visual confusions. In these 3-D MDS representations, place of articulation was an important dimension. Of the facial movements, the lips and cheeks are more important for visual perception, than the chin. The dynamic characteristics of the facial movements were included in physical distances to some degree. But the averaging effect in some cases may have not been desirable. In future work, differences due to talkers and the effect of tongue movements on visual perception of speech articulations will be examined.

5. Acknowledgements

This research was supported in part by an NSF KDI award 9996088. We wish to acknowledge the help of Brian Chaney, Jennifer Yarbrough, Sven Mattys, and Taehong Cho in data collection and conducting the perception experiments.

6. References

[1] Montgomery, A. A. and Jackson, P. L., "Physical characteristics of the lips underlying vowel lipreading performance", *JASA*, Vol. 76, 1983, pp. 2134-2144.

[2] Owens, E. and Blazek, B., "Visemes observed by hearing-impaired and normal hearing adult viewers", *J. Speech & Hearing Research*, Vol. 28, 1985, pp. 381-393

[3] Bernstein, L.E., Demorest, M.E., and Tucker, P.E., "Speech perception without hearing", *Perception & Psychophysics*, 62(2), 2000, 233-252

[4] Iverson, P., Bernstein, L.E., and Auer, E.T., "Modeling the interaction of phonemic intelligibility and lexical structure in the audiovisual word recognition", *Speech Communication*, Vol. 26, 1988, pp. 45-63

[5] 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*, Beijing, P.R.China, Vol. 1, pp. 42-45.

[6] Bernstein, L.E., Auer, E.T., Chaney, B., Alwan, A., and Keating, P.A., "Development of a facility for simultaneous recordings of acoustic, optical (3-D motion and video), and physiological speech data", *JASA*, 107(5), 2000, p2887.