AUTOMATED LIP-READING FOR IMPROVED SPEECH INTELLIGIBILITY[*]

Matthew McClain[†], Kevin Brady, Michael Brandstein, and Thomas Quatieri

MIT Lincoln Laboratory, 244 Wood St., Lexington, MA 02420-9185
mrmcclai@students.uiuc.edu, {kbrady, msb, quatieri}@ll.mit.edu

ABSTRACT

Various psycho-acoustical experiments have concluded that visual features strongly affect the perception of speech. This contribution is most pronounced in noisy environments where the intelligibility of audio-only speech is quickly degraded. An exploration of the effectiveness for *extracted* visual features such as lip height and width for improving speech intelligibility in noisy environments is provided in this paper. The intelligibility content of these extracted visual features will be investigated through an intelligibility test on an animated rendition of the video generated from the extracted visual features, as well as on the original video. These experiments demonstrate that the extracted video features do contain important aspects of intelligibility that may be utilized in augmenting speech enhancement and coding applications. Alternatively, these extracted visual features can be transmitted in a bandwidth effective way to augment speech coders.

1. INTRODUCTION

An approach to improving the effectiveness of speech processing systems, such as low-rate coders, in the presence of high acoustic noise, is to augment the signals from the acoustic microphone with signals from other sensors which are less sensitive to acoustic noise. This paper presents an exploratory study of the application of video sensors for this purpose. Specifically, the effectiveness of extracted visual features such as lip height and width is investigated as a means for improving speech intelligibility in noisy environments.

Numerous intelligibility tests and psycho-acoustical experiments have demonstrated the allure of providing visual information of a speaker's face in addition to an acoustical signal for improving the perception of speech content by a listener. The reasons for these improvements are twofold [3]. First, video can provide redundancy that can improve the performance of discriminating speech [3], such as illustrated in the well known McGurk effect [6]. This redundancy is particularly evident in the lower frequency range (<800 Hz). Voiers [8] demonstrated that a bandwidth limited audio channel (<400 Hz) where the listener can see the talker speaking provides similar intelligibility to full bandwidth audio where the listener cannot see the talker. The second reason why visual information improves speech perception is that visual cues can aid a listener in tracking the acoustical signal. This complementary information is primarily in the middle frequency range (800 Hz-2 KHz) [3].

Joint audio / video processing has been investigated for a broad range of speech technology applications, including speech recognition [7, 9], speaker identification / verification [1], and speech segmentation [5]. Comparatively, little work has been done in the area of speech enhancement utilizing visual features [2]. In this previous work, features extracted from the audio and video channels are used to train a classifier that was used to estimate the clean vocal tract transfer function of the speech. A perceptual test was used to characterize the enhancement algorithms performance in improving a listener's ability to identify speech sounds. A significant improvement was demonstrated for voiced speech, though mixed results were demonstrated for unvoiced speech.

The primary contribution of this paper is in characterizing the intelligibility content of the visual features that have been extracted from video. This intelligibility content is typically in the unvoiced component of speech that is easily degraded by noise. This paper will present a plausibility experiment based on the intelligibility testing of several sensing modalities, including one that utilizes audio + animation that will be baselined against intelligibility tests utilizing audio-only and audio + video. The importance of these experiments is twofold. First, they characterize the effectiveness of state-of-the-art visual extraction algorithms in extracting visual features that can aid speech applications addressing intelligibility issues (e.g., speech coding and speech enhancement). Second, they demonstrate an effective

[*] This work was sponsored by the Defense Advanced Research Projects Agency under Air Force Contract F19628-00-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

[†] Currently at the University of Illinois, Department of Electrical Engineering, Urbana, Illinois 61801

0-7803-8484-9/04/$20.00 ©2004 IEEE

modality (i.e., audio + animation) that can be utilized as a bandwidth-effective approach for improving speech coders.

The remainder of this paper is organized as follows: Section 2 provides a description of a plausibility experiment that investigates the utility of visual information for improving speech intelligibility. Section 3 presents the results of the experiment and analyzes the effects of a specific speech attribute. Section 4 provides concluding remarks, including a discussion of applications and future work.

2. EXPERIMENT

An audio/visual corpus has been collected to facilitate intelligibility testing. In collecting this corpus, a video feature extraction algorithm [9] developed at the Georgia Institute of Technology was applied to the video data. The algorithm detects the lip region by segmenting the hue/saturation/intensity color space and using motion information. Our experiments indicated that hue was highly effective in discriminating most of the lip region, though the saturation was necessary to discriminate lip pixels near the corners of the lips.

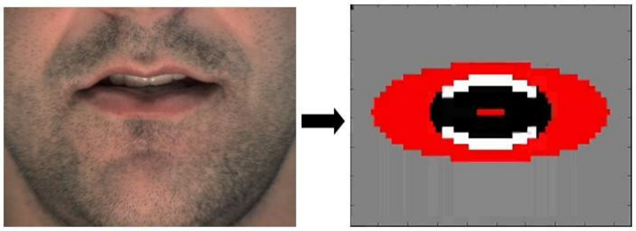


Figure 1. Lip animation extraction

The algorithm extracts points corresponding to the extremities of the inner and outer lips, as well as the amount of teeth and tongue present. Six features were derived from the video feature extraction algorithm as a function of time: inner lip height, inner lip width, outer lip height, outer lip width, amount of teeth present, and amount of tongue present. These parameters were utilized to generate an animated mouth consisting of lips, teeth, and tongue for each speaker uttering each word, as seen in Figure 1.

The original corpus was split into three different corpora. One contained only the original audio, the second contained the original audio and video, and the final contained the original audio and animated lips. Finally, three new corpora were generated by electronically adding noise to the audio channel of the above three corpora. The noise was of a M2 Bradley Fighting Vehicle at about 0 dB.

The intelligibility test used in these experiments (see Figure 2) is based on the Diagnostic Rhyme Test [8] that evaluates the perceptual aspect [7] of speech intelligibility. The DRT test consists of word pairs that differ only in their leading consonant (unvoiced) sound (eg. ‘veal’, ‘feel’). Listeners are presented with a choice from a word pair, and select which of the two words that they perceive to have been spoken. The intelligibility testing for this experiment used the same basic testing approach as the DRT, though it was an informal, in-house test. It only used a small selection of DRT word pairs (10 word pairs) and 5 male speakers, and the noise was added electronically. This resulted in larger error bars than would be typical in a DRT test.

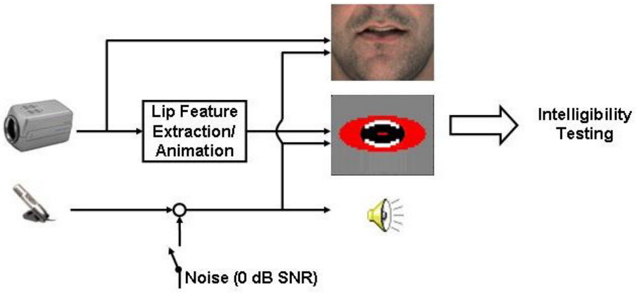


Figure 2. Intelligibility testing experiment

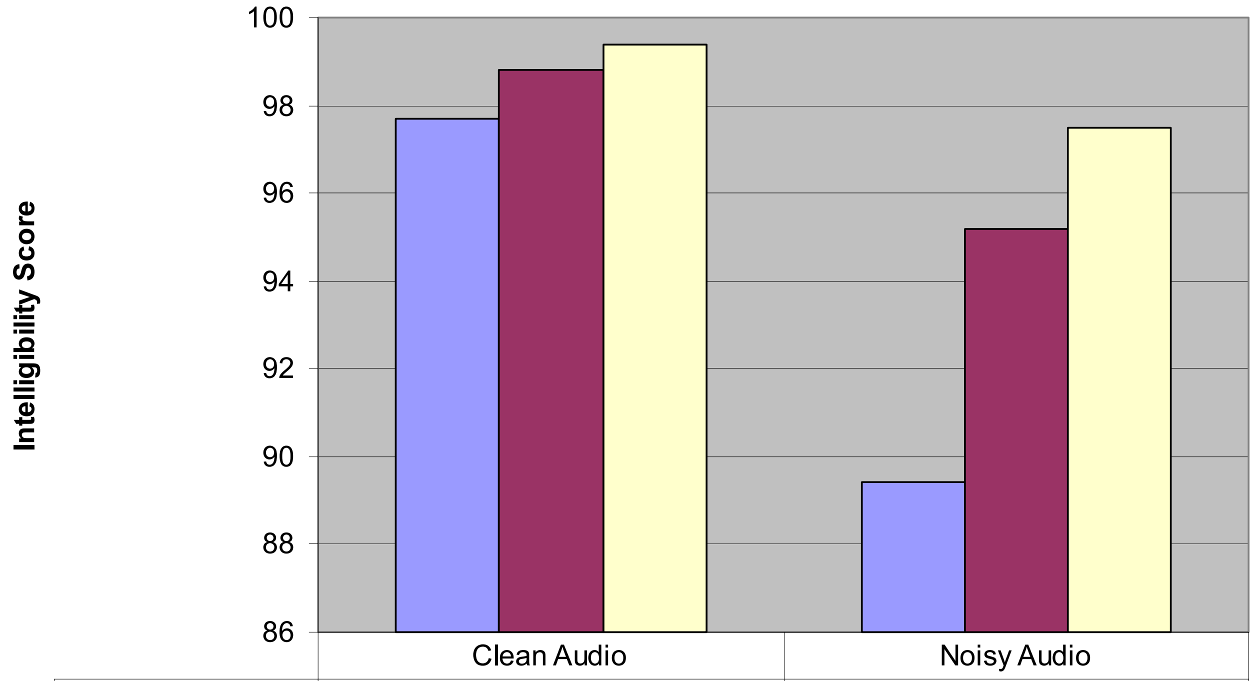
The goal of the intelligibility testing is to evaluate the three different modalities (audio-only, audio + video, audio + animation) in clean and noisy conditions. Only unencoded speech was evaluated. A total of 14 male listeners were used for this evaluation. The intelligibility scores [8] utilized in the next section are in the range of -100 to 100, with 0 being chance:

$$\text{Intelligibility Score} = \frac{\underline{(\#\ \text{Correct} - \#\ \text{Incorrect})}}{\#\ \text{Total}}$$

3. RESULTS

The intelligibility scores corresponding to the three different modalities in clean and noisy conditions are shown in Table 1. The error bars for these experiments are approximately 1.5-2.0 intelligibility points. All three sensor modalities perform very strongly in the clean environment. The audio + animation modality demonstrates an apparent intelligibility improvement over audio-only, while audio + video provides the highest level of intelligibility. There is not enough separation in the confidence intervals to make broad generalizations about the clean audio case.

In the noisy audio case the audio-only modality took a significant hit in intelligibility performance. Unsurprisingly, the audio + video modality provided the highest level of intelligibility performance. Of great interest is that the performance of the audio + animation modality was quite strong. In fact, the separation for the test results in Table 1 provides a strong and statistically meaningful intelligibility performance improvement for audio + animation over the audio-only case. These intelligibility results are almost as strong as the audio + video modality. This implies that the visual features extracted from the lip region are providing much of the intelligibility information that is lost in the audio-only modality due to noise.



	Clean Audio	Noisy Audio
▢ Audio-Only	97.7	89.4
■ Audio+Animation	98.8	95.2
▢ Audio+Video	99.4	97.5

Table 1. Audio/Video Intelligibility Experiment Results

Intelligibility testing can be refined into an evaluation of intelligibility attributes [8], such as: voicing, nasality, sustention, sibilation, graveness, and compactness. Of these attributes, graveness is one of the attributes most strongly impacted by environmental noise [8], second only to sustention. Graveness distinguishes phonemes articulated in the front versus the middle parts of the vocal cavity [8]. Five of the ten word pairs used in these experiments were specifically selected from the graveness word pair list to evaluate the effectiveness of the tested sensing modalities in addressing this attribute. Intuitively, the ability to determine the graveness attribute would be complemented by the availability of lip features that provide information on the visible articulators during the leading consonant.

The results of testing the graveness attribute for the audio-only and audio + animation modalities in the M2 noise field is seen in Table 2. All of the selected word pairs chosen from the DRT graveness list demonstrated significant intelligibility performance improvements, with one exception. The pair 'bid/did' showed a slight degradation in performance that can be attributed to its already high level of proper discrimination (100.0 for audio-only). Particularly impressive performance improvements are seen for all four of the other word pairs (7.7 - 13.8 intelligibility points). The overall average gain for the graveness attribute was 8.3 intelligibility points (85.9 to 94.2) as seen in Table 2.

4. CONCLUDING REMARKS

A plausibility experiment has demonstrated the utility of extracted video features for extending the capability of applications addressing speech intelligibility issues, such as speech enhancement and speech coding. This experiment has demonstrated that extracted video features contain information that can improve the identification of unvoiced speech in noise. Particularly impressive results were demonstrated for the graveness attribute of speech intelligibility.

These video-related modalities are specifically being investigated for their potential in improving the performance of speech coding applications. Speech coding may be improved by direct improvements to the speech coder, or by augmenting a speech enhancement algorithm that can be used as a preprocessor to the coder. For example, in speech coding applications, the video features can be used in one of two ways:

(1) The video features can be utilized to supplement audio features for generating the codebook for specific features.

(2) The relevant portions of the video channel (i.e. lip features) can be coded in parallel with the audio channel.

The importance of the audio + animation experiment is in demonstrating that the extracted video features do contain important aspects of intelligibility (1), and could be transmitted in a bandwidth effective way (2). Future work will investigate the tradeoffs of these two approaches in utilizing video to augment audio-only coding approaches in noisy environments.

Word Pair	Audio-Only	Audio+Animation
Bid/Did	100.0	98.5
Fought/Thought	70.8	83.1
Weed/Reed	80.0	93.8
Peak/Teak	90.8	100.0
Moon/Noon	87.7	95.4
Average	85.9	94.2

Table 1: Intelligibility scores in noisy audio environment

An underlying assumption in this paper is the impact of environmental noise on the collected corpus. Since the acoustic noise was added electronically, the Lombard effect [4] is not present in the audio channel. Similarly, there was no impact of the acoustic noise on the visual articulators. In fact, this is a typical assumption in audio-visual applications. An experiment is proposed where audio-video speech data is collected in clean and various noisy environments of DRT word pairs. The intelligibility of the audio-only, video-only, and audio + video modalities can then be evaluated to determine the impact of acoustic noise on speech intelligibility for video.

5. ACKNOWLEDGEMENTS

The authors would like to thank Professor Xiaozheng Zhang of California Polytechnic State University, San Luis Obispo for providing the software utilized for extracting the visual features in these experiments. We would also like to acknowledge the late Iwan Ulrich of Carnegie Mellon University whose digital firewire camera software has provided invaluable for acquiring our video clips. Likewise, we would also like to acknowledge the creators of the AviEdit, Alexander Milukov, a shareware toolbox that was used for editing our movies. Finally, we gratefully thank the volunteers at MIT/LL for participating as speakers and listeners. The results are encouraging, but the intelligibility testing was laborious and tedious.

6. REFERENCES

[1] C. C. Broun, X. Zhang, R. M. Mersereau and M. A. Clements, "Automatic Speechreading with Application to Speaker Verification", *ICASSP*, Orlando, May 2002.

[2] L. Girin, J.-L. Schwartz, and G. Feng, “Audio-visual Enhancement of Speech in Noise”, *J. Acoust. Soc. Am.*, Vol 109, # 6, pp 3007-3020, June 2001

[3] K. W. Grant and S. Greenberg, “Speech Intelligibility Derived from Asynchronous Processing of Auditory-Visual Information”, Auditory-Visual Speech Processing (AVSP) Workshop, 2001

[4] E. Lombard, “Le Signe de l’Elevation de la Boix”, Ann. Maladriers Oreille, Larynx, Nez, Pharynx, 37, 1911

[5] M. W. Mak, W. G. Allen, “Lip-motion Analysis for Speech Segmentation in Noise”, *Speech Communication*, Vol. 14, pp. 279-296, 1994

[6] H. McGurk and J. MacDonald, “Hearing lips and seeing voices”, Nature, pp. 746-748, Dec. 1976

[7] G. Potamianos, C. Neti, G. Iyengar, Eric Helmuth, “Large-Vocabulary Audio-Visual Speech Recognition by Machines and Humans”, *Proc. EUROSPEECH*, Aalborg, Denmark 3-7 September 2001

[8] W. D. Voiers, “Diagnostic Evaluation of Speech Intelligibility”, *Benchmark Papers in Acoustics*, Vol. 11: Speech Intelligibility and Speaker Recognition (M. Hawley, ed.) Dowden, Hutchinson and Ross, Stroudsburg (1977)

[9] Z. Zhang, R. M. Mersereau, and M. A. Clements, “Audio-Visual Speech Recognition by Speechreading”, *14^{th} Int. Conf. on Digital Signal Processing*, 2002