DEVELOPMENT OF A VISUAL SPEECH SYNTHESIZER VIA SECOND-ORDER ISOMORPHISM

Jintao Jiang, Justin M. Aronoff, and Lynne E. Bernstein

Department of Communication Neuroscience
House Ear Institute, Los Angeles, CA 90057, USA
{jjiang, jaronoff, lbernstein}@hei.org

ABSTRACT

The goals of this study were to evaluate the synthesis of visible speech that was based on 3-D motion data using second-order isomorphism. To do this, word stimuli were generated for perceptual discrimination and identification tasks. Discrimination trials were based on word-pairs that were predicted to be at four levels of perceptual dissimilarity. Results from the discrimination tasks indicated that visual synthetic speech perception maintained the dissimilarity structure of visual natural speech perception. This study demonstrated that the relatively sparse 3-D representations of face motion could be used to synthesize visual speech that perceptually approximate visual natural speech, suggesting that synthesizer development and psychophysics can benefit mutually when the goals are aligned.[1]

Index Terms—Visual speech synthesis, visual speech perception, second-order isomorphism, dissimilarity

1. INTRODUCTION

Research on audiovisual speech perception has demonstrated the multi-modality of human speech perception [1, 2]. People with normal hearing and vision, as well as people with hearing impairments, rely on visual speech information when the acoustic signal is not loud enough and/or is degraded by noise. An automated visual speech synthesizer would allow for generating new, spontaneous, and quantitatively controlled visual speech materials for commercial, research, and clinical applications. Our interest in a synthesized talker is primarily for perceptual and neural research, and for clinical applications. A synthetic talker would allow precise control over stimulus attributes, a fundamental requirement for human perception research.

Talking face animations are now available, but their ability to convey realistic speech information remains inadequate. Perhaps, this is because visual speech synthesis has not taken full advantage of the approach that greatly benefited development of acoustic speech synthesis for research. Acoustic speech synthesis was developed in the context of numerous careful acoustic measurements and perceptual experiments that were carried out over many years [3]. For example, early work on acoustic speech perception employed the pattern playback device, which synthesized a schematic representation of the speech signal [4]. This device assisted, for example, in discovering relationships between auditory segmental perception and formant frequencies.

Currently, meager fundamental knowledge exists concerning optical phonetics—the quantifiable attributes of the optical speech signal that perceivers use to perceive speech [5-7]. We believe that such knowledge is necessary for developing visual speech synthesis that presents the same information produced by humans. The current study was designed to investigate how results in a visual speech perception experiment might be used to give feedback to refine a synthesizer.

The need for detailed optical phonetic studies can be appreciated on consideration of several facts about visual speech perception. Although visual speech signals convey less phonetic information than do acoustic speech signals under good listening conditions, visible speech affords adequate phonetic information to recognize a high percentage of words, as demonstrated by expert deaf lipreaders [8].

The realism we seek to achieve in visual speech synthesis is the information needed to perceive words with the same accuracy as can be obtained by expert deaf lipreaders. However, current evaluation methods for visual speech synthesizers are generally not designed for evaluation of phonetic detail but instead for evaluation of the general appearance of naturalness, the boost in intelligibility obtained under noisy audiovisual conditions, and/or identification within broad viseme classes [9-11].

Nevertheless, having set as the goal phonetic accuracy, a problem is to determine what attributes of natural optical signals are perceptually relevant to phonetic perception. The visual speech stimulus is a complex display. Innumerable cues might be relevant to the expert perceiver. Shepard and Chipman [12] considered the problem of establishing the isomorphism between physical stimuli and internal (perceptual or neural) representations. They noted that internal representations are unlikely to be structurally isomorphic with stimuli in the sense that the internal representation of a square is not likely to be square. In order to approach the problem of establishing relationships between complex stimuli and internal/perceptual/neural representations, they argued that an "isomorphism should be sought – not in the first-order relation between (a) an individual object, and (b) its corresponding internal representation – but in the second-order relation between (a) the relations among alternative external objects, and (b) the relations among their corresponding internal representations. Thus, although the internal representation for a square need not itself be square, it should (whatever it is) at least have a closer functional relation to the internal representation for a rectangle than to that, say, for a green flash or the taste of a persimmon" (p. 2). That is, in the absence of a list of optical phonetic cues, the researcher would be advised to seek isomorphism between physical dissimilarities and perceptual dissimilarities.

[1] Work was supported in part by an NSF award IIS 0312434 (Bernstein, PI).

In [6], a robust second-order isomorphism relationship was established based on dissimilarities among sparse 3-D optical point representations of visible speech and perceptual dissimilarities among videorecorded consonants. In the Jiang-et-al study, subjects visually identified the 23 initial consonants of English spoken by four talkers. The resulting confusion data were transformed into perceptual spaces via multidimensional scaling. Then Euclidean distances were obtained between all pairs of phonemes. The variance accounted for in the perceptual dissimilarity measures by the physical dissimilarities computed using the 3-D point data ranged between 36% and 72% across talkers and vowels (see [6] for details). In other words, the visual perceptual structure was preserved in the sparse optical data. An implication following this demonstration is that synthetic speech should preserve the same second-order isomorphism relationship. That is, the dissimilarity relationships in natural video speech should be represented in the synthetic video speech.

The literature demonstrates that segmental dissimilarities can be used to predict the intelligibility of lipread words [13, 14]. Thus, a visual speech synthesizer should produce natural dissimilarity relationships for both segments and words. Furthermore, perceptual tests of dissimilarity typically involve discrimination tests which can be used to efficiently evaluate synthetic stimuli.

In the present study, words, instead of phonemes or consonant-vowel syllables, were used, because approximating contextual information (coarticulation) is a critical aspect of visual speech synthesis [11]. In addition to *same* word pairs, stimulus word pairs were generated that were predicted, based on segmental perceptual data, to be at *near*, *medium*, and *far* perceptual distances; participants were presented with the word pairs in a same-different discrimination task; and the word pairs comprised either two natural video tokens or one synthetic token paired with one natural video token. If the predicted dissimilarity relationships were obtained with the video-video word pairs, that result would further validate the use of segmental measures to predict word-level dissimilarity. If the predicted dissimilarity relationships were also obtained with video-synthetic word pairs, that would confirm that the sparse optical data preserved dissimilarity relationships at the word level. Deviations from the natural dissimilarity structure would point to areas for improvement in the synthesizer. Subsequent research could focus on manipulation of the sparse data so as to achieve more accurate dissimilarity relationships as evaluated using the discrimination paradigm. For those word pairs whose perceptual dissimilarities were different from the predicted dissimilarities, a pair-wise tuning on the synthesizer could also be performed.

A variety of approaches can be imagined for realizing a visual speech synthesizer [15, 16]: wireframe, muscle-based, and image-based methods. Furthermore, a visual speech synthesizer can be driven with rule-based, concatenative, acoustics-driven, or direct-physical-measures-driven methods [16]. Wireframe models are defined by a set of 3-D polygonal meshes that are controlled with simple geometric operations. Muscle models use polygonal meshes simulating muscle activities that are directly controlled by muscular activations. Image-based methods reproduce speech movements by morphing and interpolating existing speech images. The present study used a wireframe face model and focused on studying the perceptual effects of synthetic speech driven directly from the 3-D optical recordings we obtained as part of the project.

2. METHOD

2.1. Visual speech synthesizer

A face animation model was realized, incorporating a mesh of 3-D polygons that define the head and its parts [17]. The original 3-D face model was obtained from www.digimation.com. This model was later edited (addition, deletion, and modification of some vertices, polygons, and textures) to have 1915 vertices and 1944 polygons. An algorithmic layer allows the mesh to be deformed for performing facial actions as well as preventing errors (such as incursion of the lower lip into the volume of the upper lip). Optical trajectories were registered (calibrated) onto the key points on the face model, and these key points were used to deform the rest of the face vertices with a modified radial basis functions [17]. Radial basis functions have been shown to be effective [9].

Texture is re-mapped onto the deformed face that is then rendered and animated with appropriate lighting and background color using the *openGL* graphics application-programming interface. The synthetic face was scaled and shifted to have the same position and size as the natural face (see Figure 1).

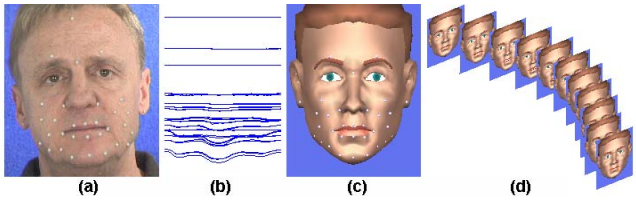

Figure 1. Synthesis: (a) face motions tracked using 33 retro-reflectors; (b) reconstructed and smoothed motion trajectories compensated for head and eyebrow motion and missing data; (c) motion data normalized and registered to key face points; and (d) whole face deformed frame-by-frame using key face points.

2.2. Stimuli

2.2.1. Recorded speech materials

An experienced American-English talker with relatively high visual intelligibility said the stimuli out loud inside a sound-treated booth. Audio, video, and 3-D optical data were recorded simultaneously and synchronized [18]. The video recording equipment was a production quality Sony digital camera and video recorder. Face motion was captured with a 120-Hz Qualisys optical motion capture system using three infrared emitting-receiving cameras and 33 optical retro-reflectors (see Figure 1). Two tokens each of 141 monosyllabic words were recorded.

2.2.2. Data processing

Seventy-six monosyllable words were manually segmented with a closed mouth in the beginning and the end. These segmented video clips were digitized, cropped to a size of 720x480 pixels, and built into 30-Hz interlaced video files.

Reconstructed 3-D motion data were processed to remove head and eyebrow motion, recover missing data, remove noise, normalize the head-size, and smooth the motion [17, 18]. For those 76 video clips, the corresponding 3-D motion data segments were retrieved and were down-sampled to 60 Hz to drive the face animation model directly with a resolution of 720x480 pixels. The resulting AVI videos were then interlaced to produce 30-Hz video.

The natural and synthetic video segments without audio were compressed, and all of the resulting compressed video clips were concatenated into a single large video file that was authored to a DVD to allow frame based searching and random access.

2.2.3. Stimulus selection

Stimuli were selected for discrimination and identification tasks. For one set of word pairs for the discrimination task, thirty-two synthetic tokens were generated. Each token was in a set with four natural video tokens chosen to vary in their perceptual distance from the synthetic token. Each set comprised a *same*, *near*, *medium*, and *far* word. The distances were computed using a perception-based cost matrix [19]. There were 62 words total comprising the natural video comparisons across trials. *Same* pairs comprised the synthetic stimulus and the natural video token that was recorded along with the 3-D optical recording.

Another set of discrimination pairs comprised the same word-pairs, but all of the tokens were the natural videorecordings. The "*same*" pairs comprised different tokens of the same words.

2.3. Perceptual experiments

Eight normal-hearing participants with above-average lipreading ability were recruited. For the discrimination (AX, same-different) trials involving synthetic speech (AXS), a synthetic token was presented, followed by a natural video token. For the discrimination trials with natural video only (AXV), two natural video tokens were presented sequentially. All tokens were presented without audio. Following the discrimination task, participants performed an open-set identification task. For the IDS condition, the 32 synthetic tokens were presented in pseudorandom order. For the IDV condition, the 62 natural video tokens were presented in pseudorandom order. Participants viewed each stimulus and typed at a PC the spoken word they had seen.

3. RESULTS

3.1. Discrimination

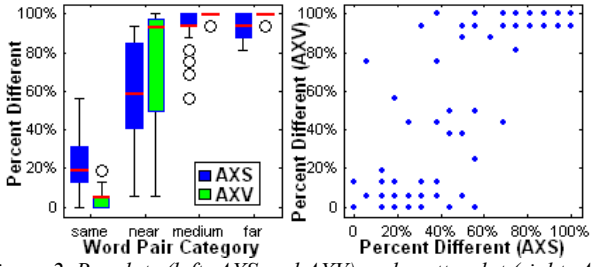

Figure 2. Boxplots (left; AXS and AXV) and scatterplot (right; AXS versus AXV) of percent different responses.

A two-way [Dissimilarity (*same*, *near*, *medium*, *far*) x Media (*synthetic*, *natural*)] repeated measures ANOVA was conducted based on the percentage of different responses with regard to word pairs. As predicted, Dissimilarity reliably affected percent different scores [$F(1.38,42.68) = 287.3$, $p<.001$; Huynh-Feldt adjustment used to correct for the violation of the sphericity assumption]. Planned comparisons confirmed that *medium* pairs were labeled different at a higher rate than *near* pairs, and *near* pairs more than *same* pairs (all *p*s<.001). *Medium* and *far* pairs did not differ.

The main effect of Media was reliable [$F(1,31)=6.2$, $p<.02$]. The Dissimilarity x Media interaction was significant [$F(2.07,64.07)=24.8$, $p<.001$], reflecting the overall higher accuracy for the AXV condition. Higher accuracy indicated that the rate of different responses was higher for different pairs but lower for *same* pairs (see Figure 2). Also in Figure 2 (right), percent different scores for AXS were correlated with those for AXV [$R^2=.78$, $F(1,126)=446.2$, $p=.000$], confirming the effectiveness of the synthetic speech in approximating visual natural speech in terms of perception.

Proportion different scores for *near*, *medium*, or *far* pairs in AXS and AXV were submitted to linear regression analyses using as the predictors the pre-computed natural log-transformed dissimilarity measures. Figure 3 shows that the dissimilarity scores accounted for a significant portion of the variance in the percent different responses for AXS [$R^2=.53$, $F(1,94)=107.0$, $p<.001$] and for AXV [$R^2=.31$, $F(1,94)=44.0$, $p<.001$].

In Figure 3, some pairs (e.g., needs-case, best-space, and sent-tax) were far from the regression line for AXS. However, these pairs were close to the regression line for AXV. This implies that the synthesizer did not adequately differentiate words in these pairs. For example, the tongue and teeth were not modeled in the present visual speech synthesizer. As a novel approach, we can focus on these outlier word pairs to fine-tune the synthesizer.

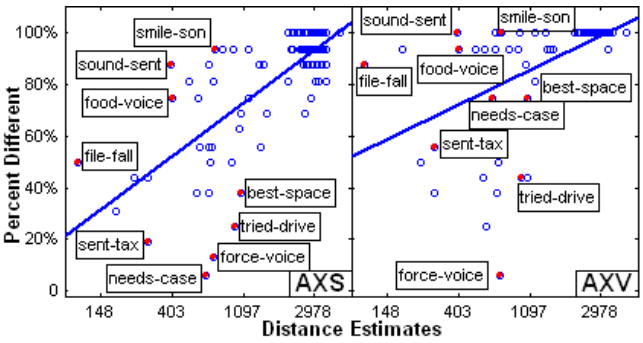

Figure 3. Percent different scores for each word pair in AXS (left) and AXV (right) were plotted against distance estimates (pre-computed using the perception-based cost matrix).

3.2. Identification

For the open-set word identification task, responses were scored in terms of phoneme accuracy and uncertainty [8]. The open-set word identification with synthetic speech was shown to be highly inaccurate. Thus, while the synthesizer preserved natural dissimilarities, it was still quite deficient in optical phonetic detail.

Analyses were undertaken to determine whether the incorrect identification responses were perceptually related to the stimuli or chosen without regard to the stimuli. The contemporary view of spoken word recognition is that identification entails discrimination of competing lexical items in long-term memory [20]. The prediction here is that perceptual errors in word identification result in closer stimulus-response distances than words selected independently of perception. To investigate this, dissimilarity was computed between stimulus words and corresponding incorrect responses. Only incorrect responses resulting from the 32 words that were presented as both synthesized and natural video words were used. Additionally, dissimilarities were computed between stimulus words and words randomly selected from an online dictionary [21]. Results indicated that the mean dissimilarities for the actual stimulus-response pairs [M=470 (SD=157) for IDS and M=242 (SD=153) for IDV] were significantly less than those from random selection [M=680 (SD=89); $F(2,54)=79.7$, $p<.001$].

Phoneme identification accuracy. Given that the synthetic stimuli grossly sampled the motion of the talking face, and that there was no indication of tongue gestures in the synthesis, the expectation was that certain phonemes would be less accurately conveyed than others. The stimulus-response pairs were submitted to an alignment procedure to obtain percent correct scores for each phoneme [19]. Results showed that for many phonemes the *natural*

and *synthetic* video stimuli were proportionally similar. In Figure 4, the distance between the two lines (*slope*=1) was fixed to be 30%, and the two lines were positioned so that the greatest number of phonemes fell between them (62.5%). Some *synthetic* phonemes performed much worse than their *natural* counterparts. Those phonemes likely did not provide sufficient information regarding lip rounding (/o, u, U, W/) and tongue position (/l, g, T, S/).

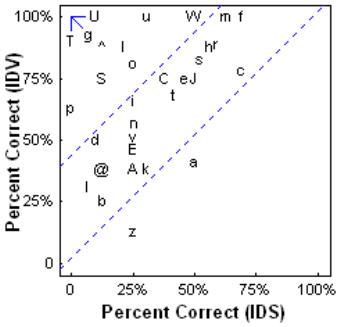

Figure 4. Percent phoneme correct for IDV versus that for IDS. The two dashed lines (slope=1) had fixed distance of 30% and were positioned to have the greatest number of phonemes fall between. Solid lines indicate the overlap of /U, g, T/ on the graph.

Phoneme uncertainty. Phoneme uncertainty [8] was also calculated to quantify the consistency with which participants responded. This analysis used both the substitutions and correct phonemes output by the alignment procedure [19]. Phoneme uncertainty was calculated as $uncertainty = \Sigma[-p \cdot \log_2(p)]$, where p is the proportion of a particular phoneme response for a given phoneme stimulus.

The results in Figure 5 indicated that uncertainty was higher for synthetic than for natural video stimuli. Also, a comparison between Figures 4 and 5 suggests that highly errorful phonemes also produced high phoneme uncertainty rather than a response bias. This supports the view that the stimuli were deficient.

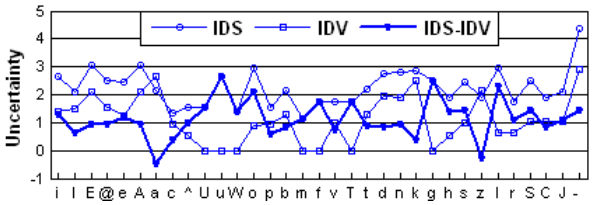

Figure 5. Uncertainty computed for IDS and IDV conditions.

4. DISCUSSION

This study demonstrated that the relatively sparse representations of speech data obtained with 3-D optical recordings can be used to synthesize and perceive speech. The perceptual tasks involved discrimination and identification. Importantly, the stimuli preserved the perceptually relevant similarity relationships among words in the mental lexicon, even though the open-set identification of the synthesized words was mostly inaccurate. Additionally, the identification task provides information that can aid improving the model at the phoneme level. Using the same methods reported here, we would like in the future to show enhanced performance with a more accurate, enhanced synthesizer. That synthesizer will in turn afford better stimuli for experiments in perception and its neural underpinnings.

5. REFERENCES

[1] H. McGurk & J. MacDonald, "Hearing lips and seeing voices," *Nature* 264, pp. 746-48, 1976.

[2] W.H. Sumby & I. Pollack, "Visual contribution to speech intelligibility in noise," *JASA* 26, pp. 212-15, 1954.

[3] D.H. Klatt, "Review of text-to-speech conversion for English," *JASA* 82, pp. 737-93, 1987.

[4] F.S. Cooper, P.C. Delattre, A.M. Liberman, *et al.*, "Some experiments on the perception of synthetic speech sounds," *JASA* 24, pp. 597-606, 1952.

[5] L.E. Bernstein, "Visual speech perception," in *Audio-visual speech processing*, E. Vatikiotis-Bateson, G. Bailly, & P. Perrier, Editors, MIT Press, Cambridge, MA, 2006.

[6] J. Jiang, E.T. Auer, Jr., A. Alwan, *et al.*, "Similarity structure in visual speech perception and optical phonetic signals," *Percept. Psychophys.* 69, pp. 1070-83, 2007.

[7] O. Govokhina, G. Bailly, & G. Breton, "Learning optimal audiovisual phasing for an HMM-based control model for facial animation," *ISCA Workshop on Speech Synthesis*, Bonn, Germany, 2007.

[8] L.E. Bernstein, M.E. Demorest, & P.E. Tucker, "Speech perception without hearing," *Percept. Psychophys.* 62, pp. 233-52, 2000.

[9] J. Ma, R. Cole, B. Pellom, *et al.*, "Accurate visible speech synthesis based on concatenating variable length motion capture data," *IEEE Trans. Vis. & Comp. Graphics* 12, pp. 266-76, 2006.

[10] C. Benoît, T. Guiard-Marigny, B. Le Goff, *et al.*, "Which components of the face do humans and machines best speechread?," in *Speechreading by humans and machines: Models, systems, and applications*, D.G. Stork & M.E. Hennecke, Editors, Springer, Berlin, pp. 315-28, 1996.

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

[12] R.N. Shepard & S. Chipman, "Second-order isomorphism of internal representations: Shapes of states," *Cog. Psychol.* 1, pp. 1-17, 1970.

[13] S.L. Mattys, L.E. Bernstein, & E.T. Auer, Jr., "Stimulus-based lexical distinctiveness as a general word-recognition mechanism," *Percept. Psychophys.* 64, pp. 667-79, 2002.

[14] E.T. Auer, Jr., "The influence of the lexicon on speech read word recognition: Contrasting segmental and lexical distinctiveness," *Psychon. Bull. Rev.* 9, pp. 341-47, 2002.

[15] F.I. Parke & K. Waters, *Computer facial animation*. A.K. Peters, Natick, MA, 1996.

[16] G. Bailly, "Audiovisual speech synthesis. From ground truth to models," *ICSLP*, Denver, CO, 2002.

[17] J. Xue, J. Borgstrom, J. Jiang, *et al.*, "Acoustically-driven talking face synthesis using dynamic Bayesian networks," *ICME*, Toronto, Canada, 2006.

[18] J. Jiang, A. Alwan, P. Keating, *et al.*, "On the relationship between face movements, tongue movements, and speech acoustics," *EURASIP JASP* 2002, pp. 1174-88, 2002.

[19] L.E. Bernstein, M.E. Demorest, & S.P. Eberhardt, "A computational approach to analyzing sentential speech perception: phoneme-to-phoneme stimulus-response alignment," *JASA* 95, pp. 3617-22, 1994.

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

[21] P.F. Seitz, L.E. Bernstein, E.T. Auer, Jr., *et al.*, *PhLex (Phonologically Transformable Lexicon): A 35,000-word computer readable pronouncing American English lexicon on structural principles, with accompanying phonological transformations, and word frequencies*, [Database], 1998.