Speech/Text Alignment in Web-Based Language Learning

Sheng-Wei Lee
National Chi-Nan University, Taiwan, R.O.C.

Hao-Tung Lin
National Chi-Nan University, Taiwan, R.O.C.

Herng-Yow Chen
National Chi-Nan University, Taiwan, R.O.C.

INTRODUCTION

After the extensive development of network technology and explosive progress of computing, it is now possible that instructors and students need not meet at the same place to begin the classroom experience. In recent years, there have been many efforts devoted to developing effective online learning systems, including tele-teaching with automated support and learning with presentation of vivid classroom experiences. As to the automated support for teaching, the primary issues lie in a robust, ubiquitous computing application to support capturing everything including teaching behaviors, multimedia authoring, and content generation. In practice, the ability to record every thing or event such as a mouse movement, clicking, and typing during a tele-presentation or a class is the key factor to reconstructing *vivid classroom experiences*. Most of the previous studies focused on the issues such as explicitly recording the synchronized replay (Abowd, 1999; Muller & Ottmann, 2000), but few efforts looked at the special properties of speech and pre-prepared transcript. The explicitly recorded media streams can be audio, video, slides, and whiteboard, and all streams have the property of time dependency. Time dependency is that each stream should be synchronized with a global clock when played back. However, with regard to language lectures or broadcasting news program, there are temporal relations that have existed between speech and text and need not be recorded explicitly. Such an existing but hidden relationship is regarded as an implicit relation; the relation recorded explicitly by automated supporting tool is an explicit relation. Explicit relations mean that the correlations between different media are pre-orchestrated as a scheduled scenario or could easily be captured by a recording tool. The Synchronized Multimedia Integrated Language (SMIL) enables simple authoring of interactive audiovisual presentations (W3C, 2004) and an SMIL-based document defines the playback scenario, including temporal, spatial, and content information to present multiple media. In contrast, implicit correlations are usually hidden and therefore could not easily be determined by a simple detecting process, so further computational analyses are needed to discover them. Suppose that the implicit relation between content and speech can be analyzed and that the structure of the recorded speech is also extracted. We can then design a friendly interface for such multimedia document navigation that is much more convenient than the traditional VCR-like navigation mechanism. The speech/text alignment is the tool that analyzes the implicit relation of time between speech and text content. The proposed Web-based Synchronized Multimedia Lectures (WSML) (Chu & Chen, 2002; Chu, Hsu, & Chen, 2001) system can exploit the combination of *implicit relation* (analysis) and *explicit relation* (capture) to provide an effective integrated presentation for teaching English as Second Language (ESL) lectures and broadcasting news programs.

BACKGROUND

The analysis, namely alignment, applied to speech/text is to compute and explore the hidden relation between speech and text. And the explored hidden relation is a key to constructing a better time-based presentation to benefit Web-based learning. In essence, the presentation for online lectures can be classified as two categories: Replay of Recorded Synchronization and Playback of Computed Synchronization

Replay of Recorded Synchronization

Replay of *Recorded Synchronization* is the synchronization of all the time-based media steams explicitly captured by the automated supporting tools to compose a tele-presentation of online lectures. Meanwhile, the manual synchronization achieved by manually adding temporal information is also one instance of Recorded Synchronization. The goal of such Recorded Synchronization is to faithfully reproduce the instruction process of a real classroom experience. Hence, the perceptual similarity between the replay of recorded synchronization and the original real instruction process is an important criterion for evaluating the capturing and playback system. Such application indeed has a significant impact on online lectures, but the criticism is that this scenario puts considerable pressure on instructors to produce as many comprehensive recorded lectures as possible. Most of the e-learning system developers focus on the area of Recorded Synchronization but ignore yet another important factor for the presentation— the existing but implicit correlation between the language text lectures and the speech.

Playback of Computed Synchronization

In practice, there have been existing hidden relations between the recorded speech of an instruction process and the corresponding text language lecture. *Computed Synchronization* is used to analyze the different media streams (e.g., text and speech, in our case) and discover the temporal clue for synchronization between multiple media streams. Although few studies have paid attention to such *implicit relation*, it remains helpful for Web-based language learning. Using this *implicit relation*, we can construct a vivid synchronized multimedia presentation with speech playback, dynamic text-highlighting, and simulated tele-pointer animation, even if no explicit annotation events are captured beforehand. The Web-based Synchronized Multimedia Lectures system has utilized this implicit relation of language lectures and of broadcasting news to provide learners with synchronized presentation of language lectures.

For some perfect cases like lecture recitation and broadcasting news based on transcript, this computation can produce a very detailed temporal index between speech and text, and then facilitate cross-media retrieval. However, in some lectures containing weaker speech/text phonetic correlation (e.g., the explanation of individual vocabulary), the complete word-level temporal index can not be precisely determined. On the other hand, although the explicit relation explorer does not produce the temporal relation that is implicit between speech and text, we incorporate it with the implicit relation explorer to compensate the missed temporal region for a more complete and better presentation of language lectures. For the Web-based presentation of language lectures, the combination of computed synchronization and recorded synchronization can provide more benefits to the learners.

THE METHODS FOR SPEECH/TEXT ALIGNMENT

There have been several solutions proposed to analyze the implicit correlations between speech and text. As with most other methods adopted, we designed our alignment algorithm by dynamic programming strategy. Before describing our method, we first examine the structures and correlations between speech and text content.

STRUCTURES/CORRELATIONS

The process of teaching English as a Second Language (ESL) lecture is typically composed of the five stages: forward, full-text recitation, single-sentence-recitation, comment, and conclusion. It can be clearly observed that three key stages—the full-text-recitation, single-sentence-recitation, and comment stages—occupy most of the teaching time. Similarly, the scenario in which an anchorperson recites news transcript also follows a particular structure. We depict those two structures in Table 1.