Consonant Confusion Structure Based on Machine Classification of Visual Features in Continuous Speech

Jianxia Xue[1], Jintao Jiang[2], Abeer Alwan[1], and Lynne E. Bernstein[2,3]
[1]Department of Electrical Engineering
University of California, Los Angeles CA 90095, USA, {jxue, alwan}@ee.ucla.edu
[2]Department of Communication Neuroscience
House Ear institute, CA 90057, USA {jjt, lbernstein}@hei.org
[3]National Science Foundation
Social, Behavioral, and Economics Directorate, Arlington, VA, 22230

ABSTRACT

This study is a first step in selecting an appropriate subword unit representation to synthesize highly intelligible 3D talking faces. Consonant confusions were obtained with optic features from a 320-sentence database, spoken by a male talker, using Gaussian mixture models and maximum a posteriori classification methods. The results were compared to consonant confusions obtained from visual-only human perception tests of non-sense CV syllables spoken by the same talker. At the phoneme level, machine classification results for the continuous speech database had worse performance than human perception with isolated syllables. However, the number of distinguishable consonant clusters by machine is equal to that by humans. To model the optic feature for continuous visual speech synthesis, the results suggest that for most consonants, modeling optic feature in phoneme level is more appropriate than modeling in phoneme clusters determined from visual-only human perception tests. For some consonants, modeling in a context-dependent manner might be helpful in improving the modeling accuracy for the talker studied in this paper.

1. INTRODUCTION

The long term goal of this study is to synthesize highly intelligible 3D talking faces using an acoustically-driven visual speech synthesis system as shown in Figure 1. The hybrid system is a combination of concatenation and model-based methods. The system requires two codebooks: one for acoustic-to-optic feature mapping, and another for concatenation. The two codebooks are obtained using statistical methods such as Gaussian mixture modeling (GMM). The present study addresses two issues that are important in designing the second codebook: What is the phoneme confusion structure in continuous speech as measured by machine learning, and how does the confusion structure differ from phoneme confusions obtained via human lipreading experiments?

In previous studies, visual features have been modeled in different ways with various units such as subphonemes [1][2][3], visemes [4][5], divisemes[6], triphones [7], and utterances [8]. In [1], optic features were first transformed into an eigen-face eigen-opticflow defined space, and then modeled by Gaussians with adjustments learned from the estimation error in the training data. In [2][4], optic features were estimated using neural networks given the acoustic HMM states, while in [3] GMMs were used in modeling the optic features. In [8], dynamical system identification techniques were used to estimate the optic features. In [6][7], direct recordings of speech were used with concatenation techniques for visual synthesis. Compared to previous systems, our design is focused on using context-dependent subword [9] units to gain system efficiency, using an acoustically-driven approach to improve audio-visual synchronization and coarticulation, and using a wire-frame 3D talking face model. The type of units used in the system has an impact on database requirements and on synthesis intelligibility. In order to improve synthesis intelligibility (which requires the units to be long, i.e., triphones) while keeping the size of the database reasonable (which requires the unit to be short, i.e., phonemes), we need to find a balanced solution. This study examines the optic features modeled by GMMs at the phoneme level in continuous speech. Consonant confusion results are compared with that obtained by human lipreading experiments (human forced-choice identification with nonsense CV syllable stimuli). This is a first step in selecting an appropriate subword unit representation for the system.

In the visual speech perception literature, phoneme identification scores for nonsense syllables have been reported to be below 50% correct [10]. However, perceptual confusion

matrices are structured and sparse matrices, resulting in the groupings of some phonemes. Some of these groupings map to the visually similar phoneme clusters, known as "visemes" [11]. In [12], 12 visemes {/f v/, /T D/, /w/, /r/, /p b m/, /S Z C J/, /s z/, /t d n/, /l/, /h/, /k g/ and /y/} (hereafter referred to as Kricos-viseme) were used, and they were mainly based on the consonants' place and manner of articulation, but not derived from perception tests. They can be considered theoretically optimum in terms of the maximum number of visually distinguishable perceptual units for consonants. When visemes are defined via cluster analysis by within-cluster response rates of 75% or greater from human forced-choice identification from nonsense syllable stimuli, consonants have been found to fall into 6 viseme clusters {/f v/, /T D/, /w r/, /p b m/, /S Z C J/, /t d s z j k n g l/} in [13], and for example, {/f v, r/, /T D t d l n/, /w /, /p b m/, /k g y h/, /s z S Z C J/}(hereafter referred to as Jiang-viseme) in [14]. Consonant clustering results differed by talker [10][14] and by vowel context in CV syllables [14]. One of the talkers studied in [14] was chosen for this study.

In this paper, classifications using visual speech information for phoneme and phoneme-cluster units were obtained from a 320-sentence database. Visual features are described by Gaussian mixture models of phoneme units. Confusions were obtained by machine classification based on the maximum a posteriori criterion. Results were compared with those from visual speech perception tests using various measurements, including average within-unit accuracy scores and unit identification entropy.

2. METHODS

2.1 Database and preprocessing

The database is an audio-visual recording of 320 sentences chosen from the IEEE/Harvard sentence corpus. The sentences were spoken by an American English talker with high visual intelligibility. Details on the recording can be found in [15].

Necessary data pre-processing were carried out for both the acoustic and optic data. For the acoustic data, phoneme transcriptions were obtained manually. Then, phoneme segmentation using HTK forced-alignment tools [16] were obtained. The accuracy of the forced-alignment tool was checked by manual segmentation on a database that has the same 320 sentences spoken by another talker. If a relative segmentation discrepancy is defined by the relative phoneme middle point difference between the machine and manual segmentation results (normalized by the manual segmentation duration), then the forced-alignment tool had an average discrepancy of 0.09. If a segmentation error is defined by having segmentation discrepancy equal to or larger than 0.5, then the forced-alignment tool had an error rate of 4.6%. The discrepancy and error rate are acceptable for this study.

The optic data were from 3D recordings of 17 markers placed on the talker's cheeks, lips, and chin [15]. Head movements were compensated for in the optic data using three anchor points on the upper face. Low-pass filtering was applied to limit the optic data within 40 Hz (to remove recording noise).

2.2 Visual feature extraction

The recorded facial movements include the lip, cheek and chin areas which, in a previous study, accounted for 55%, 36%, and 32%, respectively, of the variance in visual perception results [14]. This study explores static optic features extracted at the mid-phoneme position. Global PCA features from the 3D marker positions and a set of marker distance features were compared in the analysis. The PCA features performed poorly in classification and are not reported here. The marker distance features generated meaningful confusion patterns. Nine features extracted from 10 markers were selected for the analysis and are listed in Table 1 with the marker locations shown in Figure 2. From the correlation analysis between all the combinations of two feature channels, features 1, 8 and 9 are highly correlated (Pearson correlation scores greater than 0.99). Correlations from other feature pairs range from 0.01 to 0.86.

2.3 Modeling

For each phoneme, visual features were characterized using a multivariate Gaussian mixture model described by the following:

$$p(V|phn_i) = \sum_{m=1}^{M_i} N(\mu_{im}, \Sigma_{im}), \qquad (1)$$

where V represents the visual feature vector, phn_i represents phoneme i, M_i is the number of mixtures for phoneme i, μ_{im} represents the mean vector of multivariate Gaussian mixture m for

phoneme i, and Σ_{im} is the corresponding covariance matrix for the Gaussian. The expectation maximization (EM) algorithm [17] is used for parameter estimation. Each phoneme is represented by a different number of mixtures to describe its statistical behavior. Mixtures can be interpreted as exclusive modes automatically learned from the sample space for a phoneme. The number of mixtures provides phoneme variation information in terms of the amount of potential modes by machine learning.

Phoneme classification is obtained using the maximum a posteriori probability criteria. Given a feature vector V_{ki} that represents the k^{th} sample from phoneme phn_i, the identification decision is made by

$$\arg\max_j p(V_{ki}, phn_j), \qquad (2)$$

where

$$p(V_{ki}, phn_j) = prior(phn_j) \cdot p(V_{ki} | phn_j) \quad .(3)$$

The prior information of a phoneme, $prior(phn_j)$, is the frequency of occurrence in the database. Given the classification results for all the samples, a confusion matrix C is obtained with the element c_{ij} representing the frequency that phoneme i is classified as phoneme j.

2.4 Evaluation

To better observe the confusion structure and quantify the identification variation of each unit (e.g. phoneme, or phoneme clusters), both within-unit identification correct scores and the unit identification entropy values were computed from the confusion matrix. Correct scores reflect the diagonal elements of the confusion matrix. The entropy also takes into account the off-diagonal structures. The identification entropy value for unit p_i is

$$e_i = -\sum_{j=1}^{P_i} c_{ij} \log_2 c_{ij} \quad . \qquad (4)$$

where P_i is the total number of units that phoneme phn_i are classified into. Higher entropy indicates a more evenly distributed misclassification. The measurements are calculated using three types of representation units: Phoneme, Kricos-viseme (theoretic viseme unit), and Jiang-viseme (experimental viseme unit). To our knowledge, no previous studies had calibrated optical speech classification using perception results for different optical-phonetic representation units.

Results from human visual-only perception of consonants in nonsense CV syllables spoken by the same talker with normal hearing subjects [14] were used for comparison.

3. RESULTS

Tables 2 and 3 show the comparison between machine classification and human identification in terms of average percent correct score and entropy, respectively. Machine classifications were either done with the training and test sets being the same (A) or different (B). The performance degraded by 24% from machine A to B in terms of average correct score. Consonant confusion matrices and the corresponding dendrograms obtained from machine-B classification and human lipreading are shown in Figure 3. Figure 4 shows a comparison between machine classification and human identification using the Kricos-visemes in correct score and entropy. Figure 5 shows a similar comparison using the Jiang-visemes. Due to limited data, the phonemes /Z, J/ [18] were not included in the machine classification experiment. Consonant /G/ was not tested in the human forced-choice identification test.

Using the 75% within-cluster response rates threshold, machine classification generated 4 consonant clusters {/p, b, m/, /f, v/, /s, z, T, S, C, y/, and /t, d, D, k, g, w, r, l, n, G, h/}, and human perception for this talker generated a different set of 4 consonant clusters {/p, b, m/, /f, v, r/, /s, z, S, C, Z, J, t, d, T, D, l, n, k, g, y, h/, and /w/ [14]. Note that the cluster /p, b, m/ is the only common cluster obtained by machine and by humans. This is also the only common cluster that has been reported for human lipreading [11] [13] [14].

In the Kricos-viseme representation, the units /sz/ and /k, g, G/ had similar or better Machine-B classification results than human in both correct score and entropy. The unit /r/ and /t, d, n/ had better or equal correct score than human, but the opposite was observed in entropy results. The units /p, b, m/, /f, v/, /T, D/, /S, Z, C, J/ and /y/ had moderate degradation compared to human identification. The units /l/, /h/, and /w/ had significantly worse confusion structure by machine classification in both diagonal and off-diagonal positions. The performance degradation is more significant in terms of identification entropy in general. In the Jiang-viseme

representation, machine classification resulted in worse performance than human perception for all the units except /S, Z, C, J, s, z/.

4. DISCUSSION AND CONCLUSION

In this study, consonant confusions were obtained with optic features from a 320-sentence database, spoken by a male talker, using Gaussian mixture models and maximum a posteriori classification methods. The results were compared to consonant confusions obtained from visual-only human perception tests of non-sense CV syllables spoken by the same talker. Machine classification results for the continuous speech database had worse performance than human perception with isolated syllables. However, the number of distinguishable consonant clusters by machine is the same as that by humans.

For a few phonemes, visual information in the recorded continuous visual speech can better characterize some phoneme/phoneme-clusters from the physical data than what humans can perceive from isolated speech. In addition, the exact clustering by humans is talker and context dependent [14]. For these two reasons, for the purpose of continuous visual speech synthesis, modeling finer units such as phonemes should yield better results than modeling clusters (or visemes) as defined by human perception. For those consonants that have more machine-based confusions than those by humans, the models might not fully capture the variations that exist in continuous speech. Adding prior knowledge of context variations into the optic feature modeling might be helpful in improving the accuracy of the optic feature codebook. In addition, improvement of visual feature extraction and processing may be warranted.

Results suggest that to design an efficient visual feature codebook for an intelligible visual speech synthesizer, for the talker in this study, the consonants /s, z, t, d, n/ may be represented in a context-independent manner, while the remaining consonants need to be modeled with context-dependency. Future work will examine the generalization of these conclusions to different talkers in detail.

5. ACKNOWLEDGEMENT

This research was supported in part by NSF grants KDI 9996088 and ITR 0312434. We would like to thank B. Chaney for database collection and processing. The views expressed here are those of the authors and do not necessarily represent those of the National Science Foundation.

6. REFERENCES

[1]. T. Ezzat, G. Geiger, and T. Poggio, "Trainable video realistic speech animation," *Proc. ACM SIGGRAPH*, 2002, pp. 288-298.

[2]. E. Yamamoto, S. Nakanura, and K. Shikano, "Lip movement synthesis from speech based on hidden Markov models," *Speech Comm.*, 26, 1998, pp. 105-115.

[3]. D. Massaro, J. Beskow, M. Cohen, C. L. Fry, and T. Rodriguez, "Picture my voice: audio to visual speech synthesis using artificial neural network," *Proc. AVSP'99*, Santa Cruz, CA, 1999.

[4]. P. Hong, Z. Wen, and T. Huang, "Real-time speech-driven face animation with expression using neural networks," *IEEE Trans. on Neural Networks*, vol. 13, no. 1, Jan, 2002, pp. 100-111.

[5]. C. Benoit, T. Lallouache, T. Mohamadi, and C. Abry, "A set of French visemes for visual speech synthesis," Talking machines: Theories, models and designs, G. Bailly and C. Benoit Eds., Amsterdam, The Netherland: Elsevier Science BV, 1992, pp. 485-504.

[6]. J. Ma, R. A. Cole, B. Pellom, W. Ward, and B. Wise, "Accurate automatic visible speech synthesis of arbitrary 3D models based on concatenation of Di-viseme motion capture data," Journal of Computer Animation and Virtual Worlds, Vol. 15, 2004, pp. 485-500.

[7]. C. Bregler, M. Covell, and M. Slaney, "Video rewrite: Driving visual speech with audio," *Proc. ACM SIGGRAPH*, 1997.

[8]. P. Saisan, A. Bissacco, and S. Soatto, "Synthesis of facial motion driven by speech," ECCV, Prague, May, 2004.

[9]. L. Rabiner and B. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, Englewood Cliffs, NJ, 1993.

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

[11]. C. G. Fisher, "Confusions among visually perceived consonants," *Journal of Speech & Hearing*, 11, 1968, pp. 796-804.

[12]. P. Kricos and S. Lesner, "Differences in visual intelligibility across talkers," *The Volta Review*, 84, 1982, pp. 219-225.

[13]. B. E. Walden, R. A. Prosek, A. A. Montgomery, C. K. Scherr, and C. J. Jones, "Effects of training on the visual recognition of consonants," *Journal of Speech & Hearing Research*, 20, 1977, 130-145.

[14]. J. Jiang, "Relating optical speech to speech acoustics and visual speech perception," *Dissertation*, UCLA, 2003.

[15]. J. Jiang, A. Alwan, P. A. Keating, E. T. Auer, and L. E. Bernstein, "On the relationship between face

movements, tongue movements, and speech acoustics," *EURASIP J. Applied Signal Processing*, vol. 11, 2002, pp. 1174-1188.

[16]. S.Y. et al., *The HTK Book (version 3.1)*, Cambridge University, Engineering Department, 2001.

[17]. J. A. Bilmes, "A gentle tutorial on EM algorithms and its applications to parameter estimation for Gaussian mixture and hidden Markov models," *ICSI-TR-97-021*, 1997.

[18]. P. F. Seitz, L. E. Bernstein, and E. T. Auer. PhLex (Phonologically Transformable Lexicon), A 35,000-word pronouncing American English lexicon on structural principles, with accompanying phonological rules and word frequencies (Gallaudet Research Institute, Washington, DC), 1995.

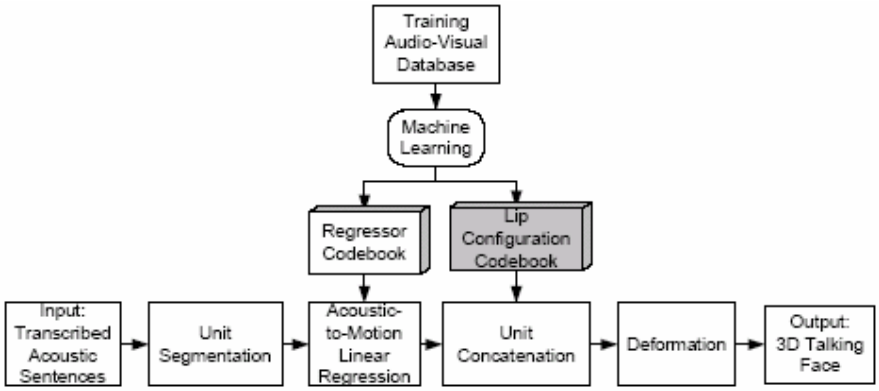

Figure 1. Framework of the visual speech synthesis system.

Ch	Feature name	Physical measurement
1	Mouth open height	ULC-LLC_xyz
2	Mouth open width	MLC-MRC_xyz
3	Lip protrusion parameter 1	LLC_z
4	Lip protrusion parameter 2	ULC-MLF_z
5	Lip protrusion parameter 3	LLC-ULC_z
6	Lip protrusion parameter 4	LLC-MLF_z
7	Cheek feature	CheekL_MLF_xyz
8	Lip rounding	Mouth open height / Mouth open width
9	Mouth open area	Area of the polygon defined by the 8 lip markers

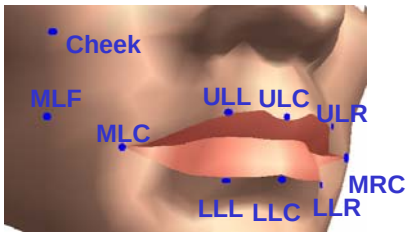

Figure 2. Locations of the markers used in the physical measurements.

Table 1. Selected physical optical features.

	Consonant	Kricos-viseme	Jiang-viseme
Machine-A	55%	64%	80%
Machine-B	26%	40%	63%
Human[14]	33%	61%	90%

Table 2. Average identification correct scores using 3 units. Machine-A refers to machine classification when the training and test sets are the same. Machine-B refers to machine classification when the training and test sets are different.

	Consonant	Kricos-viseme	Jiang-viseme
Machine-A	2.12	1.60	0.99
Machine-B	2.60	1.99	1.39
Human[14]	2.00	1.24	0.52

Table 3. Average identification entropy using 3 units.

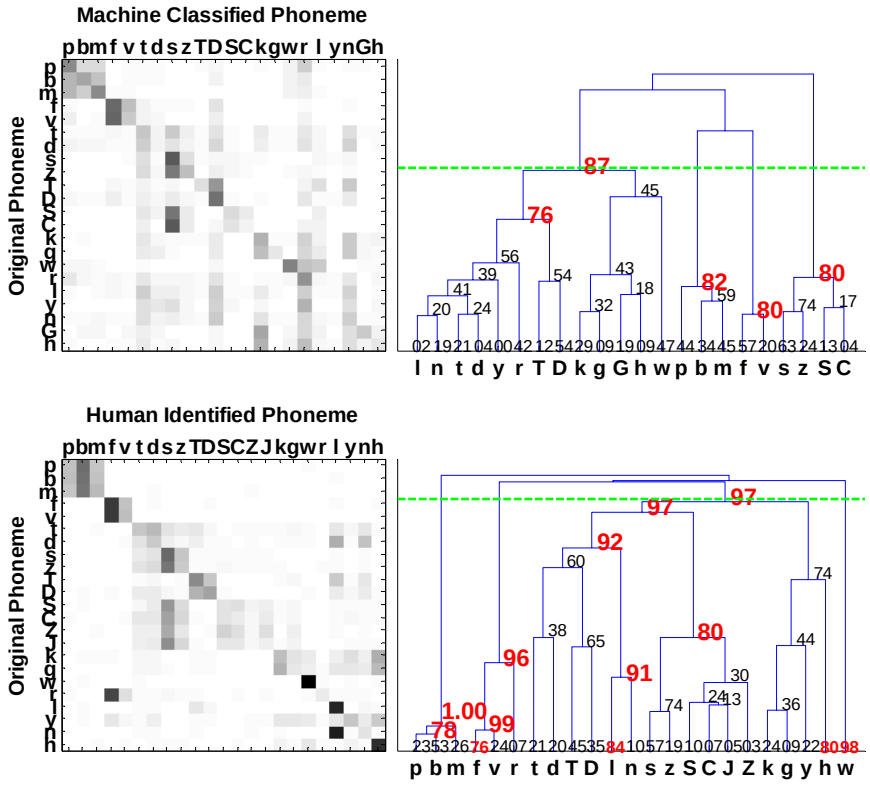

Figure 3. Consonant confusion patterns generated by machine-B classification of 22 consonants with a confusion matrix shown on the left, and a dendrogram shown on the right in the top row, and by human forced-choice identification of 23 CV syllables averaged over the /a, i, u/ vowel conditions [14] in the bottom row. The numbers at each node of the dendrogram show the percent correct score of the corresponding phoneme/phoneme cluster. Those nodes that have percent correct scores ≥ 75% are shown in larger font. The dashed line crosses the phoneme cluster branches that have inner-group classification correct scores ≥ 75% at a consistent hierarchical level.

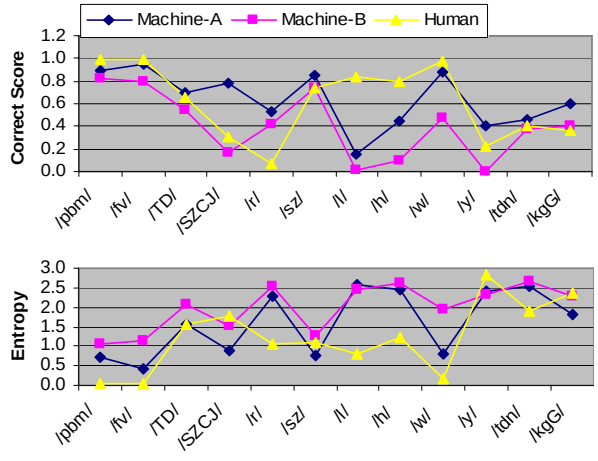

Figure 4. Comparing consonant confusion pattern in terms of identification correct scores (top plot) and entropy (bottom plot) using Kricos-visemse between machine classification and human forced-choice identification.

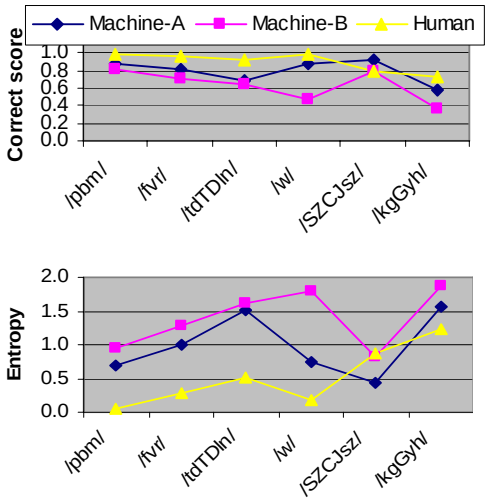

Figure 5. Comparing consonant confusion pattern in terms of identification correct scores (top plot) and entropy (bottom plot) using Jiang-visemes between machine classification and human forced-choice identification.