What can Visual Speech Synthesis tell Visual Speech Recognition?

Michael M. Cohen and Dominic W. Massaro
{mmcohen,massaro}@fuzzy.ucsc.edu

University of California, Santa Cruz
Santa Cruz, CA 95064

Abstract

We consider the problem of speech recognition given visual and auditory information, and discuss some of the ways that speech synthesis can provide assistance. Three possible contributions of synthetic visual speech are discussed: First, the use of synthetic speech to study human speech perception, second, the use of speech synthesis techniques to instantiate models of human speech production, and third, the use of these production models to help guide automatic speech recognition. Finally, we consider the reverse relationship: How can the techniques of automatic speech recognition assist in better visual speech synthesis?

1. Speech perception by human and machine

Speech perception is a human skill that rivals our other impressive achievements. Even after decades of intense effort, speech recognition by machine remains far inferior to human performance. Although some engineering approaches to speech recognition may take advantage of distinctly non-human methods (e.g. the Blackboard model in HEARSAY which does not take into account human memory constraints), we believe that an understanding of human performance is highly valuable for any machine solution. Humans have solved the problem of speech recognition; understanding these solutions would necessarily be valuable for improving machine recognition. Our understanding of the human solution has been helped immensely through the use of speech synthesis. We begin by briefly considering some aspects of human speech recognition research, especially concentrating on bimodal perception.

Experiments have revealed conclusively that our perception and understanding are influenced by the visible speech in the speaker's face and the accompanying gestural actions. These experiments have shown that the speaker's face is particularly helpful when the auditory speech is degraded due to noise, bandwidth filtering, or hearing-impairment [20]. Although the influence of visible speech is substantial when auditory speech is degraded, visible speech also contributes to performance when paired with intelligible speech sounds. Although visible speech has a strong influence in speech perception, it does not provide all of the segmental distinctions provided by auditory speech. The number of distinctive visible categories, called visemes, are fewer than the number of audible categories or phonemes.

One might think that visible speech is simply redundant with auditory speech. However, research has shown that human speech perception is robust because perceivers use multiple sources of information, not just a single source or modality. It follows that disruption of a given source will not dramatically disrupt speech perception. One additional attractive aspect of having both visible and auditory speech is the complementarity of the two channels. Visible speech is usually most informative for just those distinctions that are most ambiguous auditorily. For example, perceiving place of articulation (such as the difference between /b/ and /d/) is difficult via sound but relatively easy via sight. Voicing information, on the other hand, is difficult to see visually but is easy to resolve via sound. Thus, audible and visible speech not only provide two independent sources of information, these two sources are often productively complementary. Each is strong when the other is weak.

Recent research has also conclusively demonstrated that perceivers have continuous information from the speech cues, not just categorical information. Although one's phenomenal experience in speech perception is usually that of perceiving categories, the processing system is not limited to categorical information. Many empirical investigations have now demonstrated that perceivers are capable of perceiving differences within a speech category. When asked appropriately, people can reliably indicate the degree to which different speech tokens represent a given speech category [20, 21]. In addition, ambiguous tokens require more time for categorization than do clear tokens. These results indicate that people can discriminate differences within a speech category and they are not limited to just categorical information. The richness of the representation of a speech token is not obscured during speech perception, but retains its graded composite of information. Most likely because of the discrete structure of human communication via spoken language, however, the decision process simply maps the rich continuous information into one of the discrete categories used in our language. We note that for ASR also, there are advantages to using continuous as opposed to discrete information [5, 7].

Finally, research has demonstrated that humans naturally integrate multiple sources of information in speech perception. Within the framework of the Fuzzy Logical Model of Perception, perceptual events are processed via three operations: feature evaluation, feature integration, and decision. The sensory systems transduce the physical event and make available various sources of information called features. These continuously-valued features are evaluated, integrated, and matched against prototype descriptions in memory, and an identification decision is made on the basis of the relative goodness-of-match of the stimulus information with the relevant prototype descriptions.

2. Synthetic speech

A critical assumption of human speech perception research concerns the theoretical, experimental, and applied value of synthetic speech. Auditory synthetic speech has proven to be valuable in all three of these domains. Synthetic speech has made both theoretical and applied contributions to analysis of speech perception. It gives the experimenter control over the stimulus in a way that is not always possible using natural speech and permits the implementation and perceptual test of theoretical hypotheses, such as a) which cues are critical for various speech distinctions and b) how these cues are integrated. Synthetic speech also allows us to evaluate the adequacy of our models of human speech production. Speech production models serve to formalize a theoretical overview as well as to focus experimental research activities [27]. Eventually, such models should incorporate everything from anatomy, biomechanics, and aerodynamics, to phonology, syntax, and semantics. Finally, the applied value of auditory synthetic speech is apparent in the multiple everyday uses for text-to-speech systems for both normal and visually-impaired individuals.

We believe that visible synthetic speech will have the same value as audible synthetic speech. Its use can provide a more fine-grained assessment of psychophysical and psychological questions not possible with natural speech. For example, testing people with synthesized syllables intermediate between several alternatives gives a more powerful measure of the functional value of visible cues and how these cues are integrated with auditory information. These two questions cannot be answered as easily, if at all, with unaltered natural speech stimuli. Analogous to the valuable contribution of using auditory speech synthesis in speech perception research, visible speech synthesis permits the type of experimentation necessary to determine 1) what properties of visible speech are used, 2) how they are processed, and 3) how this information is integrated with auditory information and other contextual sources of information in speech perception.

Given the value of synthetic speech, it is worth considering some general requirements for good speech synthesis. There are basically two necessary ingredients [30]. The first is to have a highly-specified low-level model for the human speech production apparatus. In the acoustic modality, Fant’s [10] source-filter theory of speech production has provided a model, which has been instantiated for synthesis electronically [11] and in software [15]. The second necessary ingredient for speech synthesis is to have a good higher level model of the transformation from linguistic information to the control parameters for for the production apparatus [16]. Important developments in this area were made in the MITalk project [1]. Among the important higher level phenomena to be modeled are segmental timing changes dependent on phonetic and syntactic structure and coarticulation. Coarticulation refers to changes in the articulation of a speech segment depending on preceding (backward coarticulation) and upcoming segments (forward coarticulation). An example of backward coarticulation is the difference in articulation of a final consonant in a word depending on the preceding vowel, e.g. "boot" versus "beet". An example of forward coarticulation is the anticipatory lip rounding at the beginning of the word "stew". The degree to which both our low and high level models accurately capture the nature of human speech production will be reflected in the quality of our synthetic speech production. We should keep these criteria in mind during our discussion of visual speech synthesis.

3. Approaches to visual speech synthesis

Some early work by speech researchers [9] used relatively simple Lissajou’s figures displayed on an oscilloscope to simulate lip movement. Later, a model for lip shape [22] using about 130 vectors was developed which allowed computation of coarticulatory effects for short segments. More recent researchers working in computer graphics have used 3-dimensional facial models (cued by lighting, shading, and in some cases texture).

Two general strategies for generating highly realistic full facial displays have been employed: parametrically controlled polygon topology and musculoskeletal models. Using the first strategy, Parke [25] developed a fairly realistic animation by modeling the facial surface as a polyhedral object composed of about 900 small surfaces arranged in 3D, joined together at the edges, and smooth shaded. The face was animated by altering the location of various points in the grid under the control of 50 parameters, about 10 of which were used for speech animation. The control parameters used for several demonstration sentences were estimated by studying articulation frame by frame.

Parke's software and topology was given new speech and expression control software [26] in which a user could type a string of phonemes which were then converted to control parameters which were changed over time to produce the desired animation sequence. Each phoneme was defined in a table according to values for segment duration, segment type (stop, vowel, liquid, etc) and 11 control parameters such as jaw rotation, mouth width, lip protrusion, lower lip "f" tuck, etc. The program made a transition between two phonemes by interpolating in a nonlinear fashion between the values for two adjacent phonemes. Different transition speeds were used depending on the type of segments involved. Our current software [6] is a descendant of the Parke software, incorporating additional control parameters, a tongue, and a new speech synthesis control strategy. An important improvement in our visual speech synthesis software has been the development of an new algorithm for articulator control which takes coarticulation into account. Our approach to the synthesis of coarticulated speech is based on an articulatory gesture model [19]. In this model, the speech segment has dominance over the vocal articulators which increases and then decreases over time during articulation. Adjacent segments have overlapping dominance functions which leads to a blending over time of the articulatory commands related to these segments. We have instantiated this model in our synthesis algorithm using negative exponential functions for dominance. Given that articulation of a segment is implemented by several articulators, there is a separate dominance function for each articulator. The different articulatory dominance functions can differ in time offset, duration, and magnitude.

An example of the system's operation is shown in the top panel of Figure 1, illustrating the lip protrusion dominance functions for the word "stew". As can be seen, the /s/ and /t/ segments have very low dominance with respect to lip protrusion compared to /u/. Also the dominance of /u/ extends far forward in time. The lower panel gives the resulting lip protrusion. One can see how the lip protrusion extends forward in time from the vowel. Note that the figure only illustrates the dynamics for lip protrusion. Other control parameters, e.g. tongue angle, for /t/ and /u/ have equal dominance. This allows the tongue to reach its proper location against the back of the upper teeth for /t/. As part of our higher level control strategy we have integrated the MITalk system [1] to provide the segments, durations, and suprasegmental information to the visual

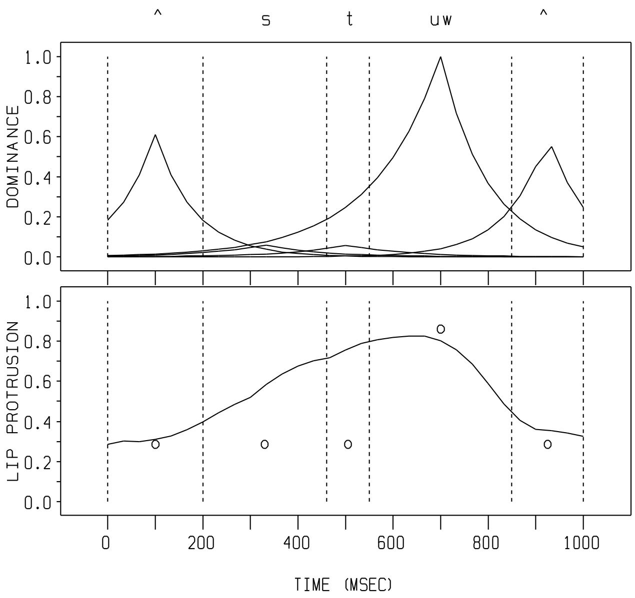

Fig. 1. Dominance functions (top panel) and parameter control functions (bottom panel) for lip protrusion for the word "stew". The circles in the bottom panel indicate the control parameter targets for each segment.

synthesis algorithms, and also to provide the auditory speech which is played in synchrony with the visual speech.

Using the second basic strategy, human faces have been made by constructing a computational model for the muscle and bone structures of the face [28, 32, 33]. At the foundation of the model is an approximation of the skull and jaw including the jaw pivot. Muscle tissues and their insertions are placed over the skull. This requires complex elastic models for the compressible tissues. A covering surface layer changes according to the underlying structures. The driving information for such a model can be defined by a dynamically changing set of contraction-relaxation muscle commands, often controlled using the "Facial Action Coding System" [8]. These codes are based on about 50 facial actions defined by combinations of facial muscle actions.

One drawback to this synthesis approach is that calculations needed for the tissue simulations take significantly longer to carry out than the calculations of the changing surface shapes in the polygon models. It also may be more difficult to achieve the desired articulations in by varying the constituent muscle actions as opposed to varying the desired shapes directly. This difference in synthesis methods is parallel to the difference between articulatory [12] and terminal-analog [13] synthesizers for auditory speech. As with visual speech, auditory articulatory synthesizers require much more computation.

Although still in its early stages, the development of realistic, high-quality, facial displays provide a powerful tool for investigation of a number of questions in auditory-visual speech perception. The analysis of the articulation of real speakers guides the development of the visible speech synthesis. In addition, perception experiments can tell us how well the synthesis simulates real speakers. The results of this research can be used to implement automatic lipreading to enhance speech recognition by machine. Just as human perceivers achieved robust recognition of speech by using multiple sources of information, the same could be true for machine recognition.

4. Using speech synthesis to aid speech recognition

As we have discussed, our speech synthesis techniques should incorporate all of our knowledge about the human speech production. The question is how to apply that knowledge, indirectly and directly to the problem of speech recognition.

Let us first consider the indirect contributions from the study of human perception using synthetic speech. As mentioned above, our perceptual research suggests that we integrate continuous valued multi-modal feature information to achieve optimal recognition performance.

Both human perceptual research as well as our synthesis algorithms tell us that coarticulatory and other phenomena occur over a supra-phoneme span [6]. This fact alone leads to a recommendation for an analysis window which takes into account feature information over several phonemes.

We can also synthesize synthetic speech to test our speech recognition systems. Synthetic speech, both unimodal and multimodal can also be used as inputs to validate recognition techniques. Precisely controlled stimuli allow a fine-grained assessment of system operation.

A somewhat more direct use for speech synthesis is illustrated in the ANGEL ASR system [7]. In this system, the lexicon consists of word pronunciations stored in the form of phonetic networks. To create these networks, baseform pronunciations were first derived using MITalk letter-to-sound rules or transcriptions. These baseforms were then transformed using more than 250 rules for word boundary effects, followed by the folding of individual words into a general graph structure. This use of speech synthesis greatly simplified the adapting the lexicon for new tasks with fairly large vocabularies.

In terms of direct application of synthesis to recognition, we return first to one of the earliest models: Analysis-by-Synthesis. This model was formulated as part of the original motor theory of speech perception [29]. The basic idea of this theory was that we understand speech by recovering the articulatory information from the acoustic signal. It was believed that the problem of mapping the acoustics to articulation was too difficult due to contextual effects. The idea of analysis-by-synthesis method [2, 31] could be used to synthesize and verify some particular hypothesis. Some typical later works on this idea are given in perception of transitions [24] and sentence recognition [4]. However, there are two major problems with this approach [17]. First, there is no satisfactory solution to how one comes up with initial hypotheses to test (but see [35]), and second, the cognitive load required by the model is too high [23]. We would also point out that the revised version of the motor theory [18] obviates the need for analysis-by-synthesis by the assumption of direct perception through a special-purpose module which directly transforms acoustic information to the intended underlying articulatory gestures. But this idea has other problems [20] including the fact that any particular acoustic signal could be produced from a variety of vocal tract shapes. Another criticism is that mediation of auditory speech perception by articulatory gestures is probably unnecessary [20]. It is interesting to note that in at least one scheme for recognition from bimodal speech [34], the visual information is translated to an auditory space before combination. All that being said, we do not argue that the articulatory information is necessarily uninformative, and in fact with visual speech, it may be easier to obtain. It remains an open question, however, whether analysis-by-synthesis will play a valuable role somewhere down the line.

How should we make use of the production models as represented in our speech synthesis techniques. We think that it would be valuable to take this information into account somehow, given the current weakness of some current approaches dealing with contextual information and the temporal character of speech. Hidden Markov Models (HMMs) for example, commonly assume that the probability that a certain feature vector will occur depends solely on the current state probability, thus disallowing the influence of contextual information.

One approach which may hold promise for integrating production knowledge is the idea of constraint surface learning [3]. In this approach, the possible feature values (e.g. representing lip shape) form a surface in a multidimensional space. Given some input feature vector which may contain noise, a revised feature vector is formed which is the closest projection from the observed vector to the surface. This is then used in a hybrid connectionist recognition system. Because the temporal resolution of acoustic information in a bimodal speech recognition situation is considerably faster than the visual information (100 frames/second vs 30 frames/sec), it may prove useful to create interpolated visual feature vectors for combination with the auditory. However, rather than using

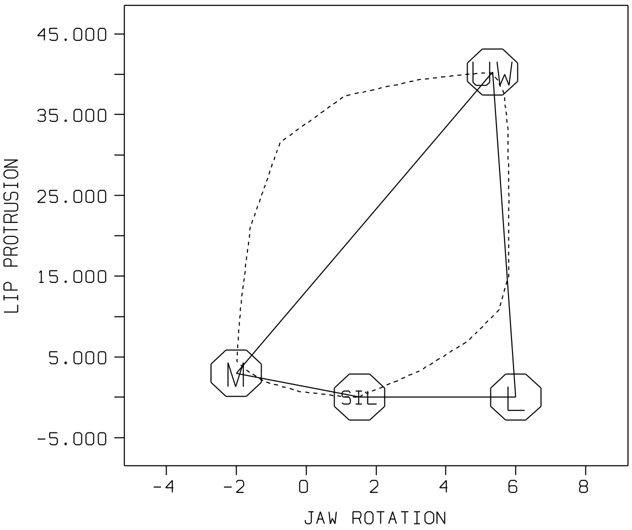

Fig. 2. Jaw rotation and lip-protrusion control parameter trajectories for the word "loom". The solid lines give the direct path between phonemes and the dashed lines give the coarticulated path.

linear interpolation, it may be better to interpolate along the surface. We are collaborating with Bregler and his colleagues in this investigation, with validation by high speed image capture.

We believe that this general approach might similarly be extended by using production information to constrain possible trajectories on the surface given contextual information. For our visual speech synthesis algorithm [6], Figure 2 illustrates the difference between linear and coarticulated control parameter trajectories. In this simple example for the word "loom", we trace the path in jaw-rotation / lip-protrusion space starting (and ending) at the silent (SIL) position, going through the phonemes /l/, /U/, and /m/. As can be seen, the coarticulated trajectory follows a much different path than the direct linear one. For example, lip-protrusion never reaches it's target for /l/. While this plot shows the trajectories in control-parameter space, using the synthesis software we can also easily generate feature information corresponding to the lip-shape measurements used by Bregler and his colleagues. We are currently evaluating methods of using such trajectory constraints.

In addition to using constraint surface trajectories directly, synthesis might also be used in an indirect fashion similar to its use in the ANGEL system. That is to say, the synthesis could be used to generate feature-vector data for learning the constraint surface. For example, by modeling different talkers in our synthesis, the recognition system could more easily adapt to speaker variability.

5. From recognition to synthesis

In this section we briefly consider the reverse relationship between synthesis and recognition: How can speech recognition techniques aid the development of speech synthesis.

The first contribution that is apparent is the use of automatic lip tracking techniques for refining our synthesis parameters, both in terms of obtaining segment target values and control parameter dynamics, which in turn allow us to evaluate the accuracy of our synthesis algorithms and their underlying production models.

A second contribution of recognition techniques is the assessment of which cues provide optimal recognition. Finn [12] used an algorithm to obtain the best weighting of facial features for recognition. A similar examination of various features using a principal components analysis and then selecting those with the highest eigenvalue has been employed [14] to pick which features would be used for recognition. Although these results do not necessarily mean that the same features are equally important to human observers, the results are suggestive (and reinforced by the fact that the recognition algorithms yield similar results to the humans). By assessing the accuracy of recognition and the nature of the confusion errors, the commonalities of human and machine recognition can be determined.

Acknowledgment

The writing of this paper was supported, in part, by Public Health Service Grant R01 DC 00236 to Dominic W. Massaro.

References

[1] J. Allen, M.S. Hunnicutt & D.H. Klatt. *From text to speech: The MITalk system* Cambridge, MA: Cambridge University Press, 1987.

[2] C.G. Bell, H. Fujisaki, J.M. Heinz & K.N. Stevens. Reduction of speech spectra by analysis-by-synthesis techniques. *Journal of the Acoustical Society of America, 33*, S15. 1725-1736, 1961.

[3] C. Bregler, S. Omohundro, Y. Konig & N. Morgan. Using surface-learning with to improve speech recognition with lipreading *Proceedings of the 1994 Asilomar Conference on Signals, Systems and Computers*.

[4] J.S. Bridle & M.P. Ralls. An approach to speech recognition using synthesis by rule. In F. Fallside and W.A. Woods (Eds.) *Computer Speech Processing* London: Prentice/Hall International. 277-292, 1985.

[5] H.A. Bourland & N. Morgan. *Connectionist Speech Recognition: A Hybrid Approach* Boston: Kluwer Academic Publishers, 1994.

[6] M.M. Cohen & D.W. Massaro. *Modeling coarticulation in synthetic visual speech.* In N. . Thalmann & D.

Thalmann (Eds.) *Models and techniques in computer animation*, Tokyo: Springer-Verlag, 139-156, 1993.

[7] J.S. Denton & C.R. Taylor. *Final Report on Speech Recognition Research, December 1984 to April 1990* Carnegie Mellon University: School of Computer Science, 1990.

[8] P. Eckman & W.V. Friesen. *Manual for the Facial Action Coding System* Palo Alto: Consulting Psychologists Press, 1977.

[9] N.P. Erber & C.L. De Filippo. Voice-mouth synthesis of /pa, ba, ma/. *Journal of the Acoustical Society of America, 64,* 1015-1019, 1978.

[10] G. Fant. *The Acoustic Theory of Speech Production*. The Hague: Mouton & Co, 1960.

[11] G. Fant & J. Martony. Instrumentation for parametric synthesis (OVE-II). *KTH Speech Transmission Laboratory: Quarterly Progress and Status Report, 2/3*, 18-19, 1962.

[12] K.E. Finn. *An Investigation of Visible Lip Information to be Used in Automated Speech Recognition* Ph.D. thesis, Georgetown University, 1986.

[13] J.L. Flanagan, K. Ishizaka & K.L. Shipley. Synthesis of speech from a dynamic model of the vocal cords and vocal tract. *Bell System Technology Journal, 54*, 485-506, 1975.

[14] A.J. Goldschen. *Continuous Automatic Speech Recognition by Lipreading,* Ph.D. thesis, George Washington University, 1993.

[15] D.H. Klatt. Software for a cascade/parallel formant synthesizer. *Journal of the Acoustical Society of America, 67,* 971-995, 1980

[16] D.H. Klatt. Review of text-to-speech conversion for english. *Journal of the Acoustical Society of America, 82,* 737-793, 1987.

[17] D.H. Klatt. Review of selected models of speech perception. In W. Marslen-Wilson (Ed.) *Lexical Representation and Process,* Cambridge, MA: MIT Press, 169-226, 1989.

[18] A.M. Liberman & I.G. Mattingly. The motor theory of speech perception revised. *Cognition, 21*, 1-36, 1986.

[19] A. Löfqvist. Speech as audible gestures. In W.J. Hardcastle and A. Marchal (Eds.) *Speech Production and Speech Modeling*. Dordrecht: Kluwer Academic Publishers, 289-322, 1990.

[20] D.W. Massaro. *Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry*, Hillsdale, NJ: Lawrence Erlbaum Associates, 1987.

[21] D.W. Massaro & M.M. Cohen. Categorical or continuous speech perception: A new test. *Speech Communication, 2*, 15-35, 1983.

[22] A.A. Montgomery. Development of a model for generating synthetic animated lip shapes. *Journal of the Acoustical Society of America, 68,* S58 (abstract), 1980.

[23] A. Newell. Harpy, production systems, and human cognition. In R. Cole (Ed.) *Perception and Production of Fluent Speech*, Hillsdale, NJ: Lawrence Earlbaum Associates, 1979.

[24] K.K. Paliwal & P.S.V. Rao. Synthesis-based recognition of continuous speech. *Journal of the Acoustical Society of America, 71,* 1016-1024, 1982.

[25] F.I. Parke. A parametric model for human faces, *Tech. Report UTEC-CSc-75-047* Salt Lake City: University of Utah, 1974.

[26] A. Pearce, B. Wyvill, G. Wyvill, & D. Hill. Speech and expression: A computer solution to face animation. *Graphics Interface '86*, 1986.

[27] J. Perkell. Models, theory, and data in speech production. *Proc. of the 12th Int. Congress of Phonetic Sciences*, Aix-en-Provence, France, 19-24 August, 1991.

[28] S.M. Platt & N.I. Badler. Animating Facial Expressions. *Computer Graphics, 15(3),* 245-252, 1981.

[29] K.N. Stevens. Toward a model for speech recognition. *Journal of the Acoustical Society of America, 32*, S15. 47-55, 1960.

[30] K.N. Stevens. Speech synthesis methods: Homage to Dennis Klatt. In G. Bailly, C. Benoit, & T.R. Sawallis, (Eds.) *Talking machines: Theories, Models, and Designs* Amsterdam: North Holland, 485-504, 1992.

[31] K.N. Stevens & M. Halle. Remarks on analysis by synthesis and distinctive features. In W. Wathen-Dunn (Ed.) *Proceedings of the AFCRL Symposium on Models for the Perception of Speech and Visual Form*, Cambridge, MA: MIT Press, 1964.

[32] D. Terzopoulous & K. Waters. Techniques for realistic facial modeling and animation. in N. M. Thalmann and D. Thalmann (Eds.) *Computer Animation '91* Tokyo: Springer-Verlag, 1991.

[33] K. Waters & D. Terzopoulous. A physical model of facial tissue and muscle articulation. *SIGGRAPH Facial Animation Course Notes*, 130-145, 1990.

[34] B.P. Yuhas, M.H. Goldstein, & T.J. Sejnowski. Integration of acoustic and visual speech signals using neural networks, *IEEE Communications Magazine,* 65 - 71, 1989.

[35] V. Zue. Models of speech recognition III: The role of analysis by synthesis in phonetic recognition. In P. Mermelstein (Ed.) *Proceedings of the Montreal Satellite Symposium on Speech Recognition (Twelfth International Conference on Acoustics),* 1986.