How Speech/Text Alignment Benefits Web-based Learning

Sheng-Wei Li
Dept. of Computer Science and Information Engineering
National Chi-Nan University, Puli, Nantou, Taiwan 545, R.O.C.
886-049-2910960#4864
kennylee@mclab.csie.ncnu.edu.tw

Hao-Tung Lin
Dept. of Computer Science and Information Engineering
National Chi-Nan University, Puli, Nantou, Taiwan 545, R.O.C.
886-049-2910960#4864
haotung@mclab.csie.ncnu.edu.tw

Herng-Yow Chen
Dept. of Computer Science and Information Engineering
National Chi-Nan University, Puli, Nantou, Taiwan 545, R.O.C.
886-049-2910960#4843
hychen@csie.ncnu.edu.tw

ABSTRACT
This demonstration presents an integrated web-based synchronized scenario for many-to-one cross-media correlations between speech (an EFL, English as Foreign Language, lecture with free-style lecturing behaviors) and the corresponding textual content. The analysis/presentation of the temporal correlations enable the vivid web-based language learning through the interactive functions: browsing speech via content, word-by-word pointer guidance, synchronized scrolling/highlighting, and listening training mode. We regularly analyze and repackage the multimedia content of VoA (Voice of America) [1], ICRT (International Community Radio Taipei) [2], and Online Lectures in our University [3]. Through the subjective experiments, this repackaged synchronized speech/text content does facilitate the learning for EFL learners.

Categories and Subject Descriptors
K.3.1 [**Computers and Education**]: Computer Uses in Education – *computer-assisted instruction, distance learning.*

General Terms
Design and Human Factors.

Keywords
Cross-Media Correlation, Analysis and Presentation, Speech-to-Text Alignment, Lips Sync, Computed Synchronization

1. INTRODUCTION
With the rapid development of multimedia technology, we can retrieve many free online learning resources with speech/text. In fact, the text and corresponding spoken content are implicitly related in temporal domain. If we can compute the implicit relation between speech and text, we could repackage the learning resources in a vivid style. Further, we can add-on many interactions to facilitate listening comprehension. The implicit temporal relation is classified as simple relation and complex relation model. Broadcasting news is the typical example of the simple relation model. Figure 1 shows the concept of the simple mapping relation where speech and text are almost one-to-one temporally related. The simple relation has been addressed by a recursive global alignment method proposed by Pedro et al. [1] and got the accuracy of 94%. A recording of real lecturing is the typical example of the complex relation model. Figure 2 shows the complex many-to-one relation. The lecture comprises stages: full-text-recitation, single-sentence-recitation, and comment where each stage may contain the same textual content. The many-to-one temporal relation deepens the difficulty of computing. We develop an iterative local alignment method to compute the many-to-one relation and got time accuracy of 99.4%.

In 2002, W.T. at al. [2] proposed a system presenting the computed synchronization of simple model, but only achieved sentence-based synchronization. In this paper, we construct a more interactive and word-based synchronized system based on the computed many-to-one relation. Through this enhanced application, we do provide more auxiliaries for web-based learners.

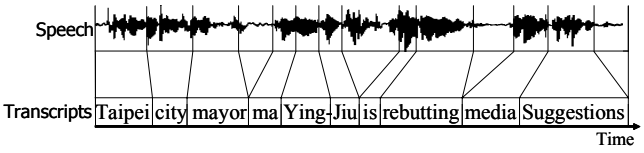

Figure 1. The simple implicit relation model

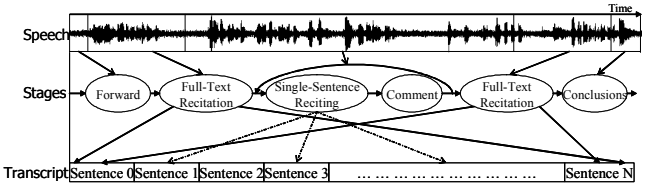

Figure 2. The Complex Implicit Relation Model.

2. SYSTEM OVERVIEW
After we got the implicit temporal relation, we constructed a fully visualized interface which repackages the learning resources and provides highly interactive functions for language learners. Figure 3 shows the complete system architecture and features. The functions include:

- **The dynamic text highlighting** – through the temporal relation computed by the alignment process, presentation system can highlight the text that is currently spoken with different levels of granularity (such as word and sentence) according to the presentation time of speech. This feature helps learners keep pace with the progress of speech playback.

MM'05, November 6–12, 2005, Singapore.
ACM 1-59593-044-2/05/0011.

- **Tele-pointer guidance** – a virtual pointer is rendered to simulate the traditional blackboard experience.
- **Region judgment/analysis** – through analysis of the alignment results, some logical lecturing stages can be defined (such as the full text recitation stage or singe sentence recitation/explanation).
- **Multiple level content accessing** – to browse the speech in our interface, we can use the annotated region in timeline, sentence, or word of interests. This function can facilitate the learning process and make the process more efficient.
- **Training mode** – to help learners to train the listening ability, we design the switching function to choose the visibility of content.
- **Automatic scrolling** –While the content of some time point is out of display region, the display region will be scrolled automatically and smoothly as possible for better human perception.
- **Lips-Sync animation** – this is an auxiliary assistance by synthesizing the mouth shapes of individual phonemes in a word of the content. With the time stamps of each word, a virtual talking head with lips animation can be presented on the web.

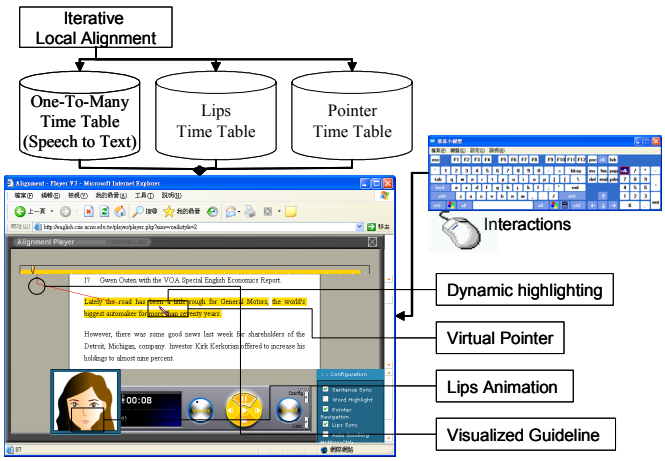

Figure 3. System overview

3. EXPERIMENTS

We discuss the experimental results in two aspects: objective and subjective. In the objective evaluation, we have regularly analyzed the VoA and ICRT data for a while and got the result that 99.3% explored content are of time skew less than 0.5 sec. The average precision rate is 92.5%.

In the subjective aspect, we refer the result proposed by Steinmetz [6] to survey the human perception on the proposed system. We classify three time regions:

• *Hardly-detected* Region: This region ranges from about -60 ms to 80ms (-45ms to +0 for lips-to-speech). In this region, very few people detect the synchronization errors about the text highlighting and the pen navigation. However, the percentage of detected sync errors about the lips synchronization is much higher than that of the text and pointer synchronization.

• *Transient* Regions: There are two such regions; one of which ranges from about -60ms to -125ms (-50ms to -80ms for lips-to-speech), and the other is between 80ms and 115ms (0ms to 60ms for lips-to-speech). In these regions, the slight non-smooth playback of media can easily be detected by human.

• *Out-of-Sync* Regions: The regions that exceed -125ms and 115ms are in out-of-sync region (exceed -80ms and 60ms for lips-to-speech). In this region, almost all people feel uncomfortable about the high sync-skew of visual elements. In this region, that human are distracted by the "out of sync" defects.

4. CONCLUSION

In this paper, we described a complex relation probing model and proposed an intuitive, but feasible and efficient, iterative algorithm to tackle the chaotic data set. Our heuristic based on the local alignment can probe more information than that of the original alignment algorithm but with the same time complexity *O(nm)*. After the analysis process, we construct a Web multimedia player to show the feasibility of an integrated vivid presentation from the probed hidden correlation. The integrated presentation is constructed based on the investigations about human perception to provide better QoE (Quality of Experience) for users. These achievements do give a better vivid presentation between speech/text/lip and provide a good interactive/access environment with full control functions for online learners. The system demonstration is available online at: http://media.csie.ncnu.edu.tw/prototype/kennylee/.

5. REFERENCES

[1] VoA: http://www.voanews.com

[2] ICRT: http://www.icrt.com.tw

[3] WSML: http://english.csie.ncnu.edu.tw

[4] Pedro J. Moreno, Chris Joerg, Jean-Manuel Van Thong and Oren Glickman (1998), A Recursive Algorithm for The Forced Alignment of Very Long Audio Segments. International Conference on Spoken Language Processing (ICSLP), Sydney (Australia).

[5] Wei-Ta Chu & Herng-Yow Chen (2002) Cross-Media Correlations: a case study of navigated hypermedia documents. ACM MM, p.p. 57-66.

[6] Ralf Steinmetz (1996), Human Perception of Jitter and Media Synchronization, IEEE Journal on Selected Areas in Communication, Vol. 14, No. 1, pp. 61-72.