An automatic construction system of CALL materials 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

{ttakashi, kmori, koba, nakagawa}@slp.ics.tut.ac.jp

Abstract

It is said that learning materials made from the mass media such as TV programs are suitable for language learning, and also it is necessary to assign a lot of exercises to a learner. Therefore a system which constructs a learning materials automatically has been eagerly anticipated. We have developed a system that makes Computer Assisted Language Learning (CALL) materials from closed-captioned TV programs semi-automatically: Japanese version and English version. The learning materials constructed on this system have many features and functions comparable to many learning materials on the market.

This learning material was evaluated by some language teachers and some students. The results states that this learning material brings attractive usage in a school lesson.

1. Introduction

Many language learning materials have been published, and the research on listening has continued[1,2,3]. In language learning, although repetition training is obviously necessary[4], it is difficult to maintain the learner's interest/motivation using existing learning materials, because those materials are limited in their scope and contents. In addition, we doubt whether the speech sounds used in most materials are natural in various situations. It is said that learning materials made from the mass media such as TV programs and video tapes are suitable for language learning[5,6], but the topics of the materials have been limited in number and scope too. Each teacher has to make learning materials from these mass media by himself, but its process is very time consuming. Nowadays, some TV news programs (CNN, ABC, PBS, NHK, etc.) have closed/open captions corresponding to the announcer's speech. We have developed a system that automatically makes Computer Assisted Language Learning (CALL) materials for both English and Japanese from such captioned newscasts. This system computes the synchronization between speech and the caption by using acoustic models of HMMs and a forced alignment algorithm, and the materials (both English and Japanese) compiled by this system have the following functions: full/partial text caption display, repetition listening, consulting an electronic dictionary, display the user/announcer sound contour, and automatic generation of a dictation test. The materials have the following advantages: polite and natural speech sound, various and timely topics, and abundant materials. And the materials have the following possibilities: automatic creation of listening/understanding tests, and storage/retrieval of the many materials. We are now evaluating this learning material and augmenting the functions as a CALL material.

We describe the outline and characteristics of the system in Section 2. Procedures to construct CALL materials are explained in Section 3. Section 4 gives an alignment method to synchronize captions with speech sounds and the assessment of the accuracy. Section 5 describes the functions of this CALL system. Section 6 shows the evaluation of this system and the results. Finally, Section 7 states 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 which are corresponding to the announcer's delivery on "News 7" [7]. It is possible to record newscasts and captions daily. In the present study, we suggest a system that makes CALL materials from those 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 preceding topic of "telecommunication", because it needs time to recognize speech by a speech recognizer, and to confirm and correct the recognition results by NHK's 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, in Japanese version, we used a teletext tuner board to record captions, and the voice and captions are synchronized to adjust for the delay for NHK news program.

On the other hand, in English version, we used an open capture or Web script of PBS news program[8]. For the former case, we typed the caption into a computer by hand. The caption is covered by a black-image as shown in Figure 2.

In addition to that, we developed CALL material using a computer. Compared with commercial learning materials, this system has the following advantages.

- Polite and natural utterances of news

announcers
- Various and timely topics
- Capability to collect a variety of topics
- Saving much time in producing learning materials
- Automatic generation 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, speech sound and captions
- Repetition listening on sentence by sentence
- Rewind and fast-forward
- Consulting a dictionary
- Various caption's display styles (untouched caption, pronunciation, only noun etc.)
- Display of waveform and F0 contours of the voice (announcer's and learner's voices)
- Automatic construction and grading of dictation tests

Figure 1: Actual newscast image (Japanese)
Upper frame: TV Image; topic of "telecommunication"
Lower frame: Closed Captions; topic of "weather"

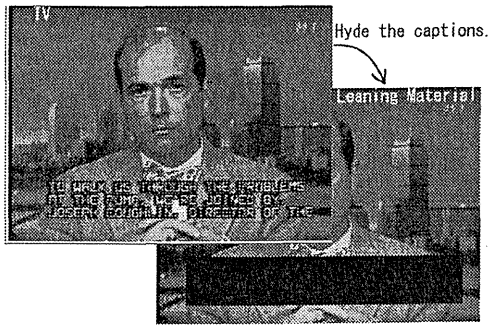

Figure 2: Actual newscast image (English)

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 3.

1. When recording voices, pictures, captions, two TV tuner boards are employed to record the TV news and 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, an operator corrects them.
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 techniques[11], the time alignment between the sequences of syllables transcribed from captions and the speech sounds is carried out. Then the synchronous information is acquired from this result.
6. The timing of each word is adjusted by this synchronous information, and this information is outputted into a file which is readable by the LLP.
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 4.

These procedures can be done automatically except for 2.

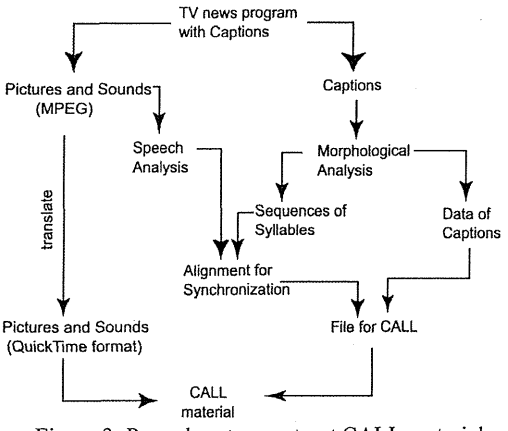

Figure 3: Procedure to construct CALL material

```
<CAI>
<OPTION LANGUAGE="english"/>
<VIDEO src="cai_data¥¥English¥¥AV¥¥anthrax1.mov"/>
<SOUND src="cai_data¥¥English¥¥AV¥¥anthrax1.raw"/>
<CAPTIONS>
<CAPTION start="0" end="6432">
<WORD POS="Adverb" NORMAL="Now" PRONOUNCE="n aw" START="0" END="364">Now </WORD>
<WORD POS="," NORMAL="," PRONOUNCE="" START="364" END="364">, </WORD>
<WORD POS="TO" NORMAL="to" PRONOUNCE="t uw" START="364" END="393">to </WORD>
<WORD POS="Article" NORMAL="some" PRONOUNCE="s ah m" START="393" END="552">some </WORD>
<WORD POS="Preposition" NORMAL="of" PRONOUNCE="ah v" START="552" END="595">of </WORD>
<WORD POS="Article" NORMAL="the" PRONOUNCE="dh ah" START="595" END="662">the </WORD>
                    :
```

Figure 4: Example of XML file for captions and time synchronous information

4. Synchronization of captions and speech

In this system, the synchronization between captions and speech sound is carried out by a Viterbi decoding algorithm between a sequence of syllable-based HMMs (phoneme-based HMMs for English) corresponding to the caption and the speech sound. To acquire a sequence of syllables from captions, we use the morphological analyzer (ChaSen) for Japanese contents or the POS analyzer (Brill's Tagger) for English contents. The speech analysis condition is shown in Table 1.

Table 1:Speech Analysis Condition

Sampling frequency	12kHz
Window function	Hamming (21.33msec.)
Frame period	8msec.
LPC analysis order	14th
Feature vectors	LPC mel-cepstrum (10dim.) +ΔCEP(10dim.) +ΔΔCEP(10dim.) +ΔPOW +ΔΔPOW
Acoustic Model	Syllable/Phoneme based HMMs with 4 mixture full covariance matrix

To measure a degree of accuracy of this synchronization, we evaluated the accuracy at ending times of noun words, by the measure of a margin of error. In Japanese contents, the average (number of samples: 39) of a margin of error was achieved to 14ms. And in English contents, the average (number of samples: 25) of a margin of error was 92ms. An SNR of Japanese speech sounds was 30~33dB, but an SNR of English speech sounds was 16~20dB because there were large back-noise. Furthermore a training data of English HMM was not enough. It is thinkable that these were the main factors of worse results for English speech data.

5. Details of LLP

In this section, we describe some functions of the LLP (See Figure 5).

Figure 5: Language Learning Player(main window)

5.1. Replay

Video images, speech sounds and the captions are replayed sentence by sentence, thus facilitating repetition listening of any part. Of course, rewind and fast-forward are also possible. In addition, each word is highlighted in red when the word is uttered on the video.

5.2. Captions

The following caption display styles are available.

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

A user can select these caption display styles in accordance with his/her level.

5.3. Phonation support

This LLP can display the waveform and F0 contours of both announcer's and learner's voices. Using this, the learner by oneself can compare one's utterance speed and accent with those of the announcer. Addition to it, boundary of each word was connected from user's one to announcer's one with a straight line. This makes easy to compare timings between user's and announcer's voices.

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 (See Figure 6).

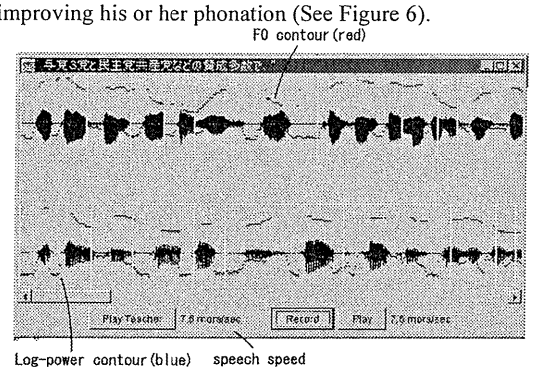

Figure 6: LLP (phonation support)
upper part: Announcer's speech sounds
lower part: User's speech sounds

5.4. Consulting a dictionary

The LLP has a function for consulting a dictionary (Japanese-English, English-Japanese and Japanese-Chinese). This LLP offers two ways to consult a dictionary. One is popup the meaning below the word when a user moves cursor to the word closer. The other is if a user click the word on a caption, then the LLP opens a dictionary window and shows the meanings in detail. (See Figure 7).

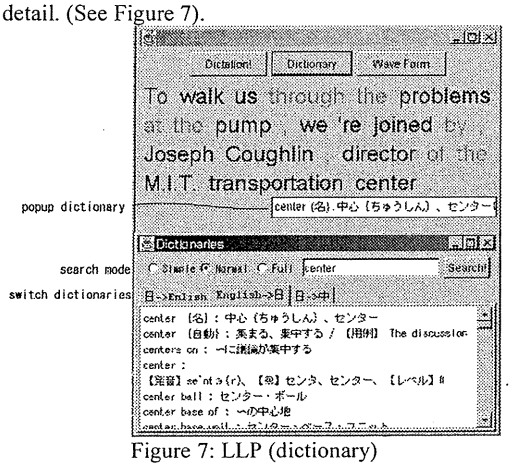

Figure 7: LLP (dictionary)

5.5. Dictation

The LLP can generate listening tests and grade them. In this function, some words specified by a POS in the captions (for example, noun) are displayed as blanks or selectable items, and user can type these words by keyboard or select from selectable items. And then graded results are displayed with effective sounds. If these results are perfect, then the LLP plays a pleasant ring of bell (See Figure 8).

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

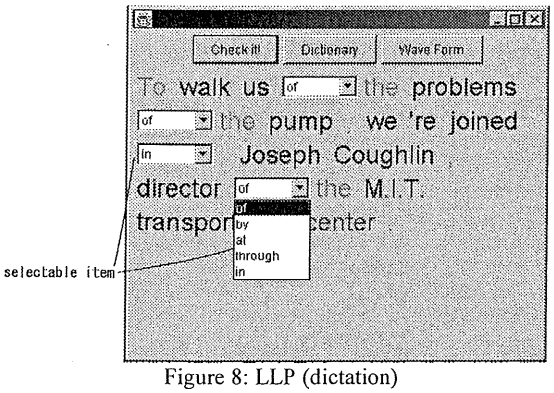

Figure 8: LLP (dictation)

6. Evaluation

This section describes the evaluation result of the learning materials created by this system. For English version system, 2 Japanese teachers teaching English and 5 Japanese students evaluated the English learning materials. All Japanese students are poor English ability. For Japanese version system, 2 Japanese teachers teaching Japanese and 6 foreign students evaluated the Japanese learning materials. The 6 foreign students consist of 3 Chinese and 3 East-West Asian. The evaluation was carried out in two times. After, at the first time, subjects used the CALL system for 30 minutes, at the second time, subjects also used the CALL system for 30 minutes, and then scored this learning material on a 1 to 5 basis for many aspects.

In conclusion, the results states that (1) synchronization accuracies are pretty good, (2) automatic dictation tests are useful, (3) displaying waveform and F0 contour is a bad score, because many trial subjects could not understand these acoustical meanings. Therefore we have to simplify displaying of these or teach these acoustical meaning for user in advance. And as these learning materials are made from TV news programs, the announcer uses a large vocabulary and speaks so fast. The beginners, therefore, are hard to listen and understand them.

7. 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 CALL materials. In addition, this system saves much time in producing learning materials.

This system is now under-development, and some problems remain. For example, automatic drawing up of a suitable curriculum for learners, and automatic scoring of a learner's phonation are not yet available.

Acknowledgement

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

References

[1] Y. Taniguchi, A. A. Reyes, H. Suzuki and S. Nakagawa: "An English conversion and pronunciation CAI system using speech recognition technology", EUROSPEECH, pp. 705-708, 1997

[2] N. Minematsu, Y. Fujisawa, S. Nakagawa: "Evaluation of Japanese manners of generating word accent of English based on a stressed syllable detection technique", ICSLP, pp. 3103-3106, 1998

[3] "Investigation of CALL Softwares (in Japanese) ", http://www2.slp.tutics.tut.ac.jp/CALLsoft/

[4] Tracey M., Darwing: "Information type and its relation to non-native speaker comprehension", Language learning, 39, pp.43-48, 2000

[5] "BBC Online Education", http://www.bbc.co.uk/education/home/

[6] "SIM super Elmer CBS course", Source Corporation, Tokyo SIM Foreign Lab, 1997

[7] A. Ando, T. Imai, A. Kobayashi, H. Isono, and K. Nakabayashi, "Real-Time transcription system for simultaneous subtitling of Japanese broadcast news programs", IEEE Trans. Broadcasting, Vol. 46, No. 3, pp.189-196, 2000.

[8] "PBS news hour", http://www.pbs.org/newshour

[9] "Eijiro", http://www.alc.co.jp/eijiro/index.html

[10] "Japanese Morphological Analyzer System ChaSen", http://cl.aist-nara.ac.jp/lab/nlt/chasen

[11] A. Kai, S. Nakagawa: "Investigation on Unknown Word Processing and Strategies for Spontaneous Speech Understanding", EUROSPEECH, pp.2095-2098, 1995.