REQUIREMENTS FOR REALISTIC REMOTE EDUCATION SYSTEMS

Geraldo Lino de Campos, Prof.
Escola Politécnica da Universidade de São Paulo, P.O.Box 8191, São Paulo 05508-900, Brazil

E-mail: geraldo@regulus.pcs.usp.br

URL: http://regulus.pcs.usp.br

Abstract

Telepresence resources are available for at least two decades, but its use is still very low, both in education and industry. There are two main reasons: to save on telecommunication costs, current systems impose restrictions on quality that precludes the transmission of minute details, essential for the transmission the subjective contents of the material, and to save on display equipment, the amount of displayed information is insufficient. This paper proposes minimum standards of quality to overcome these problems, in a remote education context.

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 [3], 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.

This paper presents the position 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 holds 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 evolution of voice synthesis systems, to substantiate the above claim. Since the same justification for image properties would be lenghty, it will not be presented here, but is available in [2]. Section 3 presents the requirements for image quality, enhancing 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, and finally, section 5 presents some 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) [1], 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.

3. Image requirements.

As stated in the introduction, there is a lack of formal studies in this area. However, informal evidence can be gathered from several sources. A few examples are presented in [2].

3.1 Requirements for transmission of behavioral details

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 considerations, however, can lead to a few implementation guidelines.

Images have the same properties observed with voice: 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.

In teleteaching applications, the low quality of the transmission of details in the teacher’s image 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 numbers 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.

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.

3.2 Resolution requirements

A movie screen can in many instances, especially when displaying images of faces in close, convey the emotional content of the scene. The same scene in a TV screen usually can not. This is a first clue about required resolution.

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 [5] and [6] are examples of papers of good technical quality leading to very different conclusions.

It is possible, however, adopt a simplifying assumption that a conventional 35mm movie has approximately 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 allow moving the teacher in front of demonstrating material, like a blackboard, in a much more natural fashion then the current situation.

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

3.3 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.

There is no substitute for images large enough; zooming and panning are incompatible with the concentration required in a learning environment. The student should concentrate on the subject matter, not on using the system. In a classroom environment, the panning must have a central control, and the individual student can not look for the information that is relevant to him when the need arises.

3.4 Color quality requirements

Color quality is an important factor. Minute changes in skin color are an important element in face-to-face communication. Unfortunately, representation of skin tones is one of the most difficult aspects of television systems [4]. 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 absolute guidelines here.

4. Volume of information

A typical real classroom has four essential elements:

1. The general environment, that should provide good working and viewing conditions.
2. Semi-static display elements, like slides, charts, maps; the information contained changes slowly and a few times during one class.
3. Dynamic elements, represented by the blackboard or flip charts, were information is dynamically constructed during one class.
4. The teacher.

Classrooms have that organization for centuries, and with small variations are used worldwide in all levels of teaching. Changing this paradigm is always uncomfortable - remember the last time you attended a class or seminar in a room that requires removing the projection screen to use the blackboard.

This organization is not arbitrary. Good general environment includes several aspects, but for our purpose the main requirement is a well-lit room. This requires bright screens and powerful projectors in teleteaching environments.

The semi-static element is a reference of content or indicates the order and sequencing of subjects; the dynamic element shows the details or the construction of the subject, and the teacher transmits still more dynamic information, and orchestrates all activities.

Contrary to this tradition, most teleteaching systems try to put all the information in a single, and usually rather small, screen.

A realistic teleteaching environment should provide the four elements described above. If wall size and hi-resolution displays were available, all elements could be displayed in that display. Since such kind of display is very expensive, it is unrealistic to suggest this solution. An alternative is to present elements 2, 3 and 4 separate screens, with size compatible with the number of students. For group audience, lifelike dimension is a good, albeit expensive, starting point. For a small group – 5 to 8 students – in experimental systems, good 21” monitors (or its HDTV successor) are good enough.

The resolution for static and dynamic elements should high. – the horizontal resolution of a simple blackboard 4 meters wide, used with chalk lines of 0.5 cm is 800 pixels. A 1024x1280 pixels display can represent the information content of a 3 meters wide blackboard – not very large by any standard. The resolution for the teacher image should be compatible with HDTV. The center figure should be the teacher, with the ability to point to left and right as it happens in a classroom. Obviously every screen may convey different information, if the flow of material so requires; but for most of the time is interesting to keep each one displaying the same element of information – just the way things happen in real environments.

An interesting research point is about the possibility of multiplexing the static and dynamic information in a single display. Informal observation of both regular classes and didactic films is that voice information is present almost 100% of time, but visual information is presented less then 50% of time. The subjective influence of switching the content on learning should be evaluated before recommending this hardware simplification.

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/presenter.

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

Quantity of information is another neglected aspect especially on teleteaching systems. All the information present in a real classroom should be present in the virtual classroom.

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.

Quality and quantity of information are essential for making teleteaching realistic, delivering the full promises of the technology. Costs of equipment and telecommunications are going down continually, and what is called hi-end video systems will be commonplace in few years. It is time to start researching high quality systems for teleteaching.

5. References

[1] 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.

[2] CAMPOS, G.L., Technical requirements for realistic telepresence, Teleteaching '98 - Distance Learning, Training and Education - Part III, XV IFIP World Computer Congress, Vienna, Austria/Budapest, Hungary, August 1998, pp 77-84.

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

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

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

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