Automatic Construction of CALL System from TV news Program with Captions

Takashi Tanaka* Kazumasa Mori* Satoshi Kobayashi** Seiichi Nakagawa*

*Department of Information and Computer Sciences, **Computer Center, Toyohashi University of Technology
takashi@slp.ics.tut.ac.jp kmori@slp.ics.tut.ac.jp koba@slp.ics.tut.ac.jp nakagawa@slp.ics.tut.ac.jp

Abstract

Many language learning materials have been published in Japan. However, they are limited in their scope and content. In addition, we doubt whether the speech sounds found there are natural in various situations. These days, some TV news programs (by NHK, CNN, ABC[1], etc.) have closed/open captions corresponding to the speech of the announcer. We have developed a system that automatically makes CALL (Computer Assisted Language Learning) materials from such captioned newscasts. Materials compiled by this system have the following functions: repetition listening, consulting an electronic dictionary, and automatic construction of a dictation test. The materials have the following advantages: polite and natural speech sound, various and timely topics, and abundant materials.

In this paper, we describe the organization of our new system.

1. Introduction

Many language learning materials on listener training have been published, and the research on listening has continued [2,3,4]. However, the topics offered listening exercises (for example, learning the differences between /r/ and /l/ in English) have been limited. In addition, the speech sound used in the material is often uttered in virtual situations. Thus, it sounds unnatural.

Various language learning materials (CALL, video, tapes, etc.) can be made from TV and radio broadcasts. Learning materials made from the mass media [5,6] are suitable for repetition training, but the topics used have been limited in number and scope.

For each teacher to put together learning materials from newscasts is very time consuming. When learners themselves record newscasts, it allows them only to practice repetition listening.

In language learning, although repetition is obviously necessary [7], it is difficult to maintain the learner’s interest using existing learning materials, because available topics have been so limited. By contrast, it seems that a learner can possibly maintain his or her interest in foreign language learning, if the learning materials can offer various and timely topics such as daily news events. However, it takes much time to make learning materials that embrace a wide range of topics. Consequently learning materials must be produced automatically.

In this study, we developed a system that automatically makes CALL materials from TV news programs with closed/open captions. We show the outline and characteristics of the system in Section 2. Procedures to construct CALL materials are proposed in Section 3, while Section 4 gives an alignment method to synchronize captions with speech sounds. Section 5 shows the functions of this CALL system. Finally, Section 6 shows the prospects of this system.

2. Outline and characteristics

Since March 2000, NHK (Nippon Housou Kyoukai: Japan Broadcasting Corporation) has been broadcasting TV news with closed captions corresponding to the announcer’s delivery on “News 7” [8]. It is possible to record newscasts and captions daily. In the present study, we suggest a system that makes CALL materials from newscast pictures, speech sounds, and captions.

However, delays occur between the actual newscast and the display of the captions as shown in Figure 1, in which the topic of the upper frame (TV image) is “telecommunication”, while that in the lower frame (display captions) is “weather” that was a former topic of “telecommunication”, because it needs time to recognize speech by a speech recognizer, and to confirm and correct the recognition results by operators. In addition, the captions are broadcast as teletext unlike video and sounds as in the NHK newscasts. Therefore, it is not possible to use the captions as a learning device only by recording the newscast normally, in other words, it is difficult to make learning materials using those videos and captions.

In this study, we used a teletext tuner board to record captions, and the voice and captions are synchronized to adjust for the delay.

In addition to that, we developed CALL material using a computer. Compared with commercial learning materials, this system has the following advantages.
Figure 1: Actual newscast image
Upper frame: TV Image; topic of “telecommunication”
Lower frame: Closed Captions; topic of “weather”

- Polite and natural utterances of news announcers in the real world
- Various and timely topics
- Capability to collect a variety of topics
- Saving much time in producing learning materials
- Automatic construction of dictation tests

A range of topics in particular helps the learner to maintain his or her interest.

This system consists of two units. One is the CALL player, called “The Language Learning Player or LLP”. Another is the synchronizer; this unit performs the synchronization of captions and speech sounds to adjust for the delay.

The LLP has mainly the following functions.

- Display video image and captions corresponding to the speech
- Repetition listening on sentence by sentence
- Rewind and fast-forward
- Consulting a Japanese-English dictionary [9]
- Various caption’s display styles (Chinese characters, Japanese KANA characters, etc.)
- Display of waveform and F0 contours of the voice (announcer’s and learner’s voices)
- Automatic construction and grading of dictation tests

3. Procedure to construct learning material

The construction and processing flow of this system to make CALL materials from captioned TV news are shown in Figure 2.

1. When recording voices, pictures, captions, two TV tuner boards are employed to record the TV news with captions. One records the video pictures and those speech sounds. The other is a teletext board for recording the captions.
2. A CALL developer checks acquired captions for error. If errors are detected, operators correct them, and then divides captions into sentence units.
3. Next, a morphological analysis [10] is done, the results of which are used to generate sequences of syllables.
4. Feature vectors of recorded speech are extracted.
5. Afterwards, using the speech recognition system [11], the time alignment between the sequences of syllables and speech is carried out. Then the synchronous information is acquired from this result.
6. The timing of captions are adjusted by this synchronous information.
7. Finally, we can use the LLP as a CALL.

We use the XML format to represent captions with synchronous information for easy comprehension and to extend this system easily as shown in Figure 3.

These procedures can be done automatically except for 2 and 5.

Procedure 5 is very difficult to do automatically[1] because of some speech sounds like filled pauses and repairs are not transcribed.
4. Synchronization of captions and speech

The speech analysis condition is shown in Table 1. In this system, concatenated syllables based on HMMs (Hidden Markov Model) which correspond to the pronunciation of captions are matched with speech sound by using one pass Viterbi algorithm. Captions are divided into sentences, and the matching is carried out sentence by sentence. The utterance speed is assumed as 5 mora/sec at latest, and the synchronizer picks up a part of speech sound corresponding to the sentence’s length, and then makes a forced alignment between the picked up sound and the sentence (i.e., the syllable sequence based on HMMs).

Decision on each sentence’s boundary of speech sound is carried out by the following method.

At first, the synchronizer gets local N maximums of likelihood from the result of a forced alignment between the sound and the first sentence (the likelihoods are denoted by \( P_a^1, \ldots, P_a^N \)), and the corresponding frames are \( \{ T_a^1, \ldots, T_a^N \} \), where \( N \) is integer. In our experiment, \( N=5 \).

Then the synchronizer makes a forced alignment between the next sentence and the next sound from each local maximum and seeks a maximum of likelihood (the next sound’s start frames are \( \{ T_b^1, \ldots, T_b^N \} \) and corresponding maximums are \( \{ P_b^1, \ldots, P_b^N \} \)).

Out of these likelihoods, the maximum likelihood of sum of \( P_a^i \) and \( P_b^i \) is found and its corresponding frame \( T_a^i \) is considered a candidate of the first sentence’s boundary.

This procedure is repeated until the end of captions.
5. **Details of LLP**

In this section, we describe some functions of the LLP.

5.1. **Replay**

Video images, speech sounds and captions are replayed sentence by sentence, thus facilitating repetition listening of any part. In addition, rewind and fast-forward are also possible. Changing the replay speed is also possible by controlling the sampling rate.

5.2. **Captions**

The following caption display styles are available.

- Only Japanese KANA characters (for beginners)
- Mixture of Chinese characters and Japanese KANA characters (for intermediate)
- Specific parts of speech such as only nouns or keywords (for intermediate)
- Without captions (for advanced)

The system will can automatically control these caption display styles in accordance with the level of learners.

5.3. **Phonation support**

This LLP can display the waveform and F0 of both announcer’s and learner’s voices. Using this, the learner by himself or herself can compare his/her utterance speed and accent with those of the announcer. However, it is not easy for the learner to distinguish his or her voice waveform from the announcer’s voice waveform.

We have a plan to add a function to evaluate the learner’s pronunciation\[2,3\] and to display hints for improving his or her phonation.

5.4. **Consulting a dictionary**

The LLP has a function for consulting a Japanese-English dictionary. Learners can choose a word in a caption and push “dictionary” button, then the LLP displays the word or phrase in English.

5.5. **Dictation**

The LLP can construct and grade listening tests. With this function, some words specified by part of speech in the captions (for example, noun) are displayed as blanks, and a window opens for entering the answers. Results are displayed with the correct answers.

In the future, we will incorporate a function to construct tests of various types, and a function to help teachers.

6. **Summary**

This paper described a system that automatically makes CALL materials from captioned TV news programs. Such materials cannot only replay news pictures and sounds, but also have some functions as a CALL system. In addition, this system saves much time in producing learning materials.

However, this system has not been completed, and some problems remain. For example, automatic drawing up of a curriculum suited for learners, and automatic scoring of a learner’s phonation are as yet unavailable as a CALL material.

Since the synchronization accuracy between caption and speech sounds is still poor, the algorithm must be improved.

We are adapting this system for Chinese and Korean students to learn Japanese, and developing a CALL system for Japanese students to learn English by using English captioned TV news.

**Acknowledgement**

We are grateful to NHK Science & Technical Research Laboratories for allowing us to use NHK’s newscast programs.

**References**


