Challenges in the Fusion of Video and Audio for Robust Speech Recognition

Jer-Sen Chen and Oscar N. Garcia
Department of Computer Science and Engineering
Wright State University
Dayton, OH 45435
jschen@cs.wright.edu
ogarcia@cs.wright.edu

Abstract

As speech recognizers become more robust, they are popularly accepted as an essential component of human-computer interaction. State-of-the-art speaker-independent speech recognizers exist with word recognition error rates below 10%. To achieve even higher and robust recognition performance, multi-modal speech recognition techniques that combine video and audio information can be used. Speech reading, the video portion of bimodal speech recognizer, introduces not only additional computataional cost of video processing, but also chanllenges in the design of the integrated audio-video recognizer.

Introduction

We have produced speech-reading (more commonly called lipreading) recognition by computer of the video taped face of a single speaker that realized 25% continuous speech recognition with a perplexity of 150 based on video alone, without audio or a grammar [Gold 93]. While this may not seem very high at first, even though this was our first experimental system, it should be compared with the absolute best human lipreading performance of some 35% on context-free nonsense words. The significance of the contribution of speech-reading to the reduction of errors in recognition is understood when the strong complementarity between the two modalities comes to play in noisy or stressed environments.

In fusing the video and audio speech information there is the problem of synchronization and of trying to achieve the same or analogous spatio-temporal granularity for the sampling of the two signals. The acoustic signal is ordinarily sampled at 8 or 16 kHz, and windowed at 10 to 20 ms frames while the video signal is captured at some 640Hx480V pixels at 30 fps or 640Hx240V pixels at 60 fps. Matters are further complicated if color is introduced with 3-CCD arrays as is desirable when we try to track the movement of the lips in contrast with the skin and other mouth features. The capture of a video frame that has the desired lipreading information is difficult if not impossible depending on the circumstances.

Visible Articulators in Speech Production

Given a vocal tract with fixed dimensions the most important articulators, after the glottal pulses are produced, are the tongue and lips. The tongue is only visible for some phones, most dramatically in labidentals, and the lips are most easily discriminated, in general, when they open or close to articulate. With a zoom lens camera or a camera positioned under the chin in a harness or on a desk top, it is possible to locate the oral cavity, most easily by finding the nostrils and safely assuming that it is located in the area under the nose and normalized for scale and rotation [Peta 96]. In addition there are variations in pose, lighting, skin and lip color, and facial hair. Acquiring the lip features from a broadcast television signal is hampered by the fact that even if the speaker is looking straight into the camera, as most news commentators are prone to do, the head moves and the mouth area constitutes a small percentage (less than 5% of the whole picture, less than 10% in the horizontal dimension) of a frame. However, if a specially placed camera(s) and microphone(s) were used for automatic close-caption during news broadcast, it is very likely that better synchronization between speech and text in speech-to-text translation (recognition) would be accomplished than what currently takes place. A picture of the mouth, however small, can give a good indication of the time proximity of one word to another and of the start and stop of the speech when the narrator is visible. Presence of teeth and tongue are also good phonetic discriminators of the manner and place of the articulation taking place.

Physically the lips act as radiators which, like an antenna, match the acoustic impedance of the mouth to

that of the external medium. Similarly, for the articulation of the nasal phones, the nasal passage shunts the radiating path of the mouth to which the tongue (for the alveolar "n"as in "none") or the lips (in the labial "m") or the velum (in the velar as in "bang") contributes. The lips, therefore, shape the effective mass of vibrating air as it moves outside the mouth, acting like a high pass RL filter with a cut off frequency of some 6 kHz.

We have investigated the visual features of the oral cavity [Garc 92, Gold 96] which contribute most to machine speech discrimination, through an unorthodox feature reduction process after a classical process of principal component analysis. In the analysis we selected those features whose eigenvalues covered 90% of the cumulative proportion measure, requiring sixteen features. Furthermore, we considered the cross-correlation of the features and finally reduced them to 13, out of the 35 features that were originally postulated for the analysis. No feature with correlation greater than 0.6 to another in a group was considered independently. This allowed the surjection of 56 phonemes (including closures) into 35 visemes (a viseme is the visual equivalent of a phoneme.) Just as in the case of the phonemes, the features required for the visemes included first and second derivatives, but we believe more prominently so, although to elicit this point a similar analysis with similar criteria would have to be carried out with the acoustic data. Since the accuracy of the derivatives depends on the frame rate, it seems that the difficult-to-achieve, degraded-resolution rate of 60 frames may be most desirable for estimation of the derivatives. The issue in question is to decide: what is worse a 30 fps rate at VGA resolution or 60 fps at half of the vertical VGA resolution? Further considerations in the process of recognition determined that there is as much coarticulation with the visemes as there was with the phones, and that triples of visemes which we called trisemes, had to be used as the units of visual speech. This places serious demands on the amount of data to be able to provide samples of 35x35x35= 42,875 triplets of visemes, assuming that there is no equivalent phonotactic reduction. While that assumption is likely to be wrong and the number of triplets may be smaller (we found that there were 2,702 triplets in 300 phonetically balanced TIMIT sentences, which might indicate strong phonotactic constraints.) Another challenge is, therefore, to have a substantially large database to provide sufficient training samples for reliable results with a learning Welsh-Baum (HMM) algorithm.

The features that we considered did not include puckering, although it is known experimentally that such phenomenon, called lateral protrusion from two orthogonal views, is important in English [Wozn 79] and in French [Cath 96]. What is not known, because we have been the only ones to experimentally map the phones into visemes in continuous speech, is whether such feature would further discriminate among our 35 visemes. We suspect that it would, and such would require more than the two-dimensional view obtained with only one camera. A system of two or more "genlocked" cameras of the proper frame rate for acquisition. A rate of 60 fps also requires a large amount of storage, as well as some fast video acquisition technology, such as the "Firewire" [Fire 96] approach. The complexity of video processing with multiple cameras, be it color or monochrome, presents another challenge for real time application. Algorithmic simplification for extraction of relevant video features for bimodal speech recognition will therefore be vigorously studied.

Architecture of Bimodal Speech Recognition Systems

Figures 1 and 2 show alternative models of bimodal human-to-machine speech recognition architectures. The alternative models differ in the point at which fusion of the video and acoustic information takes place. In these figures the block includes the classifier, which means that the code vectors in Figure 2 bring together into the Hidden Markov Model (HMM) both the acoustic and visual information together. Consideration of where and how such fusion takes place in the human speech recognition system is an area of clear interest for cognitive scientists without a definitive answer yet, although some strong theories such as the "fuzzy logical model" of Massaro and Cohen ([Mass89], [Mass94], [Mass96]) which seems supported by the work of others. These studies [Move 96] suggest that conditionally independent modality recognizers in humans which follow Morton's law are marginally superior in independent digit recognition. Many of these studies, however, have been carried out with small datasets and with tasks not involving very complex linguistic understanding tasks. However, it is an interesting issue, assuming that natural evolution reaches for optimal solutions, to resolve whether the same situation is true for machines, in particular, given the complementarity of visual (viseme) and corresponding acoustic (phoneme) signal detection by humans, as noticed long ago by Summerfield [Summ87] and by others more recently for machine perception ([Sils 93], [Stor 92]). We will investigate both architectures in the machine speech recognition domain and compare their performance. Preliminary work in this area at the ICP in Grenoble by the research group of Christian Benoit [Adjo 95]

with a single speaker uttering 54 nonsense words and with French vowels [Robe 96] seems to favor the "separate" identification method when normalized weights for the output probabilities are used in noisy environments. By contrast, the direct identification model significantly improves the audio recognition scores in all but highly degraded audio conditions, where the bimodal scores drop below the visual scores alone.

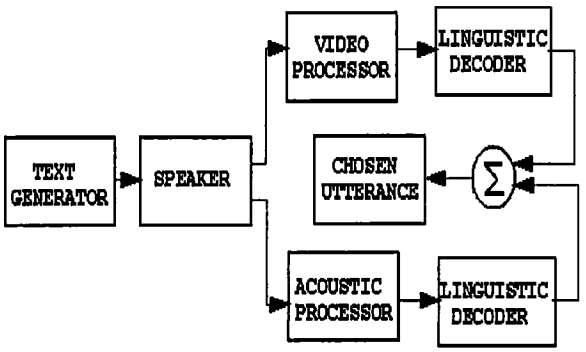

Figure 1. "Separate identification" or "late integration" model

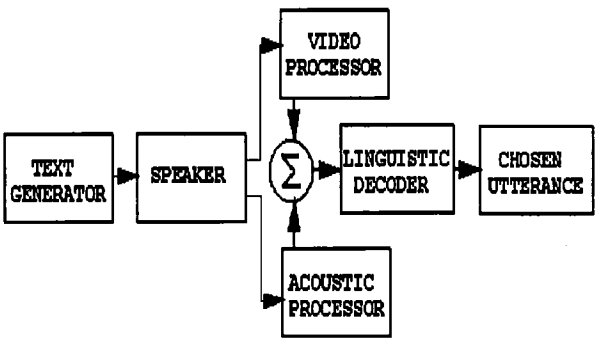

Figure 2. "Direct identification" or "early integration" model

Early work by Silsbee [Sils 95] with 22 consonants, reported in [Asil 94], was improved in [Stork 96] with a reduction of 25% in the error rate when a weight of visual vs. acoustic parameters is adaptively determined at a pre-categorical stage and then integrated in the HMM. Research on the issue of integration becomes even more relevant in recognition robustness involving multiple continuous speech samples from different speakers in noisy environments. The idea of automatic adaptation to environmental conditions is certainly worthy of continued investigation as noticed by Meir and his colleagues [Meie 96], who used it in conjunction with time-delayed neural networks. We have also learned informally [Vati 96] that there is electromyographic evidence of substantive differences between the nonsense words muscular activation and the meaningful sentence utterances, and it is not certain to what extent they affect the results obtained by researchers using nonsense words. As we can see, this topic is far from having been definitively settled.

A Light at the End of the Tunnel

With such challenges it would seem very difficult to make progress in actually combining the bimodal recognition into a manageable system. We claim it is not so. Actually the modalities complement each other, and a microphone array combined with multiple view cameras can certainly provide a system which may rely on weighted factors for use of information of each of the two modes as signal to noise rates demand the greater reliability of one or the other.

A very controversial and yet unsettled research issue, that of what is the best place for bimodal machine speech recognition for fusion of the visual and acoustic information channels, may be settled if these challenges are met. Insight on this issue, while not settling how such fusion takes place in humans, would generate fruitful ideas for experimentation in cognitive AI along similar directions.

References

[Adjo95] Adjoudani, A. and Benoit, C., 1995, "Audio-Visual Speech Recognition Compared Across Two Architectures," Institut de la Communication Parlee, Universite Stendhal, Grenoble, Proceedings of Eurospeach, Madrid.

[Asil 94] Session TA6: Visual Speech Processing Systems, 1994, organized by Garcia, O.N., Conference Record, Twenty-Eight Asilomar Conference on Signals, Systems & Computers, pp.555-590,.

[Cath 96] Cathiard, M.A., Lallouache, M.T., and Abry, C., 1996, "Does Movement on the Lips Mean Movement in the Mind?" in Speechreading by Humans and Machines, NATO ASI Series, Vol. 150.

[Fire 96] "FIREWIRE: The IEEE-1394 Standard," Lundell, A., Digital Video, pp. 30-36, August 1996.

[Garc 92] Garcia, O.N., Goldschen, A.J., and Petajan, E.D., 1992, "Feature Extraction for Optical Automatic Speech Recognition or Automatic Lipreading," IIST Report 32, George Washington

University, Washington, DC.

[Gold 93] Goldschen, A.J., 1993, "Continuous Optical Automatic Speech Recognition by Lipreading," Doctoral Dissertation, Department of Electrical Engineering and Computer Science, George Washington University, Washington, DC.

[Gold 96] Goldschen, A.J., Garcia, O.N., and Petajan, E.D., 1996, "Rationale for Phoneme-Viseme Mapping and Feature Selection in Visual Speech Recognition," in Speechreading by Humans and Machines, NATO ASI Series, Vol. 150.

[Mass 89] Massaro, D., 1989, "A Fuzzy Logical Model of Speech Perception," in Human Information Processing: Measures, Mechanisms, and Models, D. Vickers and P. L. Smith, eds., Elsevier, North Holland.

[Mass 94] Cohen, M. and Massaro, D., 1994, "What can visual speech synthesis tell visible speech recognition?" Conference Record, Twenty-Eight Asilomar Conference on Signals, Systems & Computers, pp.555-590,.

[Mass 96] Massaro, D.W., 1996, "Perceiving Talking Faces," in Speechreading by Humans and Machines, NATO ASI Series, Vol. 150.

[Meie 96] Meier, U., Hurst, W., and Duchnowski, P., 1996, "Adaptive Bimodal Sensor Fusion for Automatic Speechreading," pp. 833-836, ICASSP.

[Move 96] Movellan, J., 1996, "Channel Separability in the Audio-Visual Integration of Speech: Implications for Engineers and Cognitive Scientists," in Speechreading by Humans and Machines, NATO ASI Series, Vol. 150.

[Peta 96] Petajan, E.D., and Hans Peter Graf, 1996, "Robust Face Feature Analysis for Automatic Speechreading and Character Animation," in Speechreading by Humans and Machines, NATO ASI Series, Vol. 150.

[Robe 96] Jordi Robert-Ribes, 1996, "A Plausible and Functional Model for Audiovisual Integration," INPG, Universite Stendhal, in Speechreading by Humans and Machines, NATO ASI Series, Vol. 150.

[Sils 93] Silsbee, P.L., 1993, "Computer Lipreading for Improved Accuracy in Automatic Speech Recognition," doctoral dissertation, University of Texas.

[Sils 95] Silsbee, P.L., 1995, "Sensory Integration in Audiovisual Automatic Speech Recognition," Conference Record, Twenty-Eight Asilomar Conference on Signals, Systems & Computers, pp.555-590,.

[Stor 96] Stork, D.E. and Henessy M.E., Eds., 1995, Speechreading by Humans and Machines, NATO ASI Series, Vol. 150.

[Summ 87] Dodd B., and Campbell, R., eds., "Some Preliminaries to a comprehensive account of audio-visual speech perception," in Hearing by Eye: the Psychology of Lip-reading, pp. 3-51, Lawrence Earlbaum Associates, London, 1987.

[Vati 96] Vatikiokis-Bateson E., 1996, ATR Human Information Research Laboratories, Kyoto, Japan, personal communication.

[Wozn 79] Wozniak, V.D. and Jackson, P.L., 1979, "Visual Vowel and Diphthong Perception from Two Horizontal Viewing Angles," Journal of Speech and Hearing Research, vol. 22, pp. 354-365.