7. SUMMARY AND CONCLUSION

In this study, relationships among face movements, tongue movements, and acoustic data were quantified through correlation analyses on CVs and sentences using multilinear regression. In general, predictions for syllables yielded higher correlations than those for sentences. Furthermore, it demonstrated that multilinear regression, when applied to short speech segments such as CV syllables, was successful in predicting articulatory movements from speech acoustics, and the correlations between tongue and face movements were high. For sentences, the correlations were lower suggesting that nonlinear techniques might be more applicable or that the correlations should be computed on a short-time basis. For CV syllables, the correlations between OPT and EMA data were medium to high (correlations ranged from 0.70 to 0.88). Articulatory data (OPT or EMA) can be well predicted from LSPE (correlations ranged from 0.74 to 0.82). LSP data were better predicted from EMA than from OPT (0.54–0.61 vs. 0.37–0.55), which is expected from the speech production model point of view: the vocal tract is shaped to produce speech, while face movements are a by-product, and thus contain variance unrelated to speech acoustics.

Another fact about these correlations was asymmetry of the predictions. In general, articulatory movements were easier to predict from speech acoustics than the reverse. This is because speech acoustics are more informative than visual movements. Lipreading accuracy for these CV syllables ranged from 30% to 40% [38], while listening accuracy should be very high. Another reason may be that all frequency components were weighted equally. Articulatory movements, however, are very slow, about 15–20 Hz, and most frequency components are even lower than 5 Hz. Therefore, when dealing with acoustic data, low frequency components may need to be emphasized, or weighted differently.

The study also investigated the effect of intelligibility and gender of the talker, vowel context, place of articulation, voicing, and manner of articulation. The results reported here did not show a clear effect of intelligibility of the talker, while the data from the two males gave better predictions than those from the two females. Note that the data from talker M1, who had the largest face among the four talkers, yielded reasonably good predictions. Therefore, face size may be an effect in the predictions. For visual synthesis, talker effects should be accounted for.

Results also showed that the prediction of C/a/ syllables was better than C/i/ and C/u/. Furthermore, vowel-dependent predictions produced much better correlations than syllable-independent predictions. Across different places of articulation, lingual places in general resulted in better predictions of one data stream from another compared to bilabial and glottal places. Among the manners of articulation, plosive consonants yielded lower correlations than others, while voicing had no influence on the correlations.

For both syllable-dependent and sentence-independent predictions, prediction of individual channels was also examined. The chin movements were the best predicted, followed by lips, and then cheeks. In regards to the acoustic features, the second LSP pair, which is around the second formant frequency, and RMS energy, which is related to mouth aperture, were better predicted than other LSP pairs. This may suggest that in the future, when predicting face or tongue movements from speech acoustics, more resolution could be placed around the 2nd LSP pair. The RMS energy can be reliably predicted from face movements. The internal tongue movements cannot predict the RMS energy and LSP well over long periods (sentences), while they were predicted reasonably well for short periods (CVs).

Another question we examined was the magnitude of predictions based on a reduced data set. For both CVs and sentences, a large level of redundancy among TB, TM, and TT and among chin, cheek, and lip movements was found. One implication was that the cheek movements can convey significant information about the tongue and speech acoustics, but these movements were redundant to some degree if chin and lip movements were present. The three pellets on the tongue captured the frontal-tongue movements of certain consonants well. Data from additional movements about the vocal tract around the glottis, velar, and inner lip areas might have improved the predictions. For CVs, using one channel or all channels did not make a difference, except when predicting LSPs from OPT, where the chin movements were the most informative. For sentences, using all channels usually resulted in better prediction; lip movements were the most informative when predicting LSP or EMA; when predicting LSP or OPT, TT was the most informative channel.

In [16], the authors showed that the coupling between the vocal-tract and the face is more closely related to human physiology than to language-specific phonetic features. However, this was phoneme-dependent, and this is why it is interesting to examine the relationships using CV syllables. In [16], the authors stated that the most likely connection between the tongue and the face is indirectly by way of the jaw. Other than the biomechanical coupling, another source is the control strategy for the tongue and cheeks. For example, when the vocal tract is shortened the tongue does not have to be retracted. This is reflected in analyses obtained as a function of place and manner of articulation (in Figures 9 and 10).

A limitation of this study is that correlation analysis was carried out uniformly across time without taking into account important gestures or facial landmarks. For example, some specific face gestures or movements may be very important for visual speech perception, such as mouth closure for a bilabial sound. In the future, physiological and perceptual experiments should be conducted to define which face movements are of importance to visual speech perception, so that those movements are better predicted. So far, the results are not adequate for reconstructing speech acoustics from face movements only. Noisy speech, however, can be enhanced by using information from face movements [5]. If articulatory movements could be recovered from speech acoustics, a shortcut for visual speech synthesis might be achieved.

ACKNOWLEDGMENTS

We wish to acknowledge the help of Brian Chaney, Sven Mattys, Taehong Cho, Paula E. Tucker, and Jennifer Yarbrough in data collection. This work was supported in part by an NSF-KDI Award 9996088.

REFERENCES

[1] D. Lindsay, "Talking head," *American Heritage of Invention & Technology*, vol. 13, no. 1, pp. 57–63, 1997.

[2] P. Rubin and E. Vatikiotis-Bateson, "The talking heads site: http://www.haskins.yale.edu/haskins/heads.html," 1998.

[3] J. Luettin and S. Dupont, "Continuous audio-visual speech recognition," in *Proc. 5th European Conference on Computer Vision*, H. Burkhardt and B. Neumann, Eds., vol. II of *Lecture Notes in Computer Science*, pp. 657–673, Freiburg, Germany, June 1998.

[4] G. Potamianos, J. Luettin, and C. Neti, "Hierarchical discriminant features for audio-visual LVCSR," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Salt Lake City, Utah, USA, 2001.

[5] L. Girin, J. L. Schwartz, and G. Feng, "Audio-visual enhancement of speech in noise," *J. Acoust. Soc. Am.*, vol. 109, no. 6, pp. 3007–3020, 2001.

[6] A. Alwan, S. Narayanan, and K. Haker, "Towards articulatory-acoustic models of liquid consonants part II: the rhotics," *J. Acoust. Soc. Am.*, vol. 101, no. 2, pp. 1078–1089, 1997.

[7] B. S. Atal, J. J. Chang, M. V. Mathews, and J. W. Tukey, "Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer sorting technique," *J. Acoust. Soc. Am.*, vol. 63, no. 5, pp. 1535–1556, 1978.

[8] P. Badin, D. Beautemps, R. Laboissiere, and J. L. Schwartz, "Recovery of vocal tract geometry from formants for vowels and fricative consonants using a midsagittal-to-area function conversion model," *J. Phonetics*, vol. 23, pp. 221–229, 1995.

[9] G. Fant, *Acoustic Theory of Speech Production*, S-Gravenhage, Mouton, The Netherlands, 1960.

[10] J. L. Flanagan, *Speech Analysis Synthesis and Perception*, Springer-Verlag, Berlin, 1965.

[11] S. Narayanan and A. Alwan, "Articulatory-acoustic models for fricative consonants," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 3, pp. 328–344, 2000.

[12] J. Schroeter and M. Sondhi, "Techniques for estimating vocal-tract shapes from the speech signal," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 1, pp. 133–150, 1994.

[13] K. N. Stevens and A. S. House, "Development of a quantitative description of vowel articulation," *J. Acoust. Soc. Am.*, vol. 27, pp. 484–493, May 1955.

[14] K. N. Stevens, "The quantal nature of speech: Evidence from articulatory-acoustic data," in *Human Communication: A Unified View*, E. E. David and P. B. Denes, Eds., pp. 51–66, McGraw-Hill, New York, USA, 1972.

[15] J. P. Barker and F. Berthommier, "Estimation of speech acoustics from visual speech features: A comparison of linear and non-linear models," in *Proc. AVSP '99*, pp. 112–117, Santa Cruz, Calif, USA, 1999.

[16] H. C. Yehia, P. Rubin, and E. Vatikiotis-Bateson, "Quantitative association of vocal-tract and facial behavior," *Speech Communication*, vol. 26, no. 1, pp. 23–43, 1998.

[17] N. Sugamura and F. Itakura, "Speech analysis and synthesis methods developed at ECL in NTT—from LPC to LSP," *Speech Communication*, vol. 5, pp. 199–215, 1986.

[18] E. Agelfors, J. Beskow, B. Granström, et al., "Synthetic visual speech driven from auditory speech," in *Proc. AVSP '99*, pp. 123–127, Santa Cruz, Calif, USA, 1999.

[19] E. Yamamoto, S. Nakamura, and K. Shikano, "Lip movement synthesis from speech based on hidden Markov models," *Speech Communication*, vol. 26, no. 1-2, pp. 105–115, 1998.

[20] H. C. Yehia, T. Kuratate, and E. Vatikiotis-Bateson, "Using speech acoustics to drive facial motion," in *Proc. the 14th International Congress of Phonetic Sciences*, pp. 631–634, San Francisco, Calif, USA, 1999.

[21] A. V. Barbosa and H. C. Yehia, "Measuring the relation between speech acoustics and 2D facial motion," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Salt Lake City, Utah, USA, 2001.

[22] A. Sen and M. Srivastava, *Regression Analysis*, Springer-Verlag, New York, USA, 1990.

[23] Carstens Medizinelektronik GmbH, "Articulograph AG100 User's Handbook," 1993.

[24] P. W. Schönle, K. Gräbe, P. Wenig, J. Höhne, J. Schrader, and B. Conrad, "Electromagnetic articulography: use of alternating magnetic fields for tracking movements of multiple points inside and outside the vocal tract," *Brain and Language*, vol. 31, pp. 26–35, 1987.

[25] L. E. Bernstein, E. T. Auer Jr., B. Chaney, A. Alwan, and P. A. Keating, "Development of a facility for simultaneous recordings of acoustic, optical (3-D motion and video), and physiological speech data," *J. Acoust. Soc. Am.*, vol. 107, no. 5, pp. 2887, 2000.

[26] J. Jiang, A. Alwan, L. E. Bernstein, P. A. Keating, and E. T. Auer, "On the correlation between orofacial movements, tongue movements and speech acoustics," in *International Congress on Spoken Language Processing*, pp. 42–45, Beijing, China, 2000.

[27] P. A. Keating, T. Cho, M. Baroni, et al., "Articulation of word and sentence stress," *J. Acoust. Soc. Am.*, vol. 108, no. 5, pp. 2466, 2000.

[28] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*, Prentice Hall, Englewood Cliffs, NJ, USA, 1978.

[29] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, New York, NY, USA, 1985.

[30] B. Efron, *The Jacknife, the Bootstrap, and Other Resampling Plans*, vol. 38 of *CBMS-NSF Regional Conference Series in Applied Mathematics*, Society for Industrial and Applied Mathematics, Philadelphia, Pa, USA, 1982.

[31] R. G. Miller, "The jacknife-a review," *Biometrika*, vol. 61, no. 1, pp. 1–15, 1974.

[32] D. Bertsimas and J. N. Tsitsiklis, *Introduction to Linear Optimization*, Athena Scientific, Belmont, Mass, USA, 1997.

[33] P. Ladefoged, *A Course in Phonetics*, Harcourt College Publishers, Fort Worth, Tex, USA, 4th edition, 2001.

[34] L. Sachs, *Applied Statistics: A Handbook of Techniques*, Springer-Verlag, New York, USA, 2nd edition, 1984.

[35] Q. Zhu and A. Alwan, "On the use of variable frame rate analysis in speech recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 3, pp. 1783–1786, Istanbul, Turkey, June 2000.

[36] M. C. Langereis, A. J. Bosman, A. F. Olphen, and G. F. Smoorenburg, "Relation between speech perception and speech production in adult cochlear implant users," in *The Nature of Speech Perception Workshop*, Utrecht, Nederlands, 2000.

[37] K. W. Grant and P. F. Seitz, "The use of visible speech cues for improving auditory detection of spoken sentences," *J. Acoust. Soc. Am.*, vol. 108, no. 3, pp. 1197–1208, 2000.

[38] J. Jiang, A. Alwan, E. T. Auer, and L. E. Bernstein, "Predicting visual consonant perception from physical measures," in *Proc. Eurospeech '01*, vol. 1, pp. 179–182, Aalborg, Denmark, 2001.

Jintao Jiang received his B.S. (with honors) and M.S. degrees in electronic engineering from Tsinghua University, Beijing, China, in 1995 and 1998, respectively. In 1998, he joined University of California at Los Angeles where he is a Ph.D. candidate in the Electrical Engineering Department. Mr. Jiang has been with Speech Processing and Auditory Perception Laboratory at UCLA since 1998. His research interests include audio-visual speech processing, visual speech perception, and speech recognition.

Abeer Alwan received her Ph.D. in electrical engineering from MIT in 1992. Since then, she has been with the Electrical Engineering Department at UCLA as an Assistant Professor (1992–1996), Associate Professor (1996–2000), and Professor (2000–present). Prof. Alwan established and directs the Speech Processing and Auditory Perception Laboratory at UCLA. Her research interests include modeling human speech production and perception mechanisms and applying these models to speech-processing applications such as automatic recognition, compression, and synthesis. She is the recipient of the NSF Research Initiation Award (1993), the NIH FIRST Career Development Award (1994), the UCLA-TRW Excellence in Teaching Award (1994), the NSF Career Development Award (1995), and the Okawa Foundation Award in Telecommunications (1997). Dr. Alwan is an elected member of Eta Kappa Nu, Sigma Xi, Tau Beta Pi, and the New York Academy of Sciences. She served, as an elected member, on the Acoustical Society of America Technical Committee on Speech Communication (1993–1999), on the IEEE Signal Processing Technical Committees on Audio and Electroacoustics (1996–2000), and on Speech Processing (1996–2001). She is an Editor-in-Chief of the Journal Speech Communication.

Patricia A. Keating received her Ph.D. degree in Linguistics in 1980 from Brown University, and then held an NIH postdoctoral fellowship in the Speech Communications Group at MIT. She is Professor of Linguistics and Director of the Phonetics Laboratory at UCLA. Her main areas of research and publication are experimental and theoretical phonetics, and the phonology-phonetics interface, including topics on speech production, prosody, and dyslexia. She has contributed to several encyclopedias, handbooks, and textbooks, such as the MIT Encyclopedia of the Cognitive Sciences, Linguistics: The Cambridge Survey, and Linguistics: An Introduction to Linguistic Theory. She has been on the editorial boards of the journals Language and Phonology and of the book series Phonetics and Phonology and Oxford Studies in Theoretical Linguistics. A past member of the committee for Conferences in Laboratory Phonology and of the National Science Foundation Linguistics Advisory Panel, and currently a member of the Speech Communication Technical Committee of the Acoustical Society of America, she represented the Linguistic Society of America at the 1998 Coalition for National Science Funding Exhibition and Reception for Members of Congress.

Edward T. Auer Jr. received his Ph.D. degree in cognitive psychology in 1992 from the State University of New York at Buffalo. Currently, he is a scientist in the Department of Communication Neuroscience at the House Ear Institute in Los Angeles, California. He also holds an adjunct faculty appointment in the Communicative Disorders Department at the California State University, Northridge. His research uses combination of behavioral, computational, and functional brain imaging methodologies to investigate the function and development of the spoken language processing system in hearing impaired and hearing individuals. His research includes studies of visual spoken word recognition in deaf adults; novel signals to enhance speech reception (e.g., vibrotactile aids, cued speech); communication-channel-specific effects of word experience; and auditory spoken word recognition. To learn more about his work visit the House Ear Institute website at http://www.hei.org.

Lynne E. Bernstein received her doctorate in Psycholinguistics from the University of Michigan in 1977. She is currently a senior scientist and head of the Department of Communication Neuroscience at the House Ear Institute. She is also an Adjunct Professor in the Department of Linguistics at the University of California, Los Angeles. She is a Fellow of the Acoustical Society of America. Between receiving her doctoral degree and the present, she has investigated a range of topics associated with speech perception. They have included the most typical area, which is auditory speech perception, the less typical area, which is visual speech perception (lipreading), and the least typical area, which is vibrotactile speech perception. A fundamental issue in her work is how the perceptual systems preserve and combine the information needed to represent the spoken words in a language. In recent years, with the advance of brain imaging and electrophysiological methods, her studies of this issue have extended to investigating its neural substrates. The current paper fits into the scheme of her work, providing a better understanding of the signals that are processed during multimodal speech perception. More about this work and the National Science Foundation project that supports it can be found at http://www.hei.org/research/projects/comneur/kdipage.htm.