Measuring the Dynamic Performance of Video Conferencing Systems

David I. Havelock[1], and David Green[2]
[1]Institute for Microstructural Science, NRC, 1200 Montreal Rd., Ottawa ON K1A 0R6, david.havelock@nrc.ca
[2]Institute for Information Technology NRC, 1200 Montreal Rd., Ottawa ON K1A 0R6, dave.green@nrc.ca

1. INTRODUCTION

The effectiveness of a video conferencing system is determined by diverse factors such as video and audio quality, interactive system control mechanisms, and the ability of the system to track conversation. System performance is often based on the subjective assessment of criteria that lack the rigor necessary to obtain definitive comparisons between different systems and strategies.

An automated video conferencing system emulating the skill of an expert videographer in capturing the visual and audio dynamics of a presentation is of great value for video conferencing [1], tele-presentations [2], meeting archiving [3], and possibly surveillance.

2. OBJECTIVE

We are concerned here with tracking talkers visually and acoustically during a conversation and obtaining performance metrics that can be conveniently measured, compared between systems, and related to subjective judgments of performance. As an initial step, a metric is defined for the ability of a video conferencing system to follow the talker transitions in conversation with multiple participants. The methodologies are applicable to a generic multimodal system [4] and are implemented in a system independent manner to determine the audio and video switching delays independently, and are demonstrated using a prototype system that combines independent audio and video talker localization [1].

3. METHOD AND APPARATUS

Pre-recorded 'conversations' are used with two talkers (or noise sources) speaking alternately. Audible and visible markers are inserted into the scene at a variable delay t_d from the onset of each transition (event), as in **Fig. 1**. If the steering mechanisms are able to switch fast enough to display the marker then the event is recorded as a 'hit', otherwise it is a 'miss'. The probability of a hit as a function of marker delay gives the system latency probability distribution function (PDF).

The visible marker is an LED flash and the audible marker is a pair of 1 kHz tone bursts 15 ms long. The markers are synchronized as in **Fig. 1** and are presented through speakers and an LED light system as shown in **Fig. 2**. Two sets of speakers and lights are positioned at the same distance (~90 cm) from the array and camera center, with 120° angular separation. The LED flash duration t_d is varied and if the system latency is less than t_d then the observer sees the flash (a 'hit'). The first of a pair of tone bursts occurs at the event time, and the other occurs t_d later. The response of the microphone array is loudest when steered at the source so, if the system steers to the source fast enough, the second pulse is heard to be louder than the first and a 'hit' is declared for that event.

During a synthesized conversation scenario, the observer indicates when a hit or miss occur and the responses are collected on-line. About 30 runs, each with a range of 40 delays, were used to estimate each PDF. Multiple audio pulses and video flashes were used at each event to enhance the statistical significance of the observations. Both noise and recorded speech are used as audio sources. For speech, randomized phrases from a list of Harvard Sentences were used, with a male and a young female voice.

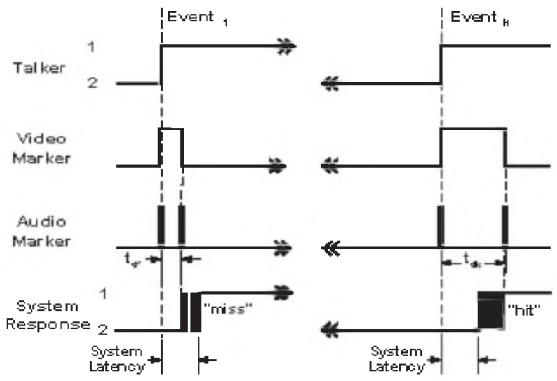

Figure 1 Timing Sequence. On the left (a 'miss'), the video and audio markers are finished before the system (lower trace) steers to the active talker. On the right (a 'hit'), the system responds fast enough to catch part of the video indicator and the second of the two audio markers.

4. OBSERVATIONS

Figure 3 shows the video and audio latency PDF with a sigmoid function overlaid. The video and audio PDF have similar slope and spread but are offset by about 600 ms. The PDF data for noise and speech are similar in both cases. The video PDF curve has a positive zero-delay intercept because the system video and data buffers allow non-memoryless processing.

The mean system response delay is estimated by the midpoint

of the sigmoid and the maximum delay by projecting the tangent at the mid point to the top axis, as shown by the large solid dots in **Fig. 3**. The proposed metric comprises these four values (mean and maximum delay for audio and video system response).

Figure 2 System setup. The 16-microphone array, panoramic camera (center of array), speaker on tripod, and LED light system are shown. Another speaker and light system are located 120° about the array.

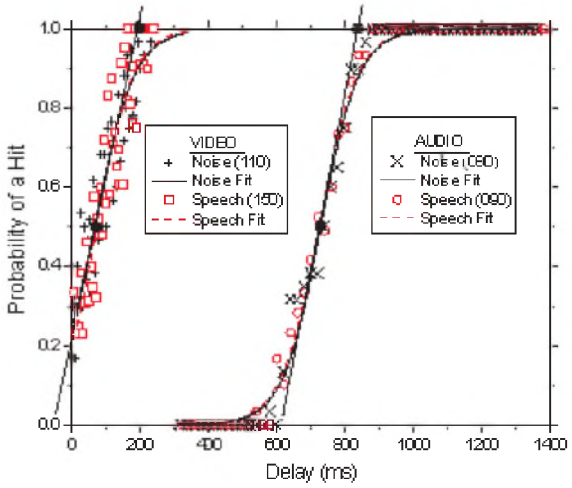

Figure 3 System response probability distribution function (PDF) indicating the probability that the system will response at least as fast as the ordinal delay. Estimated PDF values (points) from measurements are fit to a sigmoid function (lines). The metric for system response comprises the delays at the four solid dots. (Numbers in the legend are dataset codes.)

5. CONCLUSIONS AND DISCUSSION

The metric comprising the mean and maximum systems delays has been measured in a system-independent manner and is suitable for intercomparisons. The relevance of the metric for subjective performance is not yet confirmed.

The sigmoid function is a reasonable data model when many factors contribute to the delay, but its asymptotic nature is problematic for estimating the maximum delay accurately. The tangent method is a convenient approximation.

The shift of the audio PDF relative to the video is due mainly to the message packet queue used for steering the microphone array in the prototype Panocam system.

System processing is synchronous with the video frame rate (about 7 fps), yet the sloping PDF indicates variations in response time in excess of 250 ms. There are three identifiable contributors to this variation; the timing of the event relative to the start of the video frame, the asynchrony of the audio buffer and video frame, and jitter in the messaging system between the audio and video subsystems.

The video frame buffer memory allows a PDF with a non-zero value at zero delay. The video display, however, is delayed by the frame buffer interval and contributes to the system display latency.

The choice of sound source (noise, or male or female speech) has been observed to make minor differences in the PDF. Our scenario assumes a conversation is underway, with talkers identified, but the PDF may vary at the introduction to a conversation. A half-second of silence was inserted between events to model 'polite' conversation; other scenarios, such as interrupting or overlapping conversation, may give different results. The video and audio markers are thought to have little impact on the system operation.

REFERENCES

[1] M. Fiala, D. Green, and G. Roth, (2004) "A panoramic video and acoustic beamforming sensor for videoconferencing," 3rd IEEE Int'l Workshop on Haptic, Audio, and Visual Environments and Their Applications, 2004:47-52.

[2] Yong Rui, Anoop Gupta, and Jonathan Grudin (2003), "Videography for Telepresentations," Proc. SIGCHI Conf. On Human Factors in Computing Systems 2003, Ft. Lauderdale, FL, 5-10 Apr. 2003, Vol. 5, Issue 1, pp. 457-464.

[3] Eric Meurville and David Leroux (2004), "Collection and annotation of meeting room data," m4- Multimodal Meeting Manager Report D1.2, 2 April 2004. [Website on 4 August 2005: http://www.m4project.org/publicDelivs/D1-2.pdf]

[4] D. Lo, R.a. Goubran, R.M. Dansereau, "Multimodal talker localization in video conferencing environments," 3rd IEEE Int'l Workshop on Haptic, Audio, and Visual Environments and Their Applications, 2004:195-200.

ACKNOWLEDGEMENTS

We wish to thank Mark Fiala for help with the video subsystem.