The Effect of Region of 'Activity' Measures on Automatic Audio-Visual Speech Recognition

Andrei P. Makhanov
The Image and Video Processing Laboratory
2145 Sheridan Rd., Evanston, IL
apm714@ece.northwestern.edu

ABSTRACT

Automatic Speech Recognition (ASR) has made tremendous progress in the last few decades. Even so, audio-only speech recognition (A-ASR) does not work well in noisy environments. The standard approach to dealing with this shortcoming is to use visual information along with the audio. Many approaches to using the visual modality have been devised. In this paper, I propose a method that will try to mimic human perception of visual speech by measuring activity in regions of the lower face and applying the measures to train Hidden Markov Models.

1. RELATED WORK

From intuition, it is apparent that humans use visual speech information in day to day interactions. This has also been demonstrated experimentally and is known as the McGurk effect [2]. In any proposed audio-visual automatic speech recognition (AV-ASR) system, two fundamental questions arise: how is the visual information to be used and what are helpful features that should be extracted. Much work has been done to find the most useful features and then to combine these with audio in order to create noise-robust AV-ASR systems.

Over the past 30 years, there have been three main approaches to choosing visual features: *Appearance* based features, *shape* based ones, or some combination of both [4]. In systems that utilize the first approach, pixels in the regions of interest around the mouth, where the majority of the lipreading information is thought to be contained, are transformed using principal component analysis or by taking the discrete cosine transform. This makes it possible to represent each instance of a mouth region with relatively few parameters. In theory, the advantage is that every pixel in the mouth region is utilized, however the dimensionality of the data is greatly reduced. The parameters obtained are then merged with audio features and then fed into a speech recognition system. Systems that utilize the second approach, perform procedures such as finding the edges in an image and then extracting the approximate lip contours. These contours are then fit to pre-defined parameterized models, again, greatly reducing the amount of information needed to describe the pixels in the mouth region. As in the previous case, the model parameters are combined with audio parameters and fed into a recognition system.

An example of the second approach is the work done by Aleksic [1]. In this system, edges in the mouth region are used to find the general outline of the lips. A gradient vector field is found, and elastic 'snakes' finally settle on an approximation of the outline of the lips. Facial animation parameters, as defined by the MPEG-4 standard, are obtained from the changes of the shapes of the snakes from frame to frame. These parameters are then combined with the Mel frequency cepstral coefficients of the audio waveforms and fed into a Hidden Markov Model based system. While this system performs well on the female speaker of the Bernstein Lipreading Corpus, it does not do nearly as well on databases in which the video sequence has less ideal conditions (such as shadows). More recently, much of the AV-ASR work has combined elements of both appearance and shape based features to describe regions of interest that are thought to be useful in speech recognition.

2. PROPOSED METHOD

The proposed method of feature extraction is inspired by the idea that human perception of speech does not rely on specific anatomical features. That is, it appears that humans are not consciously (or subconsciously) aware of the precise borders of lips or the exact geometrical relationships between the various parts of the face. I propose that humans infer meaning from regions of activity in certain parts of the face. Clearly, this idea needs experimentation to be proved or at least better defined. Some insight can be obtained by examining the images in Figures 1 and 2. In the first figure, the outline of the lips can be clearly picked out. However, in the second figure, the top lip is merged with the inner mouth, while the bottom lip is missing with only its shadow present. However, when the sequences of both speakers are played in a video, both segmentations seem to be just as helpful to a human observer. This leads me to believe that the segmentation performed picks up the useful, dominant features that are necessary for a human being. My method attempts to pick out some regions around the mouth area and measure the activity in these regions to get an estimate of how useful they are in AV-ASR.

The feature extraction method proceeds as follows: a) The face-tracking component from [1] is used to get a rough estimate of the region of interest, in this case, defined as the region containing the mouth. b) The Adaptive Clustering Algorithm [3] is applied to the region of interest to obtain a two-level (binary) image c) Find and track lip corners. d) Based on lip corners, segment the six regions shown in Figure 3 e) For each frame, sum all of the black pixels in each region

Figure 1: Example of database in which lip contours are well segmented (nevermind the rest of the face)

Figure 2: Example of database in which outline of lip is not directly segmented

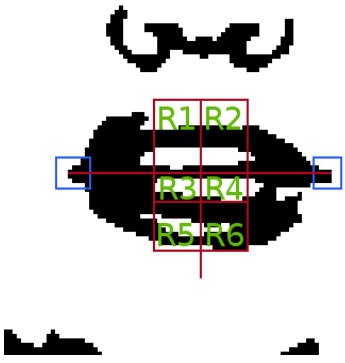

Figure 3: Regions used to define visual features

Once the features have been extracted, they are combined with audio features consisting of 12 Mel frequency cepstral coefficients and an energy coefficient. These are then passed in as observations to the Machine Learning component of the algorithm.

The machine learing portion used the Hidden Markov Model Toolkit (HTK) [5]. Before any learning occurred, HTK was used to take the combined audio and visual features and find the frame to frame 'difference' and 'acceleration' values for each observation. The difference and accelerations for the visual features can be seen as an interpretation of the 'activity' in a given region of the face.

The system made HMMs with 5 states for each phoneme in the English language, with an additional silence model. Because some of the observations had floating point values, all of them were treated as continuous distributions. Several combinations of multidimensional Gaussian mixtures were used to describe the observation distributions for each state; i.e. 3 mixtures for each audio coefficient and 5 mixtures for each visual coefficient. The models were trained on one portion of the data and tested on the the other. The learning task consisted determining which (phoneme) model was the most likely to have produced each observation sequence in the test set.

The audio waveforms used in this experiment were sampled at a rate of 90 Hz, while the video was sampled at 30 frames per second (30 Hz). In order to have the same number of audio and visual samples, the visual features had to be interpolated between each frame. This trippled the total number of features. As an aside, it was found that performance of a system that simply repeated the visual features 3 was similar on in which interpolation occurred. The integration of the audio and visual features utilized what is known as 'late integration' [1], a technique in which HMMs are trained in a way that assigns differing weights to multiple input streams. In this method, there were two streams: the audio stream and the video stream. The stream weights applied to each stream added up to 1.

3. EXPERIMENT

The experiment utilized the Bernstein Lipreading Corpus as the data set. This high quality audio-visual database includes two speakers, one male, one female. A total of 474 sentences uttered by the female speaker were used for this machine learning task. For each of the sentences, the database contains a speech waveform, a word-level transcription, and a video sequence time synchronized with the speech waveform. The vocabulary size is approximately 1,000 words. In order to extract visual features from the database, the video was sampled at a rate of 30 frames/sec (fps) with a spatial resolution of 320 x 240 pixels, 24 bits per pixel.

To simulate a noisy environment, additive Gaussian white noise was applied to the audio waveforms. The highest level of noise in this experiment reduced the audio signal to noise ration (SNR) to 0 dB. Though other noise levels were experimented with, only the results for the 0 dB case are displayed. It was observed, as has been shown many times, that the visual features gave the greatest performance boost at the

highest noise levels.

Of the 474 sentences, 95% were used to train the system, while the other 5% were used to test it. Each word-level transcription was broken down into the phoneme component parts to be used for both testing and training. Once the system was trained, the test set was used to evaluate how much these 'activity' features contributed over the audio-only system. The detected phonemes were grouped into words. These words were compared to the original transcriptions for each sentence. The system finally obtained the percentage of the words that were correctly identified.

The results obtained are shown in Figure 4. As mentioned previously, multiple combinations of varying numbers of mixtures were tested for both the audio and visual features. There was no single ideal combination that was better than all of the others for all noise levels and all combinations of stream weights. The graph displays the case in which each audio parameter was modeled by 5 mixtures and each visual parameter was modeled by 7 features.

The results show that with these features, an 80-20 distribution of weight yields the highest performance. With the same dataset and noise level, the audio only speech recognition got about 42% of words correct. Adding the visual features boosted the recognition performance by over 10%.

4. CONCLUSION

This experiment attempted to capture activity in the mouth region for a set of video sequences. This activity did not depend on finding the exact outline of the lips or any other anatomical feature – something that is quite difficult to accomplish in less ideal databases. It was hypothesized that the resulting segmentation of the adaptive clustering algorithm preserves important dominant features that are useful to human lipreading. I showed that these same segmentation results are also useful for automatic speech recognition. It should be noted that the visual features used in this experiment are quite simple. They are simply summations of pixels of one color for constant sized regions. Even so, these scalar values seem to carry a lot of information. The basic goal of this experiment was to see if these features add any improvement to the automatic speech recognition. The fact that such simple features boosted the performance by over 10 percent is a very promising prospect for future work. It should be noted that in most cases the performance of this system was only a few (2-4) percent poorer than that of the significantly more complex system proposed by [1].

5. FUTURE WORK

It has been demonstrated that the described method is certainly not shape based. It is also a bit different than most appearance based methods, particularly because it uses binary images. I plan to investigate the difference between applying appearance based techniques to gray-scale (256 shades or more per pixel) images, and the binary images obtained by using ACA. I would also like to better define the concept of measuring activity in a region of the face. The segmentations seem to carry much useful information, but the challenge is in finding the best way to fully utilize this information. Lastly, better techniques of face localization can be looked into in order to more precisely extract the desired region in the face.

6. REFERENCES

[1] P. Aleksic, J. Williams, Z. Wu, and A. K. Katsaggelos. Audio-visual speech recognition using mpeg-4 compliant visual features. *EURASIP J. Appl. Signal Processing*, 2002(11):1213–1227, November 2002.

[2] H. McGurk and J. MacDonald. Hearing lips and seeing voices. *Nature*, 264(5588):746–48, 1976.

[3] T. Pappas. An adaptive clustering algorithm for image segmentation. *Signal Processing, IEEE Transactions on [see also Acoustics, Speech, and Signal Processing, IEEE Transactions on]*, 40(4):901–914, 1992.

[4] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. Senior. Recent advances in the automatic recognition of audio-visual speech. *Proc. of the IEEE*, 91(9), September 2003.

[5] S. Young, G. Evermann, M. Gales, and ... *The HTK Book (for HTK Version 3.4)*. Cambridge University Press, Cambridge, 2001-2006.

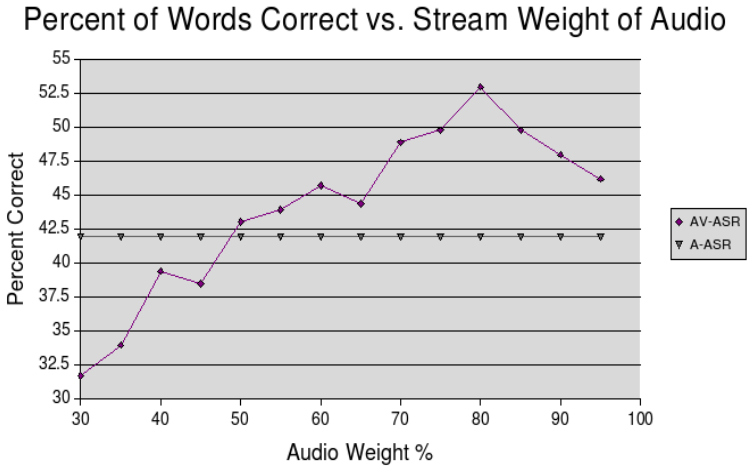

Figure 4: Some preliminary results