# A MULTIMODAL COMPLEXITY COMPREHENSION-TIME FRAMEWORK FOR AUTOMATED PRESENTATION SYNTHESIS

*Harini Sridharan*　*Ankur Mani*　*Hari Sundaram*

Arts Media Engineering, Arizona State University

Tempe, AZ 85281

Email: { harini.sridharan, ankur.mani, hari.sundaram }@asu.edu

## ABSTRACT

*In this paper, we present a joint multimodal (audio, visual and text) framework to map the informational complexity of the media elements to comprehension time. The problem is important for interactive multimodal presentations. We propose the joint comprehension time to be a function of the media Kolmogorov complexity. For audio and images, the complexity is estimated using a lossless universal coding scheme. The text complexity is derived by analyzing the sentence structure. For all three channels, we conduct user-studies to map media complexity to comprehension time. For estimating the joint comprehension time, we assume channel independence resulting in a conservative comprehension time estimate. The time for the visual channels (text and images) are deemed additive, and the joint time is then the maximum of the visual and the auditory comprehension times. The user studies indicate that the model works very well, when compared with fixed-time multimodal presentations.*

## 1 INTRODUCTION

In this paper we present a joint multimodal framework that estimates the comprehension time for a multimodal element (audio, images and text) based on the element information complexity. The problem is important in interactive presentations (slide shows, electronic games) where adaptive multi-sensory display mechanisms are needed. This is also important in consumer photo products such as [2], where the consumers create automated audio-visual slideshows.

There has been prior work on mapping the visual content to presentation time [9,14]. Both the models are limited to the visual comprehension and attention. While [14] discusses a relationship between visual complexity and comprehension time, [9] discusses a simple spatial attention model for the images. There has been prior work in auditory analysis [3,11]. In [4], the creates an audio skim by shortening pauses and by detecting segments of high-pitch activity. In [11] uses the idea that auditory perception is related to the identification of structure. There has been prior work on sentence complexity [5,8]. They show the dependence of comprehension on sentence structure and working memory usage. In prior work, there is no formal mechanism to map complexity to presentation time.

In our approach, we develop a joint multimodal model for comprehension. We build upon our early work on visual complexity [14] and map the normalized image complexity to comprehension time. The sound clips are analyzed using a psychological experiment, and the normalized sound complexity is then mapped to comprehension time, by determining upper and lower comprehension time bounds. The sentences are categorized into eleven categories, and are additionally limited to two-clause sentences. Another experiment is conducted to map the category to comprehension time.

In our joint model, we assume that the audio, visual and text modes are independent. This is a simplistic assumption, and yields a conservative model. The joint compression is determined as maximum of the visual comprehension time (including sum of the time for reading text and seeing images), and the audio comprehension time. The user-studies indicate that the joint model works very well.

The rest of this paper is organized as follows. In the next section, we discuss insights into the problem of complexity and comprehension. Then in sections 3,4,5 we develop the comprehension model for images, sound and text. We present the joint model in section 6. We then discuss our experiments and present our conclusions.

## 2 COMPLEXITY AND COMPREHENSION

There is empirical and experimental evidence that suggests that there exists a relationship between the complexity of a media element and its comprehensibility. In auditory scene analysis [10], there are grouping rules for the perception of sound. In film-making, there is a relationship between the size of the shot and its apparent time (i.e. time perceived by the viewer):

*"Close-ups seem to last relatively longer on the screen than long shots. The content of the close up is immediately identified and understood. The long shot on the other hand, is usually filled with detailed information which requires eye-scanning over the entire tableau. The latter takes time to do, thus robbing it of screen time"*[12].

**Figure 1:** How much time is needed to comprehend this image?

Recent results in experimental psychology [7] indicate the existence of an empirical law: the subjective difficulty in learning a concept is directly proportional to the Boolean complexity of the concept. Boolean complexity of a concept is defined as the number of literals, '*n*' in its irreducible form (the length of the shortest prepositional formula representing the concept – i.e. its logical incompressibility).

Feldman's work on the relationship between the compressibility and the concept learning, as well the wealth of empirical evidence has motivated our work on comprehension time of media based upon media compressibility. In this paper we have developed a joint model relating the comprehension time for a multimedia element (visual, audio and text), to their respective normalized complexities.

*0-7803-9332-5/05/$20.00 ©2005 IEEE*

## 3 VISUAL COMPLEXITY

We now summarize the key findings of our earlier work [14] on the relationship between visual complexity and comprehension, as it is a key element of this paper. In [14] we defined the visual complexity of an image to be its Kolmogorov complexity [6]. Thus the visual complexity is defined as follows:

$$K_U(x/n) \triangleq \min_{p:u(p)=x} l(p), \qquad <1>$$

where, *U(p)* denotes the output of the program *p* on an universal Turing machine, *x* is the string of length *n* and $K_U(x/n)$ is the Kolmogorov complexity of the string *x* given the length *n*. Further, since Kolmogorov complexity is non-computable [6], we showed that the Kolmogorov complexity of any string is shown to be asymptotically upper-bounded by the compression ratio provided by any universal lossless image coding such as the Lempel-Ziv coding [14].

$$\lim_{n\to\infty} \frac{1}{n} l_{LZ}(X) \to \frac{1}{n} K_U(X \mid n), \qquad <2>$$

where $l_{LZ}$ is the length of the Lempel-Ziv codeword and where *X* is a binary string of length *n.* A psychological experiment was used to map the visual complexity to the comprehension time based on the average times taken to answer *who, where, what and when* for each image [14]. The experiments showed that the comprehension time for an image with complexity *c* was bound as follows:

$$\begin{aligned} U_b(c) &= 2.40c + 1.11, \\ L_b(c) &= 0.61c + 0.68, \end{aligned} \qquad <3>$$

Where $U_b$ the upper bound is the 95th percentile bound. This means that 95% of the time, the images with the complexity c can be comprehended within this time. And where $L_b$ is the lower bound and *c* is the normalized image complexity. The experiment ignored the temporal correlation in films, and hence the upper bound is a conservative bound on comprehension.

## 4 AUDITORY ANALYSIS

We define the audio complexity of a sound clip as its Kolmogorov complexity. We can derive a formula similar to equation <2>, since it holds for any binary string. In our framework we use FLAC [1], the lossless audio encoder to compute the normalized audio complexity. The normalized audio complexity is just the ratio of the length of the FLAC compressed file to the length of the uncompressed sound file.

We conducted a simple experiment to derive a mapping from audio complexity of a clip to its comprehension time. We created a corpus of 300 sound clips with compression ratios ranging from 0.4 to 1.0. Each sound clip was 20 seconds long and was sampled from the author's personal music store to make sure that the user had heard them earlier. We ensured that the collection of clips were diverse.

Most of the original clips had a compression ratio ranging between 0.4-0.8. The audio sequences with a higher complexity were generated by adding Gaussian noise to the original audio sequences such that the SNR was between 0db to 1 db.

In the experiment, a sound clip was chosen at random and presented to the user. The experiment involved a simple identification task – we asked the user to determine if she could identify the sound. This was done in multiple sessions of five minutes each to avoid fatigue. The response time was recorded.

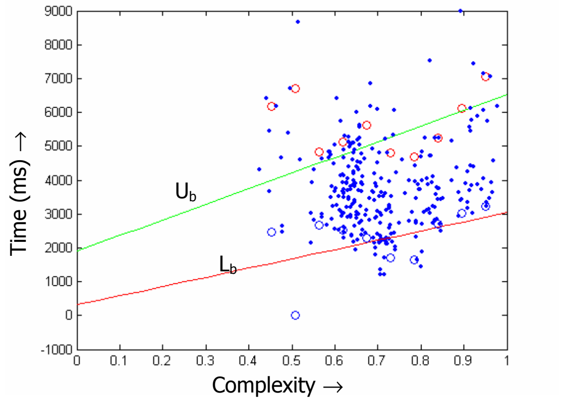

**Figure 2:** Comprehension time plotted against audio complexity and the upper-bound

The response time for each audio clip was plotted against its normalized complexity. The complexity axis was divided into bins and the histogram of the response times for each bin was plotted. For bins with sufficient number of samples, the histogram showed similarity to the Rayleigh distribution. By using the 95th percentile cut-off for each histogram we get an upper bound on the comprehension time for each bin. The upper bound for the comprehension time for each value of complexity was then estimated by the least squares fit to the upper bound in each bin. Similarly, a lower-bound for the comprehension time was estimated using the 10th percentile. The equation of the bounds are as follows:

$$\begin{aligned} U_b(c) &= 4.62c + 1.90, \\ L_b(c) &= 2.72c + 0.32, \end{aligned} \qquad <4>$$

where *c* is the normalized complexity and $U_b$ is the upper bound and $L_b$ is the lower-bound on the comprehension time. The upper bound signifies that 95% of the time, the audio clip can be comprehended in this time. We use the upper bound to estimate the comprehension time of sound clips.

## 5 TEXT COMPREHENSION

We build upon prior work on sentence complexity [5,8]. They show the dependence of comprehension on sentence structure and working memory usage. The authors classify the sentences based upon their sentence complexity and rank them. The authors suggest several guidelines –center embedded sentences are more complex than right branching sentences and that object relative sentences are more complex than subject relative sentences. Examples of such sentences:

- *Right branching sentence*: The boy is robbing the woman who is standing by the pole
- *Center embedded sentence*: The woman who is standing by the pole is being robbed by the boy.
- *Subject relative sentence*: The policemen chased the thief.
- *Object relative sentence*: The thief was being chased by the policemen.

Note that these complexities do not have order relationships amongst them. There is no relationship between the sentences

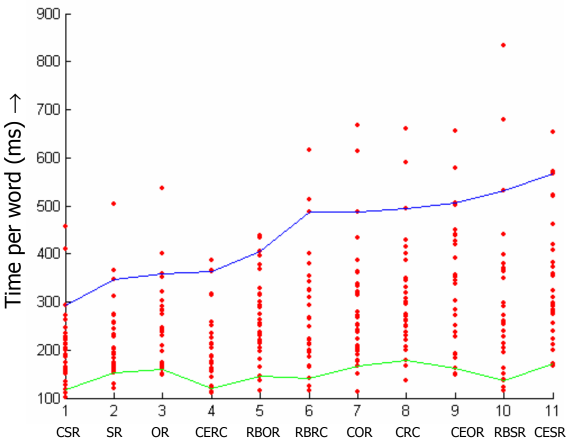

**Figure 3:** Sentence comprehension times for each category and the respective upper bounds.

that differ in two of these properties or two degrees of freedom (thus the Right branching object relative may not be compared to center embedded subject relative). It is important to note that prior work *did not investigate the relationship between these sentence categories and comprehension time*.

We conducted a simple experiment to quantify the time required for comprehension for different categories of sentences. We created a corpus of sentences where each sentence belongs to one of these classes with a maximum of two clauses. We defined eleven sentence categories as follows: *Subject relative* (SR), *Object relative* (OR), *Conjoined subject relative* (CSR), *Conjoined object relative* (COR), *Conjoined role changing* (CRC), *Right branching subject relative* (RBSR), *Right branching object relative* (RBOR), *Right branching role changing* (RBRC), *Center embedded subject relative* (CESR), *Center embedded object relative* (CEOR), *Center embedded role changing* (CERC).

The corpus sentences were presented to a set of six users in a random sequence and we measured time taken by the user to comprehend the sentence. The 95th percentile of the comprehension time for each category was calculated and was fixed as the upper-bound on the comprehension time for the sentence class. The comprehension times and their upper (95th percentile) and lower (5th percentile) bounds for each class, normalized by the length of the sentence, are as shown in Figure 3. Note that the mapping is per class, the classes themselves are not ordered.

Our experiments are consistent with prior results. We use the upper bound comprehension time per class, to compute the comprehension time per sentence. Note that in our framework, we are assuming that the sentences would be classified in one of eleven categories – this is likely in simple, interactive environments where the creator has complete control over the text.

## 6 JOINT COMPREHENSION MODEL

In the previous three section we determined the comprehension time for visuals, sound and text independently. This is an unrealistic scenario, since produced media involves highly correlated elements. Also, in the natural environment sound and vision are highly correlated.

Our approach is very useful in interactive environments where media (audio, visuals and text) are being generated as a consequence of user interaction – in such cases the comprehension time of the multimodal element cannot be known *a priori*. Secondly, the uncorrelated estimates form a highly conservative estimate of the time for comprehension.

We now present a model for estimating the joint comprehension time for the set of media elements representing a particular concept. Then, using the uncorrelated estimates the joint comprehension time of the three elements is as follows:

$$t_J = \max\left(t_{text} + t_{vision}, t_{audio}\right), \qquad <5>$$

where $t_J$ is the comprehension time estimate of a multimodal element comprising text, sound and image. $t_{text}$ is the average comprehension time for the text (averaged over all the sentences in the multimedia element), $t_{vision}$ is the comprehension time of the image, and $t_{audio}$ is the comprehension time estimate of the sound clip.

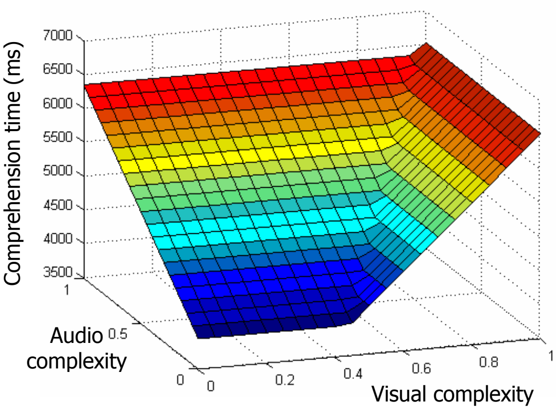

**Figure 4:** Joint Comprehension time plotted against audio and visual complexity.

The formula (eq. <5>) assumes that the audio and the visual channels are processed in parallel. *This would imply that the time to comprehend the text and the image must be necessarily additive.* We use the maximum of the two times, as a conservative estimate. The estimate of the joint comprehension time is plotted against the audio and visual complexity axes for a fixed textual complexity is shown in Figure 4.

## 7 EXPERIMENTS

In this section we discuss our experiments with users and analyze the experimental results. We conducted two experiments that evaluated our model under two different settings. We evaluated our models through a pilot user study with five users.

In the first experiment, two different automatic presentation systems were created, the first one with the media elements being presented for a duration modeled using the joint complexity analysis discussed above and the second with each media element being presented for a fixed amount of time. The time for fixed case was set at 3 sec. – a common duration setting in slideshows. The media elements were presented in the same order for both presentations.

Each user was shown both presentations, in random order. The experiment was double blind. The users were then asked to evaluate the presentation duration. They were asked to rate how many of the media elements were presented for a duration that according to them was adequate and comprehensible. The rating was on a scale of 1-7, 1 representing none of the media elements were presented adequately and 7, all were presented for an adequate duration. The results obtained are tabulated below:

**Table 1:** Average Rating of users for evaluating the media presentation duration

| Presentation type | Adequacy of duration | Comprehensibility |
|---|---|---|
| **Our joint complexity model** | **6.0 / 7** | **6.5 / 7** |
| Fixed media presentation duration | 1.0 / 7 | 2.0 / 7 |

The users felt that in the case of the fixed presentation duration system, the media were shown either too fast or too slow, in most cases. This validates the joint comprehension time model, for the non-interactive media presentation.

In the second experiment, we introduced interactive environments, to test the model. The hypothesis was that interactive frameworks give the users much greater control over the presentation and would hence be less likely to notice the improvements due to our framework. We used an interactive system developed by our group [13] for testing. It was modified to serve the needs to this experiment.

The system comprised of a interactive audio-visual environment. User interaction lead to additional media elements being shown. Three interactive environments were created, with the presentation duration of the media elements in each case being (a) fixed and very low (b) in accordance to our model and (c) fixed and very high respectively. The 'low' duration was fixed to be 1 sec. and the high duration was fixed to be 10 sec. Note the 'optimal durations' for the entire data set lie between 1.5 sec and 8 sec. Hence the 'low' and the 'high' bounds are reasonable.

We allowed users to interact with each of these three systems and asked the users if the presentation time affected their interaction experience adversely. They rated the systems on a scales of 1-7. All users felt that their experience with the system that incorporated our model of media comprehension duration was better than the other two. The results are tabulated below.

**Table 2:** Average Rating of users for evaluating media presentation duration

| Presentation type | Interaction experience |
|---|---|
| Very low fixed presentation duration | 1.30 / 7 |
| **Our joint complexity model** | **6.67 / 7** |
| Very high and fixed presentation duration | 5.67 / 7 |

The rating was good for the presentation system with the high presentation duration because the users had the capability to interact with the system to have it present another media element if they felt they had seen enough – this was not possible with the system with the low duration. The experimental results indicate that our conservative joint comprehension time framework works well, both in non-interactive and interactive frameworks.

## 8 CONCLUSIONS

In this paper, we presented a joint complexity-comprehension model to determine multimedia presentation durations. The work was motivated by observations in film-making and recent result is cognitive psychology. In our framework we assumed that audio, visual and text to be uncorrelated. We showed how the visual complexity as well as the audio complexity cane measured by their Kolmogorov complexity. We conducted experiments on text using sentence categories and measured the normalized comprehension time. The joint comprehension time was derived as the maximum time required for comprehension via the auditory and visual (including time for text) channels. We conducted a variety of experiments on both interactive and non-interactive presentations, and the results indicate that our framework outperforms static-time based presentations. We plan on developing a model that incorporates explicit correlations amongst media, for better comprehension time estimates.

## 9 REFERENCES

[1] *FLAC* http://flac.sourceforge.net/format.html.

[2] *iPhoto* http://www.apple.com/ilife/iphoto/.

[3] B. ARONS (1993). *SpeechSkimmer:Interactively Skimming Recorded Speech*, Proc. of ACM UIST '93, pp. 187-196, 1993, Atlanta GA.

[4] B. ARONS (1994). *Pitch-Based Emphasis Detection For Segmenting Speech Recordings*, Proc. ICSLP, 1931-1934, Sep. 1994, Yokohama Japan.

[5] D. CAPLAN and G. WATERS (1999). *Verbal working memory and sentence comprehension.* Behavioral and brain Sciences **22**: 77-126.

[6] T. M. COVER and J. A. THOMAS (1991). Elements of information theory. Wiley New York.

[7] J. FELDMAN (2000). *Minimization of Boolean complexity in human concept learning.* Nature **407**: 630-633.

[8] E. GIBSON (1998). *Linguistic complexity: locality of syntactic dependencies.* Cognition **68**: 1-76.

[9] Y.-F. MA, L. LU, H.-J. ZHANG, et al. (2002). *A user attention model for video summarization*, Proceedings of the tenth ACM international conference on Multimedia, pp. 533-542, Dec. 2002, Juan-Les Pins, France.

[10] S. MCADAMS and E. BIGAND (1993). Thinking in Sound: The Cognitive Psychology of Human Audition. Clarendon.

[11] E. D. SCHEIRER (1999). *Structured Audio, Kolmogorov Complexity, and Generalized Audio Coding.* IEEE Transactions on Speech and Audio Processing **9**(8): 914-931.

[12] S. SHARFF (1982). The elements of cinema : toward a theory of cinesthetic impact. Columbia University Press New York.

[13] H. SRIDHARAN, A. MANI, H. SUNDARAM, et al. (2005). *Context-Aware Dynamic Presentation Synthesis For Exploratory Multimodal Environments*. Arts Media and Engineering, ASU, AME-TR-2005-02, Jan. 2005.

[14] H. SUNDARAM and S.-F. CHANG (2001). *Condensing computable scenes using visual complexity and film syntax analysis*, Proc. IEEE International Conference on Multimedia and Expo, pp. 273-276, Aug. 2001, Tokyo, Japan.