PAPER 4.

Massaro, D.W., Beskow, J., Cohen, M.M., Fry C.L., Rodriquez, T. (1999). Picture My Voice: Audio to Visual Speech Synthesis using Artificial Neural Networks. In D. W. Massaro (Ed.) *Proceedings of International Conference on Auditory-Visual Speech Processing (AVSP'99)*, Santa Cruz, CA. pp. 133-138.

PICTURE MY VOICE: AUDIO TO VISUAL SPEECH SYNTHESIS USING ARTIFICIAL NEURAL NETORKS

Dominic W. Massaro, Jonas Beskow, Michael M. Cohen, Christopher L. Fry, and Tony Rodriguez

Perceptual Science Laboratory, University of California, Santa Cruz, Santa Cruz, CA 95064 U. S. A.

ABSTRACT

This paper presents an initial implementation and evaluation of a system that synthesizes visual speech directly from the acoustic waveform. An artificial neural network (ANN) was trained to map the cepstral coefficients of an individual's natural speech to the control parameters of an animated synthetic talking head. We trained on two data sets; one was a set of 400 words spoken in isolation by a single speaker and the other a subset of extemporaneous speech from 10 different speakers. The system showed learning in both cases. A perceptual evaluation test indicated that the system's generalization to new words by the same speaker provides significant visible information, but significantly below that given by a text-to-speech algorithm.

1 INTRODUCTION

Persons find it hard to communicate when the auditory conditions are poor, e.g. due to noise, limited bandwidth, or hearing-impairment. Under such circumstances, face-to-face communication is preferable. The visual component of speech can compensate for a substantial loss in the speech signal. This so-called superadditive combination of auditory and visual speech can produce a bimodal accuracy, which is greater than the simple sum of their separate unimodal performances [9]. An even more striking result is that the strong influence of visible speech is not limited to situations with degraded auditory input. A perceiver's recognition of a noise-free auditory-visual phrase reflects the contribution of both sound and sight. For example, if the (non-meaningful) auditory sentence, "My bab pop me poo brive", is paired with the visible sentence, "My gag kok me koo grive", the perceiver is likely to hear, "My dad taught me to drive". Two ambiguous sources of information are combined to create a meaningful interpretation [9,10].

1.1 Faces in Telecommunication

Although we benefit from face-to-face dialog, current technology precludes it when the conversationalists are at a distance and must communicate electronically, such as over the telephone or over the Internet. One option is videoconferencing, but the visual quality and frame-rate provided by such systems with reasonable bandwidth constraints are normally too poor to be useful for speech-reading purposes.

Having developed a three-dimensional talking head, we are interested in its application in telecommunications. As has been shown by several researchers [3,9,10], animated talking faces can account for significant intelligibility gains over the auditory alone condition, almost comparable to a real speaker's face. There are two methods to exploit the real time use of talking faces in human-human dialog. The most obvious involves text-to-speech (TtS) synthesis. By transmitting the symbolic message over the phone line or the Internet, this information could be used to animate a talking face at the receiving station of the participant. A

standard text-to-speech engine would translate the symbolic (written) message into a string of spoken segments [14]. The face movements of the talking head would be aligned with these synthetic speech segments. Texture mapping technology [9] would potentially allow a person's email to be spoken aloud by a talking head, which resembles the original sender. The downside of this technology is that the voice would not correspond to the voice of the sender and furthermore, synthetic auditory speech is heard as robot-like with very little prosodic and emotional structure.

The second approach to audible/visible speech synthesis uses the original auditory speech in its output. With this technique, the animated talking head is generated from and aligned with the original speech of the talker. In order to do this, it is first necessary to *identify* the segments in the utterance either directly or via recognition of the words, so that the appropriate mouth and facial movements can be determined. A potential limitation of this approach is that automatic speech recognition is not accurate enough to provide a reliable transcription of the utterance. For more reliable performance, the user can type the actual utterance in addition to saying it. By aligning the speech waveform and its phonetic transcription [15], it would then be possible to determine and implement the appropriate facial movements of the talking head, a function currently available in the CSLU toolkit [http://cslu.cse.ogi.edu/toolkit/].

1.2 Previous Research

Several researchers have investigated techniques for fully automatic generation of lip-movements from speech. The research fits within the two methods described in the previous section. The first method is based around a discrete classification stage to divide the speech into language units such as phonemes, visemes or syllables, followed by a synthesis stage. This approach has been employed by several investigators [8,12,16] In one study [16], auditory/visual syllable or phoneme/viseme HMMs were trained with both auditory and visual speech features. Context dependent lip parameters were generated by looking ahead to the HMM state sequence that was obtained using context independent HMMs.

The second group of methods does not attempt a direct classification into discrete meaningful classes, but rather tries to map the acoustics directly to continuous visual parameters, using some statistical method. Visible speech control parameters for either lip movement [13,16] or a complete talking head [11] are computed from the auditory speech signal directly. Morishima [11] trained a network to go from LPC Cepstrum speech coefficients to mouth-shape parameters. He trained on 75 speakers and included only one time step of speech information for his network. Another approach is a vector-quantization (VQ) based method maps a VQ code word vector of an input acoustic speech signal to lip parameters frame-by-frame [16].

1.3 Baldi, the Talking Head

Our talking head, called Baldi, is shown in Figure 1. His existence and functionality depend on computer animation and text-to-speech synthesis. His speech is controlled by about 3 dozen parameters. With our completely animated, synthetic, talking head we can control the parameters of visible speech and determine its informative properties. Experiments by Cohen, Walker, and Massaro [5] and Massaro [9] have shown that visible speech produced by the synthetic head, even in its adumbrated form, is almost comparable to that of a real human.

The talking head can be animated on a standard PC, and requires no specialized hardware other than a good 3D graphics card, which is now standard on many computers. In addition, we have a desktop application in which any person's face can be manually adjusted and mapped onto the talking head. A single image of a person, once adjusted to fit on the talking head, can be moved appropriately [9].

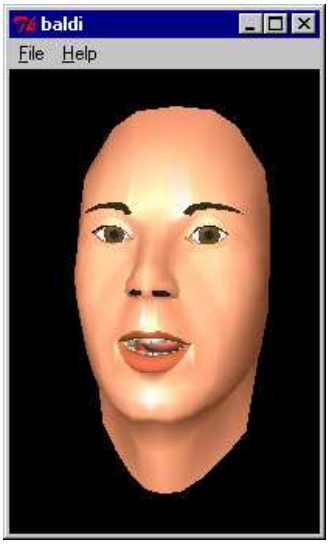

Figure 1. The animated talking head called Baldi.

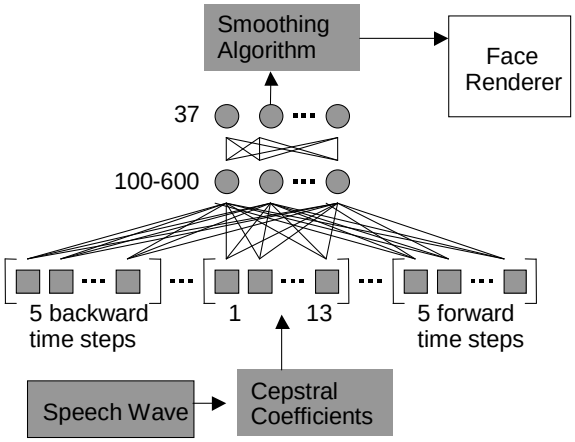

Figure 2. The model architecture of our parameter estimator.

1.4 An Acoustic Speech to Visual Speech Synthesizer

A system that reliably translates natural auditory speech into synthetic visible speech would normally require the following components.

1. A labeled data base of auditory/visual speech,
2. A representation of both the auditory and visual speech
3. Some method to describe the relationship between two representations, and
4. A technique to synthesize the visible speech given the auditory speech.

There are several labeled databases of auditory speech but no readily available labeled databases of visual speech. Given the lack of databases for visible speech, investigators have created their own in order to carry out auditory-to-visible speech synthesis. In some cases, 3D motion capture systems are used utilizing reflective markers on the face [2]. In other cases lip contours are traced using image processing techniques [3,12]. The resulting measurements can be used as inputs to the visible speech synthesis.

An alternative to a recorded auditory/visible speech data base, is to define the properties of the visible speech a priori in terms of synthesis parameters for each speech segment. Given our previous research and current technology, we know which facial movements should be made for each spoken speech segment [6, 9, Chapters 12 & 13]. For example, the mouth is closed at the onset of /b/ and open at the onset of /d/. In our development we have determined synthesis parameters that create intelligible speech approximating the visible speech produced by a natural speaker. The facial movements are realistic because they have been fine-tuned to resemble a natural talker as much as possible [9]. These control parameters then serve as labeled representation of the visible speech.

Our system takes natural auditory speech and maps it into movements of our animated talking head that are aligned appropriately with the auditory speech. Our goal is to go directly from the auditory speech to these specific movements. We determined the mapping between the acoustic speech and the appropriate visual speech movements by training an arti-

ficial neural network to associate or map fundamental acoustic properties of auditory speech to our visible speech parameters. Neural networks have been shown to be efficient and robust learning machines which solve an input-output mapping and have been used in the past to perform similar mappings from acoustics to visual speech. We report the results of training the network against two different databases: isolated words and extemporaneous speech.

2. EXPERIMENT 1: WORDS

2.1 Method

We used a bimodally recorded test list in natural speech that is available to the speech and animation communities. This data set existed in the form of a corpus of one-syllable words presented in citation speech on the Bernstein and Eberhardt [4] videodisk. This laser-man data set represents a potentially straightforward task for the network; the words are isolated and had a predictable structure. The training set (about 10 minutes worth of speech) consisted of 400 words, randomly selected out of the 468 words, leaving 68 words for testing. The audio was digitized with a PC soundcard at 8 bit/16 kHz.

From the acoustic waveform we generated cepstrum coefficients at 50 frames per second. 13 coefficients were generated using 21 Mel-scaled filters, using overlapping hamming windows with a width of 32 ms.

Desired output parameters were generated as follows: The digitized waveforms and the corresponding text, were input into a Viterbi-based forced alignment program, that produced time-aligned phoneme labels for all of the words in the database. Using the time-aligned phoneme labels, 37 control parameters for the talking head were generated at 50 frames per second, using our current visual speech TtS algorithm [9, pp. 379-390]. Two sets of tongue parameters for the simple and complex tongue models and the three visible cues used in our training studies [9, pp. 437-442] are included as outputs of the network. Furthermore, since the activation values of the networks' output nodes are constrained to lie in the range 0.0 to 1.0, each parameter was normalized relative to it's minimum and maximum values over the entire data set in such a way that all parameters varied between 0.05 and 0.95.

We used a feed-forward artificial neural network (ANN) with three layers, as shown in Figure 2. The acoustic input is streamed at 50 frames a second. At every frame, 13 cepstral parameters serve as the input to 13 input units. All of the 13 input parameters were taken at eleven consecutive time frames (current + five frames back + five frames forward) yielding a total of 143 input nodes and 37 output nodes. Networks with 100, 200, 400 and 600 hidden units were trained using the back-propagation algorithm with a learning rate of 0.005 during 500 iterations. We found that increasing the number of hidden units improved generalization to the data in the test set. The network with 600 hidden units produced the best overall correlation between desired and generated output for the test set. We therefore report the results with 600 hidden units.

When using the network outputs to drive the articulation of the synthetic face, we found the motion to be somewhat jerky due to instability in the output values. Empirically it was found that a simple post-hoc filtering, using a triangular averaging window with a width of 80 ms significantly reduced these disturbances without notably impairing the temporal resolution.

2.2 Results

The network trained on laser-man's isolated words showed fairly good learning and generalized fairly well to novel words. We computed a correlation between the target and learned parameter values across the complete training and test data sets. The overall average correlations between the target and learned parameter values were 0.77 for the training set and 0.64 for the test set.

2.3 Perceptual Evaluation

In order to evaluate the quality of the ANN versus TtS synthesized speech, a perceptual identification study was carried out with human participants. For this experiment, 131 short English words were tested, 65 of which had been used to train the network and 66 completely new words. Each of these 131 words was presented using ANN and text-to-speech (TtS) based synthesis, for a total of 262 trials per participant. Students from introductory psychology classes (5 male, 12 female, average age 18.7 years) with either normal or corrected vision served as subjects. All were native English speakers. The subjects were tested individually in sound attenuated rooms. On each trial of the experiment, a word was presented silently and then the subject typed in what word was presented. The size of the talking face was about 7 inches vertically viewed from about 12 inches. Only valid single syllable words were accepted as responses. If the typed word was not on the program's list of 11,744 words, the subject was cued to enter a new response. The next trial started 1 second after a response was entered. The experiment was presented in two sessions of about 20 minutes each.

The results of the experiment were scored in terms of the proportion of correct initial consonant, medial vowel, and final consonant, both for phonemes and visemes. Figure 3 shows the proportion correct initial consonant, vowel, and final consonant phonemes for the test words that did not occur in the training set. As can be seen in the figure, performance was well above chance for both conditions, but the TtS synthesis supported much better speechreading than the ANN synthesis. Figure 4 shows the corresponding viseme performance. Correct phoneme identification averaged 21% for TtS synthesis and 12% for ANN synthesis. Identification performance is, of course, much better when measured by viseme categories, as defined in previous research [6, Chapter 13]. Replicating the results at the phoneme level, performance given the ANN synthesis falls significantly below TtS synthesis. Overall, correct viseme identification was 72% for TtS synthesis and 46% for ANN synthesis. The discrepancy between the two presentation modes was largest for the vowels. At this time, we have no explanation for this difference between vowels and consonants.

3. EXPERIMENT 2: EXTEMPORANEOUS SPEECH

3.1 Method

Ten speakers from the CSLU stories database [http://cslu.cse.ogi.edu/corpora/stories/] were used to train ten different ANNs. The stories corpus is made up of extemporaneous speech collected from English speakers in the CSLU Multi-language Telephone Speech data collection. Each speaker was asked to speak on a topic of their choice for one minute. This database has been labeled and segmented so that the identity and duration of the spoken language segments are known.

The input data sets had approximately 50 seconds of natural speech; 40 seconds were used as training data for the networks. The remaining 10 seconds were used as a test set for the trained networks. The restricted amount of training data avaliable from each speaker makes this data set a hard test for the networks.

The training and generalization tests followed the same general procedure as with the isolated words. The networks were trained from 500 to 5000 epochs (passes through the data set) with momentum set to 0.0 and a learning rate of 0.1, .005 or .001. We experimentally determined that 100 hidden units were able to learn the mapping by training several networks with 10, 50 and 100 hidden units.

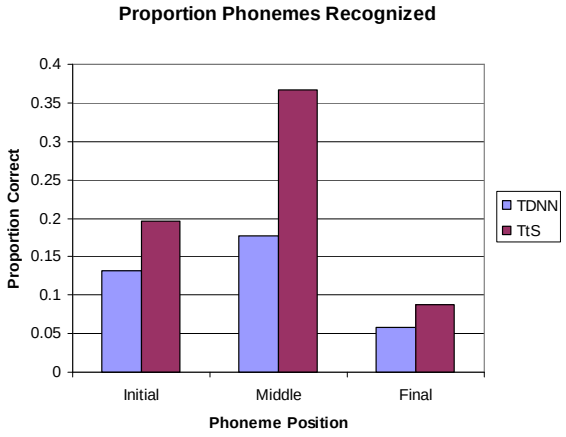

Figure 3. Proportion correct of initial consonant, vowel, and final consonant phoneme recognition for ANN and TtS synthesis.

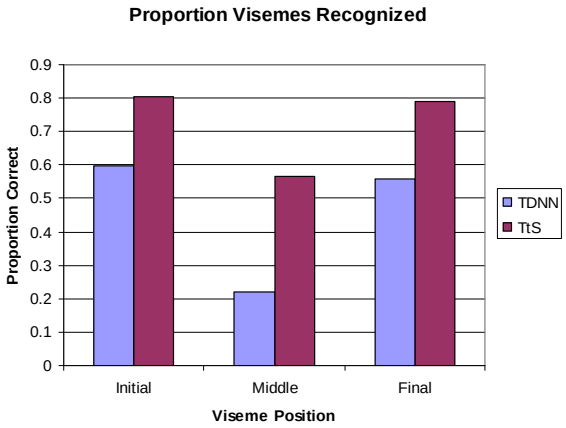

Figure 4. Proportion of initial consonant, vowel, and final consonant viseme recognition for ANN and TtS synthesis.

3.2 Results

The networks were evaluated using the root mean square (RMS) error over time and the correlation of each output parameter with the corresponding training values. The average correlation of all parameters was also used as an indicator of network performance. The networks varied somewhat in their abilities to reproduce the output parameters of each speaker (0.75 to 0.84 mean correlation across all parameters).

Each network was tested on novel speech from the speaker it was trained on. The average correlation over every parameter for each network was calculated on the corresponding test set for each network. The ability to generalize to novel speech varied across the 10 speakers. Speaker 8 and speaker 4 generalized to novel speech from their test set best with an average correlation of 0.27 and 0.28. We believe that these generalization values are low because of the paucity of training data and the restricted number of hidden units (100).

4. CONCLUSIONS

In a typical application, natural auditory speech can be used to generate an animated talking head that will be aligned perfectly with the natural auditory speech utterances as they are being said. This type of approach ideally allows for what is called graceful degradation. That is, the acoustic analysis is not dependent on a speech recognizer that could make catastrophic errors and therefore misguide the visible speed synthesis. The mapping between the acoustic parameters and the visible speech parameters is continuous and a slight error in the analysis of the input utterance will not be catastrophic because the parameters will still approximate the appropriate visible speech parameters for that utterance.

There are many potential applications for this technology primarily because bandwidth is highly limited in communication across the Internet and therefore video teleconferencing and other means of face-to-face communication are still very limited [7]. However auditory speech can be represented accurately with very little bandwidth requirements. The user could have the talking heads stored locally and controlled and animated locally by the auditory speech that is being streamed over the Internet.

This application could work for video teleconferencing as well as for email in that user could send an auditory message that would control the talking head located on the receiver's desktop. In addition the message could contain information about the sender and could either provide a texture map of the sender that would be mapped over the talking head on the receiver's computer or the appropriate texture could be stored permanently and retrieved on the receiving computer.

Currently, our system looks 5 frames or 100 ms ahead to generate the appropriate visible speech parameter values. In an actual application, it would, therefore, be necessary to delay the auditory speech by 100 ms. Another possibility is to train a network with fewer frames ahead of the current one. In either case, the network solution is preferable to any speech recognition systems that delay their decisions until at least several words have been presented.

5. ACKNOWLEDGEMENT

The research is supported by grants from PHS, NSF, Intel, the Digital Media Program of the University of California, and UCSC. Christopher Fry is now at Department of Psychology, University of California - Berkeley. The authors thank Chris Bregler, Bjorn Granström, and Malcolm Slaney for their comments on the paper.

6. REFERENCES

1. Agelfors, E., Beskow, J., Dahlquist, M., Granstrom, B., Lundeberg, M., Spens, K-E., Ohman, T (1998): "Synthetic Faces as a Lipreading Support", in Proceedings of *ICSLP'98, International Conference on Spoken Language Processing*, November 1998, Sydney, Australia.

2. Arslan, L.M. & Talkin, D. (1998) 3-D Face Point Trajectory Synthesis using an Automatically Derived Visual Phoneme Similarity Matrix. In D. Burnham, J. Robert-Ribes & E. Vatikiotis-Bateson (Eds.) *Proceedings of the International Conference on Auditory-Visual Speech Processing*, December 1998, Terrigal, Australia.

3. Benoit, C., & Le Goff, B. (1998). Audio-visaul speech synthesis from French text: Eight years of models, designs and evaluation at the ICP. *Speech Communication, 26,* 117-129.

4. Bernstein, L.E. & Eberhardt, S.P. (1986). *Johns Hopkins lipreading corpus videodisk set.* Baltimore, MD: The Johns Hopkins University.

5. Cohen, M. M., & Massaro, D. W. (1993) Modeling coarticulation in synthetic visual speech. In N.M Thalmann and D. Thalmann (Eds.) *Models and Techniques in Computer Animation*. Tokyo: Springer-Verlag.

6. Cohen, M. M., Walker, R. L., & Massaro, D. W. (1995) Perception of Synthetic Visual Speech, *Speechreading by Man and Machine: Models, Systems and Applications,* NATO Advanced Study Institute 940584, Aug 28-Sep 8, 1995, Chateau de Bonas, France.

7. Cromarty, A. (1999). Keynote address: Internetworked Multimedia and Entertainment Distribution. EnterTech Conference, April 25-27, La Costa Resort and Spa.

8. Goldenthal, W., Waters, K., Van Thong, J-M., & Glickman, O. (1997): "Driving Synthetic Mouth Gestures: Phonetic Recognition for FaceMe!", in proceedings of EUROSPEECH'97, September 1997, Rhodes, Greece

9. Massaro, D. W. (1998) *Perceiving Talking Faces: From Speech Perception to a Behavioral Principle*. Cambridge, Mass.: MIT Press.

10. Massaro, D. W., & Stork, D. G. (1998) Speech Recognition and Sensory Integration. *American Scientist, 86*.

11. Morishima, S. (1998) Real-time Talking Head Driven by Voice and its Application to Communication and Entertainment. In D. Burnham, J. Robert-Ribes & E. Vatikiotis-Bateson (Eds.) *Proceedings of the International Conference on Auditory-Visual Speech Processing*, December 1998, Terrigal, Australia.

12. Reveret, L. & Benoit, C. (1998). A new 3D lip model for analysis and synthesis of lip motion in speech production. In D. Burnham, J. Robert-Ribes & E. Vatikiotis-Bateson (Eds.) *Proceedings of the International Conference on Auditory-Visual Speech Processing*, December 1998, Terrigal, Australia.

13. Tamura, M., Masuko, T., Kobayashi, T. & Tokuda, K. (1998) Visual speech synthesis based on parameter generation from HMM: Speech driven and text-and-speech driven approaches. In D. Burnham, J. Robert-Ribes & E. Vatikiotis-Bateson (Eds.) *Proceedings of the International Conference on Auditory-Visual Speech Processing*, December 1998, Terrigal, Australia.

14. The Need for Increased Speech Synthesis Research (1999) Report of the 1998 NSF Workshop for Discussing Research Priorities and Evaluation Strategies in Speech Synthesis. Editors: R. Sproat, M. Ostendorf, and A. Hunt. March, 1999.

15. Wightman, C. W. and Talkin, D. T., (1997). "The Aligner: Text-to-Speech Alignment Using Markov Models", in Progress in Speech Synthesis, J. P. van Santen, et al, (Eds.). Springer Verlag, New York: (pp. 313-323).

16. Yamamoto E., Nakamura, S., & Shikano, K. (1998). Lip movement synthesis from speech based on Hidden Markov Models. *Speech Communication, 26,* 105-115