AN AUDIO-VISUAL APPROACH TO SIMULTANEOUS-SPEAKER SPEECH RECOGNITION

E.K. Patterson

Department of Computer Science
University of North Carolina at Wilmington
Wilmington, NC 28403, USA
pattersone@uncw.edu

J. N. Gowdy

Electrical and Computer Engineering
Clemson University
Clemson, SC 29634, USA
jgowdy@ces.clemson.edu

ABSTRACT

Audio-visual speech recognition is an area with great potential to help solve challenging problems in speech processing. Difficulties due to background noises are significantly reduced by the additional information provided by extra visual features. The presence of additional speech from other talkers during recording may be viewed as one of the most difficult sources of noise. This paper presents a study using audio-visual speech recognition for simultaneous-speaker speech recognition. The desired goal is to separate and potentially recognize speech from several simultaneous speakers.

Speaker pairs from the CUAVE multimodal speech corpus are used in this work. Audio-visual techniques are compared against speaker-independent and speaker-dependent audio-only methods for speech recognition of individuals from these pairs. For information on obtaining CUAVE, please visit the following web page (http://ece.clemson.edu/speech).

1. INTRODUCTION

The power of computing has risen over the past few years to the level where separate modalities such as audio and video can be used in a complementary method to improve desired results. Audio-visual speech processing has shown great potential, particularly in areas such as speech recognition and speaker authentication. For speech recognition the addition of information from lipreading or other features helps make up for information lost due to corrupting influences. Because of this, audio-visual speech recognition has the potential to outperform audio-only recognition, particularly in noisy environments. Researchers have demonstrated this potential of the audio-visual approach using various experimental methods [1, 2, 3, 4]. Typically, the addition of information from visual features improves recognition rates, particularly in the presence of background noise.

Another potential application of audio-visual speech methods is to recognize speech from multiple, simultaneous speakers, a task difficult to perform during audio speech recognition. Speech from another user is one of the most challenging sources of noise, as all the characteristics are similiar to the speech of the desired user to be recognized. Improved performance in this area would help solve the speech babble problem and aid in several applications where crowds or other non-user talkers are present.

This paper entails a study of the potential of using audio-visual speech recognition to improve results in multispeaker environments. Recognition of speech from simultaneous speaker pairs is tested using audio-visual features and compared to audio-only recognition performance. Both concatenated and multi-stream fusion methods are used to improve recognition performance.

2. MULTIPLE, SIMULTANEOUS SPEAKER SPEECH RECOGNITION

Although it is an important problem, research to date on speech recognition of multiple, simultaneous speakers has been fairly limited. This is in part because it is an extremely difficult task. It is similar to the "babble" or "cocktail" problem where speech needs to be separated from a background of similar acoustic features, but it also has the additional requirement that we desire to potentially recognize the "stream" of speech from any particular speaker. Some work has been performed that focuses on blind signal separation through nonlinear means such as the use of neural networks with inputs from two microphones (close-talking and omni-directional) to help distinguish multiple signals [5]. Although limited, these techniques have demonstrated some success. Another difficulty with testing multiple, simultaneous speech is recording data. One audio database, ShATR, has been recorded, using 8 microphones to aid research in this area [6]. A drawback of these approaches, though, is the requirement of headset microphones to help isolate a desired speaker's speech. An ideal approach could focus on a chosen speaker who is not wearing a microphone, either for comfort or because of the application environment. The purpose of this work is to investigate the performance of audio-visual speech recognition as a means to recognize speech from multiple, simultaneous speakers.

3. TEST SETUP AND METHODS

This work uses the last task in the CUAVE database, where speaker pairs pronounce strings of digits similar to telephone numbers. Individuals alternately speak two strings of digits separately but in the same field of view, then speak another string of digits simultaneously. The audio-visual speech recognizer tracks multiple faces and extracts features from each lip region. This additional visual information should provide a means to help separate individual speakers for continuous digit recognition in this case. Results of a speaker-independent, simultaneous-talker, audio-visual recognizer are compared against those of a traditional audio-only recognizer. Results are included for an ideal, speaker-dependent audio recognizer and a more realistic, speaker-independent recognizer.

The face-and-lip tracking routine was coded so that it could track multiple faces. The process begins with searching for the

Figure 1. Speaker Group (DCT, 2 females)

Figure 2. Speaker Group (DCT, female, male)

largest segment of face-classified blocks. Once this block is found, it is searched for the corresponding lip region. The tracker then returns to search the remaining area of the video frame for an additional large segment of face-classified blocks. This is assumed to be a second face and searched for corresponding lips, as well. Features are extracted for each set of lips and recorded based on which speaker was on the left or right of the frame. The features chosen for these tests were the standard, 2-D DCT coefficients as detailed in previous work [7]. Figures 1 and 2 are frames taken from the tracking program that illustrate tracking of two separate faces/lips and downsampling for extraction of the 2-D DCT. Three arbitrary speaker groups are actually used in this work as representative pairs: male/male, female/male, and female/female.

The DCT difference coefficients were concatenated with standard audio features (MFCC) and passed to the audio-visual recognizer. For audio-only testing, several recognizers were constructed. Six speaker-dependent recognizers were created by training on recordings of the test speakers from other portions of the CUAVE corpus. Also, one speaker-independent recognizer was trained on all speakers but those used in these tests. A speaker-independent scenario is a more realistic setup, but the speaker-dependent recognizers are included for additional comparison. The speaker-dependent training should allow the recognizers to perform almost as "matched filters" for each of the speakers. The audio-visual recognizer results are compared against each of these audio recognizers. The initial audio-visual results are based on simple, concatenated early-integration features.

Speaker	*Audio Ind.*	*Audio Dep.*	*Joint AV*	*Optimal AV*
S01 M	0.00 %	23.33 %	20.80 %	36.67 %
S02 M	13.33 %	36.67 %	30.00 %	30.00 %
S04 F	20.00 %	23.33 %	23.33 %	26.67 %
S20 F	16.67 %	13.33 %	36.67 %	46.67 %
S33 M	10.00 %	13.33 %	20.00 %	23.33 %
S34 F	3.33 %	10.00 %	50.00 %	56.67 %

Table 1. Results (Word Accuracy) for Speakers from Multiple, Simultaneous Speaker Tests (DCT).

To improve upon these results, multistream HMMs were implemented as well. Multistream HMMs allow weighting of the audio and visual features for superior performance. In this technique, two streams of features enter the recognizer. There are separate models for each of the streams, but they are aligned in a state-synchronous manner. The probabilities generated by each model are weighted by coefficients that give the "strength" of each stream. These can be changed to stress the visual or audio information.

4. EARLY-INTEGRATION RESULTS AND DISCUSSION

The results of the simultaneous speaker tests for the audio recognizers and early-integration audio-visual recognizer are given in Table 1, where each row represents the results of recognizing that speaker out of a pair. The scores given are recognition accuracies, based on the following formula:

$$Accuracy = (H - I)/N * 100\%, \quad (1)$$

where N is the total number of words expected, H the total correctly recognized, and I the number of insertions. Recognition accuracy is a more practical measure, as it is typically lower because of insertion errors. The audio recognizer performance tended to suffer more from insertions, due to the other speaker's voice or non-speech sounds that interfere. The visual recognizer, however, seemed more prone to delete words by not recognizing sufficient mouth movement, apparently, to recognize a word in a particular time segment.

The results in Table 1 reveal very poor performance for recognition accuracy for the speaker-independent, audio-only recognizer. These scores are similiar to comparable 0 dB noise scores from previous work [8], since the other speaker may be viewed as a noise source. The speaker dependent recognizers, trained specifically for each speaker, do perform significantly better. The first audio-visual scores presented are obtained with a speaker-independent, early integration recognizer. Audio and visual features are merely concatenated with no weighting on either information stream. These scores allow a significant performance improvement over both the independent and even more ideal dependent recognizers. In all cases, the joint recognizer outperforms the independent audio recognize by a large margin and only falls short of the dependent recognizer in one case, for speaker 2. Another conclusion from the results follows intuition that the female speaker should be easier to distinguish from the male speaker in that test group. The audio-visual recognizer performs significantly better in this case. Interestingly, though, recognizing the speech of the male speaker does not gain the same performance boost.

5. MULTISTREAM RESULTS AND DISCUSSION

In an attempt to yield more improvement, multistream audio-visual recognition was implemented using the same test sets. In multistream recognition, the audio and visual features maintain separate information streams coming into the recognizer. Separate HMMs are constructed for each information stream. During training, the streams receive equal weighting, but during recognition, the streams may be weighted according to confidence. This method has been used successfully on subband recognition, and there has been some success to date with audio-visual recognition as well. The concept of multistream recognition is also similar to that of the fuzzy weighting often used in late-integration approaches to audio-visual speech recognition. Intuitively, the visual information should become more useful as the noise increases, and a similiarly volumed, second speaker could be viewed as a case of 0 dB noise. Based on this, audio stream weights of around 0.25 should produce improved results, as this is roughly a proper normalized weighting of the audio information in noisy cases based on previous noise fusion work [8]. Results, however, do not indicate the same trend as before in this regard. Table 1 also contains the results of the optimal-stream-weighted, multistream audio-visual recognizer in the last column. These results are optimal because the highest performance was chosen for each speaker, regardless of what the λ weighting value was. The optimal performance is shown to achieve a significant improvement over the early integration recognizer.

An obstacle to overcome for practical implementation, though, is that there is not a strong trend between the stream weighting and optimal recognition performance based on these results. The stream-weighting performance for each of the speaker pairs is demonstrated in Figures 3 - 5. There is no particular pattern about which λ values achieve the highest recognition scores. For a few of the speakers, highest performance is achieved with a low-audio ratio, around $0 - 0.25$, such as for the noise cases. The most regular peaks, though, over all six speakers appear in the $0.5 - 0.7$ range where the audio features are favored with some influence from the visual information. Leaning slightly more toward the audio decision seems to coincide with the lower potential of the visual recognizer in this continuous-speech, speaker-independent study. Based on this, the joint recognizer achieves the best performance when using the stronger audio information while gaining some information from the visual stream to help separate recognition decisions for each speaker.

Figure 6 includes a chart of all speakers that demonstrates that there is no apparent trend. If each line is followed, though, most peak around the 0.5 to 0.7 range, except for speaker 2. Interestingly, this is the same speaker where the audio-visual recognizer fails to exceed the dependent audio performance. Figure 7 demonstrates averaged performance for recognition rate and accuracy over all six speakers. The recognition rate does not include insertions, and is thus much higher. Insertions, however, are the most likely problem when attempting to recognize simultaneous speech. The peak in the average recognition accuracy is just above 0.5, an even weighting of audio and video streams. Figure 8 is the information from Table 1 in chart form. It can be seen that the independent audio recognizer has the lowest performance, and that the joint multistream recognizer achieves the highest performance, by a large margin in most cases.

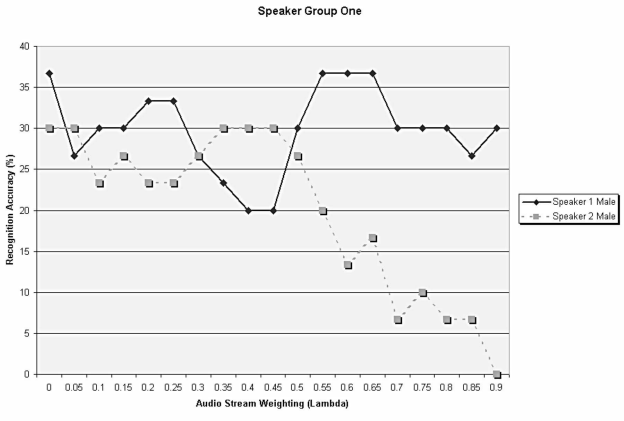

Figure 3. Recognition Accuracy versus Audio Stream Weight

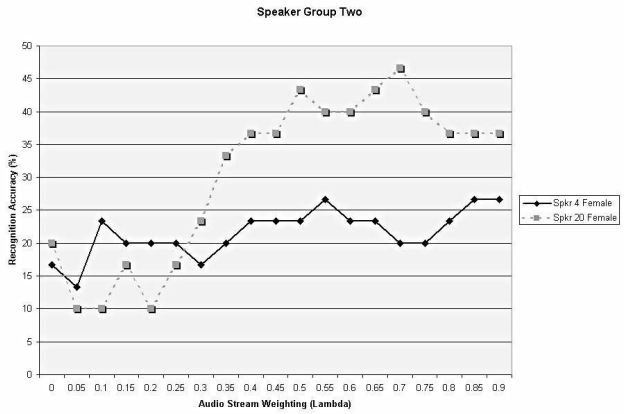

Figure 4. Recognition Accuracy versus Audio Stream Weight

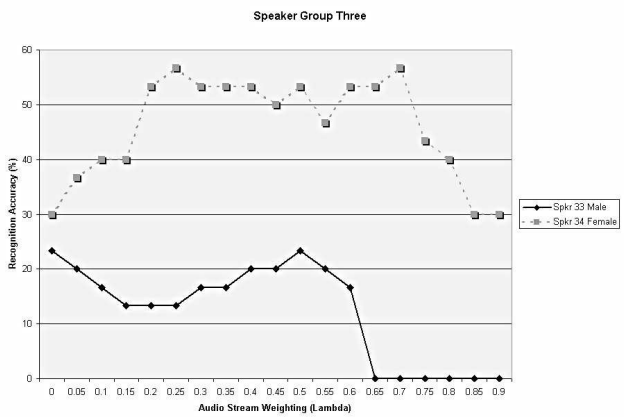

Figure 5. Recognition Accuracy versus Audio Stream Weight

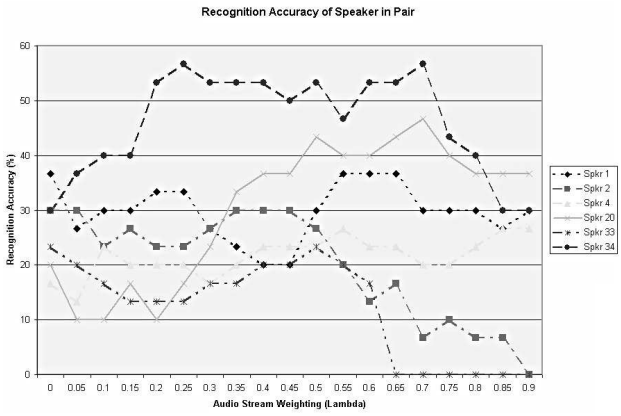

Figure 6. Recognition Accuracy for Several Speakers from Pairs

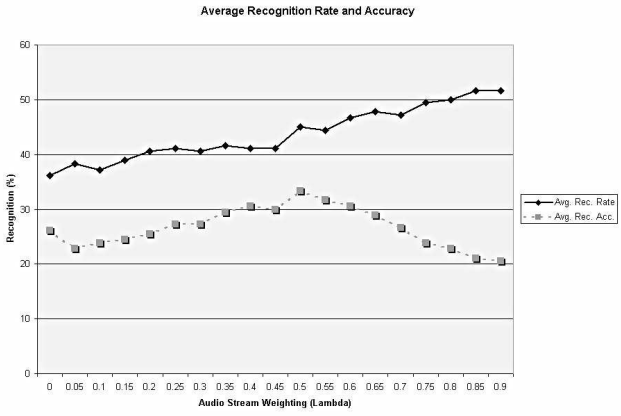

Figure 7. Recognition Rate and Accuracy Averaged over Speakers

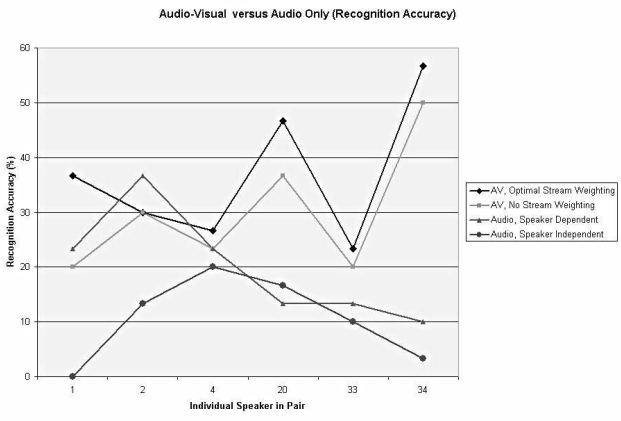

Figure 8. Recognition Accuracy for Audio-Visual and Audio Methods

6. SUMMARY AND CONCLUSIONS

This work attempted to test and improve performance for simultaneous, multiple-speaker speech recognition. Recordings from the CUAVE database were used. A speaker-independent audio-only recognizer was implemented and trained on all the speakers of the database except those used in the speaker-pair testing. Also, several speaker-dependent recognizers were created for each of the individual speakers. Joint audio-visual recognizers both based on early-integration and state-synchronous multistream fusion were implemented and tested as well. The speaker-dependent audio recognizers outperform the speaker-independent recognizers as expected. The joint audio-visual recognizer, though, outperforms the speaker-dependent audio recognizer in all cases except one where it nearly matches performance. The multistream recognizer exceeds this performance and illustrates the best ability to "separate" and recognize speech from the simultaneous speakers, but an important issue is choosing the optimal fusion ratio. In the case of these experiments, the best multistream ratios relied slightly more on the audio than video. This is likely because the audio recognizer was the more reliable of the two, but further study in this area could be useful for finding optimal fusion techniques in regard to simultaneous speech recognition.

7. REFERENCES

[1] E. Petajan, B. Bischoff, D. Bodoff, and N. Brooke, "An improved automatic lipreading system to enhance speech recognition," in *ACM SIGGHI, pp. 19-25*, 1988.

[2] P. L. Silsbee and A. C. Bovik, "Computer lipreading for improved accuracy in automatic speech recognition," *IEEE Transactions on Speech, and Audio Processing*, vol. 4, no. 5, September 1996.

[3] C. Neti, G. Potamianos, J. Luettin, I. Matthews, H. Glotin, D. Vergyri, J. Sison, A. Mashari, and J. Zhou, "Audio-visual speech recognition final workshop 2000 report," Center for Language and Speech Processing, Johns Hopkins University, Baltimore, 2000.

[4] G. Potamianos, C. Neti, G. Iyengar, and E. Helmuth, "Large-vocabulary audio-visual speech recognition by machines and humans," in *Eurospeech, Denmark*, 2001.

[5] A. Koutras, E. Dermatas, and G. Kokkinakis, "Continuous speech recognition in a multi-simultaneous-speaker environment using decorrelation filtering in the frequency domain," in *3rd International Workshop SPECOM, St. Petersburg, Russia*, 1998.

[6] M. Crawford, G. Brown, M. Cooke, and P Green, "Design, collection, and analysis of a multi-simultaneous-speaker corpus," in *Proceedings of the Institute of Acoustics, Vol. 16*, 1994.

[7] E.K. Patterson, S. Gurbuz, Z. Tufekci, and J.N. Gowdy, "Moving-talker, speaker-independent feature study and baseline results using the cuave multimodal speech corpus," *EURASIP Journal on Applied Signal Processing*, November 2002.

[8] E.K. Patterson, S. Gurbuz, Z. Tufekci, and J.N. Gowdy, "Noise-based audio-visual fusion for robust speech recognition," in *International Conference on Auditory-Visual Speech Processing, Denmark*, 2001.