TECHNICAL REQUIREMENTS FOR REALISTIC TELEPRESENCE

Geraldo Lino de Campos[1]

Abstract

Telepresence resources are available for at least two decades, but its use is still very low. The main reason is that current systems, to save on telecommunication costs, imposes restrictions on quality that precludes the transmission of minute details, very important for the transmission the subjective contents of the material. This paper proposes minimum standards of quality to overcome this problem.

1. Introduction

This paper uses the word *Telepresence* to mean all the instances a person is being represented by an imaging system, like in teleconferences, teleteaching and teleworking.

The first practical video conferencing systems were introduced at the Fourth World Telecommunications Forum held in Geneva in 1983 [Eva95], but their use is still rare. By comparison, the personal computer, introduce about the same time, is now a commodity sold by the millions.

That is surprising, considering the cost and time savings associated with teleconferencing and the economy of scale achievable by teleteaching.

[1] Departamento de Engenharia de Computação e Sistemas Digitais, Escola Politécnica da Universidade de São Paulo, Edifício de Engenharia de Eletricidade, Cidade Universitária, São Paulo 05681-900, BRAZIL. E-mail: geraldo@regulus.pcs.usp.br.

This paper presents the opinion that lack of quality in the transmission of details precludes the transfer of information of subjective nature, present in behavioral details, frustrating the communication process. Studies have shown that, in interpersonal communication, about 7% of the communication is transmitted by the content of the word uttered, 38% by the utterance intonation, and 55% by image. The exact number may be disputed, but it is clear that much of the information present in interpersonal communication is conveyed by gestures, small physiological reaction, changes in skin tone, and other minute details.

The importance of subjective of components is easily shown by the study of the evolution of voice synthesis systems. They only achieved practical and widespread use after the subjective aspects of the synthesized voice were fully understood. The same should hold true for the image aspects, and this is the key point to be accessed to build practical telepresence systems.

This paper is organized as follows. Section 2 presents the nature and evolutions of voice synthesis systems, to substantiate the above claim. Section 3 presents some evidence for the importance of the minute details in the transmission of information in human communication. Section 4 suggests some guidelines for the dimensioning of telepresence systems; finally, section 5 presents come concluding remarks.

2. What was learned from voice synthesis

Telepresence is characterized by transmission of voice and image. Voice reproduction was never a problem, but the study of voice properties is very illustrative of the telepresence problem. The usual forms of voice reproductions easily preserve the relevant properties, but this was explained only after the development of voice synthesis.

The first voice systems were able to produce clearly intelligible voice, but in practice people rejected the early systems. The comprehension of this phenomenon requires some knowledge about the natural production of voice.

Human speech is generated by the excitation of the vocal tract either by a train of pulses (voiced sounds), generated by the vibration of the vocal cords, or by white noise generated by a constriction in the vocal tract (unvoiced sounds).

The vocal tract behaves acoustically as a time-varying acoustic tube; its characteristics can be modeled by a sequence of small acoustic tubes, with time varying diameters, generating resonance peaks called formants. Modeling the resonance of the vocalic tract is easily and faithfully performed by a technique called Linear Predictive Analysis (LPA) [Ata71] , which generates an all-pole model form the resonance of the vocal tract. The difficult part is modeling the excitation.

Until recently, the excitation was modeled as a simple train of pulses or pure white noise; the resulting speech was intelligible, but quality was very poor and acceptable only for specialized applications, like secure communications. Moreover, the systems were very sensitive to ambient noise.

In recent years, several algorithms have been proposed for modeling the excitation in a more faithful way. The main innovation is to substitute artificial excitation models, like crudely simulated glottal pulses or white noise, by a better representation for the excitation. They vary in details of quantization, number of parameters, computational requirements (always high) and offer now good to excellent quality.

The problem with the early synthesizers was lack of knowledge about the exact structure of the glottal pulses. Everyone is able to distinguish a word pronounced with anger from a word pronounced calmly or passionately, but the physical characteristics were unknown. Research for determining the reasons for synthetic voice rejection determined that this difference is related to the regularity of the time interval between glottal pulses (technically called glottal pulses "phase jitter"). When the pulses are regularly spaced, the voice produced with a meaning of anger. As the time between the pulses starts to change slightly from pulse to pulse, voice turns to "normal voice", and when that difference increases even more, we get a sweet voice.

That is why people rejected the early voice synthesis systems: but not knowing this property, synthesizers used a very stable frequency for simulating the glottal pulses; people heard an angered

voice, and immediately reacted with anger against the voice. Latter systems introduced a controlled amount of phase jitter, and voice synthesizers became accepted and useful.

Many other physical properties are relevant for the subjective perception of voice. For instance, the shape of the glottal pulse also has influence in the properties of voice - that is reason why we can speak loudly in soft voice, and can shout with a low volume voice. Every aspect is interesting in itself, but they will not be further examined here, by space limitations.

The important conclusion is that there are many minor aspects of the properties of voice that are quite significant for the subjective perception and acceptance of voice synthesis and transmission systems. These aspects were poorly understood because the usual transmission systems tend to preserve very well the properties of the glottal shape, frequency and phase jitter; most problems happen with the formants properties, that are important for the musical quality of voice, but not for the subjective interpretation of its meaning. It is also important to put in relief that these important aspects were not well understood before the introduction of voice synthesis by computers.

Analogous phenomena should happen with images. Subtle properties are perceived by people as the core of the communication process, but the physics of these changes are not yet studied in detail, and many may be still unknown. The lack of this information is the probable cause of low acceptance of the usual means of telepresence.

3. Informal evidence for the relevance of small details in images

In this section two informal evidences for the relevance of small details in human communication will be presented. As stated in the introduction, there is a lack of formal studies in this area.

3.1 A renaissance example

Most people had had eye examinations with pupil dilatation, caused by some drops of a solution containing a natural alkaloid, atropine, obtained from a vegetal called belladona or deadly nightshade. If one has the curiosity to ask what is the meaning of belladona, it is the Italian word for "beautiful woman".

Belladona was used, in the form of a lotion, by the renaissance courtesans in Italy, and perhaps in other European countries as well. It was said that it clarified and softened the skin, but its role was really to enlarge the pupils, simulating one of the secondary aspects of sexual arousal. I do not know if the courtesans were aware of the phenomenon, but it certainly increased their revenues. It was so important to promote the association of plant name to its use as a cosmetic. And, most probably, the customers weren't aware at all.

This is a good example of minute details in image making a big difference in the subjective reaction. Most of the current telepresence systems will not be able to represent such a small modification in the pupil's diameter, failing to give an important clue about the intention and meaning of the person represented.

3.2 A contemporary example

As a second example, the of effects of seeing a movie in a large screen and watching the same movies in a TV set will be compared, based on a casual experience. Consider the Ingmar Bergman's movie "Scenes from a marriage". Originally designed for a TV series, it consists mostly of closes.

The film has a long sequence showing only the couple in a typical wife-husband discussion. Seen in a large screen movie theater, we can observe that the room is in absolute silence. After the scene ends, every spectator moves in his chair, or coughs, or makes other noises, showing that the scene called everyone attention to a degree that immobilized the spectators. Watched in a TV set, the scene does not have any special impact - just two people arguing.

What is happening is that minute details present in the movie screen convey much more meaning to scene; meaning that is the essence of communication in that situation. With the lower resolution in the TV set, the meaning is lost. (This observation rises another interesting question: What differentiates a good and a bad actor? Perhaps the unconscious capability of making this subtle gestures and movements that convey the emotional content of the scene represented?)

Again, most of the current telepresence systems will not be able to represent such minute details.

4. Image requirements

There are no conclusive studies about the requirements for transmission of behavior details. That is not surprising, since the behavioral details themselves are only superficially known. A few consideration, however, can lead to a few implementation guidelines.

This topic requires extensive experimental studies, in a broad range of circumstances. In the absence of such studies the present guidelines should be regarded at most as educated guesses, and only as a starting point.

The first consideration is about resolution: minute details must be reproduced. These details fail into two categories: spatial resolution, important for the small movements, and color resolution. This last category is important, because small changes in superficial blood circulation are an important conveyor of information; in usual circumstances, they are perceived by subtle skin color changes. This component of emotional reaction is of such importance that is one of the 3 parameters recorded by polygraphs (lie detectors).

A second consideration relates to viewing angles; it is useless to represent minute details in a small screen. If the telepresence is intended for a reasonably sized audience (a meeting group or a classroom), lifelike size is important.

3.1 Resolution requirements

From the renaissance example, and supposing a typical face of 150 mm across, occupying one third of the viewing area, a horizontal resolution of about 1000 pixels (picture elements) per line will be required to represent a small change in pupil diameter. From the contemporary example, a standard movie can convey the required information.

It is extremely difficult to compare the resolution of movies and television; images produced by the two media will never appear exactly the same (that intrinsic difference is the reason why we can tell if a TV set is showing a direct image or a movie). There are many technical difficulties, aggravated by the strong positions of practitioners of both fields. References Ken91 and Tho94 are examples of papers of good technical quality leading to very different conclusions.

We can, however, adopt a simplifying assumption that a conventional 35mm movie has about the same resolution of HDTV (High Definition Television), 1920 by 1080 pixels. This is something like 6 times the resolution of conventional TV, but is achievable and it is expected to become commercial in the next few years. It is also about twice the resolution of the best Personal Computer monitor of today. The amount of data required to feed such a display is enormous, but there are efficient compression techniques that preserve most of the image details.

Since HDTV is intended for the consumer market, it can be expected that equipment will be readily available and of reasonable cost. The aspect ratio (relation of the width to the height) of the image, 16x9, is larger then the usual computer monitor of TV set, and it can be expected that it may show both the speaker and some graphical information, like the contents of a blackboard or a slide.

An important research issue is to determine what level of compression will preserve the significant details.

3.2 Angle of view requirements

The image should be of adequate size to allow the identification of details by the viewer. It is also important that the image occupies most of the viewer visual field, avoiding distractions. For group audience, lifelike dimension is a good, albeit expensive, starting point.

3.2 Color quality requirements

Color quality is an important factor, as explained above about the minute changes in skin color. Unfortunately, representation of skin tones is one of the most difficult aspects of television systems [Ing96]. It is unclear at this point how precisely color must be represented, or if only the change in color is significant. It is also required to do some research in the capacity of the usual compression algorithms to faithfully represent such small changes in color. There are no guidelines here.

5. Conclusion

Telepresence is being around for about 20 years, but it failed to become a mainstream component of communication. We believe that the reason is lack of sufficient quality, causing the loss of important subjective properties of the speaker.

Rising the quality to the movie standard seems to allow to overcome the problem. The forthcoming HDTV standards and equipment also seem to satisfy the requirements for effective telepresence.

The subject has many unsolved problems, and much research is still needed. Some directions are the measure of the minimum resolution requirements, the amount of tolerable color error, and, in the long run, the determination of the details that convey subjective information in business meetings and classroom environments.

5. References

[Ata71] ATAL, B. S. and HANAUER, S. L. "Speech analysis and synthesis by linear prediction of the speech wave", Journal of the Acoustical Society of America 50:2, 637-55, Aug 1971.

[Eva95] EVANS, B. "Understanding Digital TV", IEEE Press, Piscataway, NJ, 1995

[Ing96] INGLIS, A. F. & LUTHER, A. C. "Video Engineering", McGraw-Hill, New York, 1996.

[Ken91] KENNEL, G, DeMARSH, L & NORRIS, J., "A Comparison of Color Negative Films and HDTV Cameras for Television Program Production", SMPTE J., May 1991.

[Tho94] THORPE, L., NAGUNO, F. & USHIKAWA, K., A Comparison Between HD Hyper-HAD Cameras and Color Film for Television Program Production", SMPTE J., June 1994.