# Modeling and Evaluation of Multimodal Perceptual Quality

**Petersen, Kim T; Hansen, Steffen Duus; Sørensen, John Aasted**

*Published in:*
I E E E - Signal Processing Magazine

*Link to article, DOI:*
10.1109/79.598591

*Publication date:*
1997

*Document Version*
Publisher's PDF, also known as Version of record

man perception of speech is bimodal in that acoustic speech can be affected by visual cues from lip movements. For example, one experiment shows that when a person "sees" a speaker saying /ga/, but "hears" the sound /ba/, the person perceives neither /ga/ nor /ba/, but something close to /da/.

Due to the bimodality in speech perception, audio-visual interaction is an important design factor for multimodal communication systems such as video telephony and video conferencing. A prime example of this interaction is lip reading or speech reading. Lip reading is not only used by the hearing-impaired for speech understanding. In fact, everyone utilizes lip reading to some extent, in particular in a noisy environment such as at a cocktail party. Communication systems must be able to provide the full motion necessary for speech reading by the hearing-impaired. Researches have studied the importance of frame rates with impaired listeners [2] and analyzed the effects of frame rates on isolated viseme recognition [3]. Research in these areas will lead to multimedia systems that account for the perceptual boundaries of the hearing-impaired. Researchers have also tried to teach computers to lip-read [5]. Based on computer-vision techniques for tracking lip movements of a speaking person, a computer can be trained to understand visual speech. In addition, automatic lip reading has also been used to enhance acoustic speech recognition.

What can one do if the frame rate is not adequate for lip synchronization perception, which is a typical situation in video conferencing equipment due to the bandwidth constraint? One solution is to extract the information from the acoustic signal that determines the corresponding mouth movements, and then process the speaker's mouth image accordingly to achieve lip synchronization [4]. On the other hand, it is also possible to warp the acoustic signal to synchronize with the person's mouth movements. The latter approach is very useful in nonreal-time applications, such as dubbing in a studio.

One key issue in bimodal speech analysis and synthesis is the establishment of the mapping between acoustic parameters and the mouth shape parameters. In other words, given the acoustic parameters, such as the cepstral coefficients, one needs to estimate the corresponding mouth shape, and vice versa. A number of approaches have been proposed for this task that utilize vector quantization [7], neural networks [8], Gaussian mixtures, and hidden Markov models [9].

Audio-visual interaction can be exploited in many other ways. The correlation between audio and video can be utilized to achieve more efficient coding of both audio and video [6, 7]. Audio-visual interaction can also be used for person authentication and verification [10, 11, 12]. Other applications include dubbing of movies, segmentation of image sequences using video and audio signals [13], human-computer interfaces, and cartoon animation. All these clearly demonstrate that the joint processing of audio and video provides additional capabilities that are not possible when audio and video are studied separately. It is clear that once we break down the artificial boundary between audio/speech and image/video processing, many new research opportunities and innovative applications will arise.

## References


1. H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, pp. 746-748, December 1976.

2. Frowein, et al., "Improved speech recognition through videotelephony: Experiments with the hard of hearing," *IEEE Journal on Selected Area in Communication*, vol. 9, no. 4, May 1991.

3. J. Williams, J. Rutledge, D. Garstecki, A. Katsaggelos, "Frame Rate and Viseme Analysis For Multimedia Applications," *Proc. of IEEE Multimedia and Signal Processing Conference*, Princeton, NJ, June 1997.

4. T. Chen, H.P. Graf, and K. Wang, "Lip-synchronization using speech-assisted video processing," *IEEE Signal Processing Letters*, vol. 2, no. 4, pp. 57-59, April 1995.

5. D. Stork, "'I could see your lips move': HAL and Speechreading," HAL's Legacy, The MIT Press, 1997.

6. D. Shah, and S. Marshall, "Multi-modality coding system for videophone application," WIASIC'94, Berlin, Germany, October 1994.

7. S. Morishima, K. Aizawa, and H. Harashima, "An intelligent facial image coding driven by speech and phoneme," ICASSP, p. 1795, Glasgow, UK, 1989.

8. F. Lavagetto, "Converting speech into lip movements: A multimedia telephone for hard of hearing people," *IEEE Trans. on Rehabilitation Engineering*, pp. 114, March 1995.

9. T. Chen and R. Rao, "Audio-Visual Interaction in Multimedia Communication," vol. 1, pp. 179-182, ICASSP, Munich, April 1997.

10. J. Luettin, N.A. Thacker, S.W. Beet, "Speaker identification by lipreading," ICSLP, October 1996.

11. M.R. Civanlar and T. Chen, "Password-free network security through joint use of audio and video," *SPIE Photonic East*, November 1996.

12. *Proceedings of the First International Conference on Audio and Video Biometric Person Authentication*, Crans-Montana, Switzerland, March 12-14, Springer-Verlag, Berlin, 1997.

13. J. Nam and A.H. Tewfik, "Combined Audio and Visual Streams Analysis for Video Sequence Segmentation," *ICASSP*, vol. 4, pp. 2665-2668, Munich, April 1997.


# Modeling and Evaluation of Multimodal Perceptual Quality


*Kim Tilgaard Petersen, Steffen Duus Hansen, John Aasted Sørensen, Tech. Univ. of Denmark*


The increasing performance requirements of multimedia modalities, carrying speech, audio, video, image, and graphics, emphasize the need for assessment methods of the total quality of a multimedia system and methods for simultaneous analysis of the system components. It is important to take into account still more perceptual characteristics of the human auditory, visual, tactile systems, as well as combinations of these systems. It is also highly desirable to acquire methods for analyzing the main perceptual parameters, which constitute the input for the total quality assessment. Altogether, this is necessary for opti-

mization of interacting modalities and the associated multimedia system transmission bandwidths. Examples of systems with interacting modalities are given in [2] and [3]. Today there is only little established consensus about methodologies for design and quality assessment of multimedia systems [3]. There is an increasing effort to incorporate perceptually important properties in the design and analysis of communication systems in general, as exemplified by [4] and [1].

In the following a framework is suggested for assessing the quality of modalities and their combinations. It is based on models for the total quality of given modalities together with their perceptually important parameters. The models can either be based on two input signals consisting of a reference and its encoded/decoded representation, as exemplified in [1] for speech, or a single input signal just consisting of the latter one.

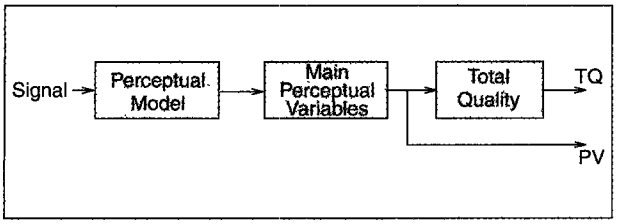
A quality model for a single modality (Fig. 3) can be structured in three layers. The first layer receives either the above two or one input signals and estimates, based on a perceptual model, a selection of parameters that can be used in the second layer for the estimation of the perceptually important parameters from which the total quality measure is derived in the third layer. A speech coding example of this model is given in [5]. Such models require design based on subjective tests and factor analysis.

In Fig. 4 a possible multimedia quality assessment model is shown, which delivers a selection of variables for audio- and video-system analysis. PVA/V are the perceptually most important variables for audio/video and TQA/V are the corresponding total quality of audio/video. All these are independent characterizations of audio/video.

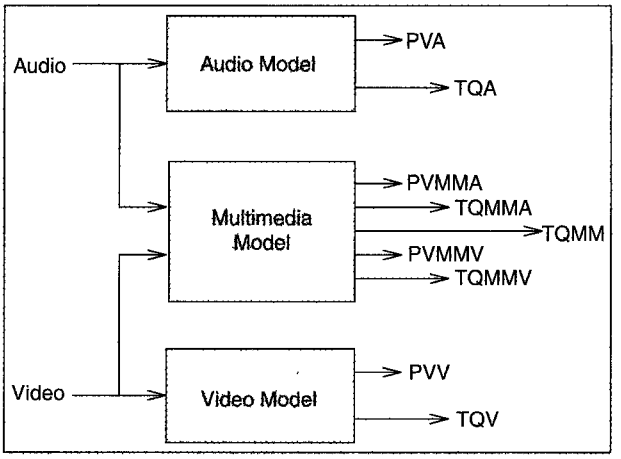
The variables PVMMA/V describe the most important multimedia perceptual variables for audio and video. Accordingly, the measures TQMMA/V describe the total quality of audio/video, and finally TQMM describes the total quality of the multimedia system, taking into account simultaneous audio and video. The construction of such models represents substantial efforts, but this might be necessary to carry out simultaneous optimization of the modalities in multimedia systems.

## References


1. J.G. Beerends, J.A. Stemerdink, "A Perceptual Speech-Quality Measure Based on a Psychoacoustic Sound Representation," *J. Audio Eng. Soc.*, vol. 42, no. 3, 1994 March.


*3. Three-layer quality model of a modality.*




*4. Multimedia quality assessment model.*


2. T. Chen, R.R. Rao, "Audio-visual Interaction in Multimedia Communications," *Proc. of 1997 Int. Conf. on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 179-182.

3. James Flanagan, Ivan Marsic, "Issues in Measuring the Benefits of Multimodal Interfaces," *Proc. of 1997 Int. Conf. on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 163-166.

4. N. Jayant, J. Johnston, and R. Safranek, "Signal Compression Based on Models of Human Perception," *Proc. of the IEEE*, Oct. 1993, pp. 1385-1422.

5. K.T. Petersen, J.A. Sørensen, and S.D. Hansen, "Objective Speech Quality Assessment of Compounded Digital Telecommunication Systems," IEEE 1997 Workshop on Multimedia Signal Processing, Princeton, June 1997.


## Signal Processing for Networked Multimedia


*Reha Civanlar and Amy Reibman, AT&T Labs*


Real-time transmission of multimedia data over packet networks poses several interesting problems for signal processing research. Although the range of these problems covers a large variety of topics, currently two groups appear to attract the most attention. The first group concerns adapting the signal compression techniques to address the special requirements imposed by the packet networks, including accommodating for packet losses, delays, and jitter; providing capability for multipoint; and coping with the heterogeneous nature of today's networks. The second group of problems is related to protecting the intellectual property rights (IPRs) associated with the transmitted multimedia data. The increasing availability of high-bandwidth networking makes it extremely easy to illegally duplicate and disseminate digital information. Unless a mechanism can be established to protect the rights of the content providers, commercial use of networked multimedia will remain extremely limited.

Adapting signal compression techniques to networked applications may require some changes in the fundamental approach to this problem. The goal of classical signal compression is to achieve the highest possible compression ratio. The compression and transmission aspects have