Japanese Broadcast News Transcription and Information Extraction

Developing a system for close-captioning and automatic information extraction for Japanese broadcast news speech.

Sadaoki Furui, Katsutoshi Ohtsuki, and Zhi-Peng Zhang

Inspired by the activities within the DARPA research community, we have been developing a large-vocabulary, continuous-speech recognition (LVCSR) system for Japanese broadcast news speech transcription [4]. This is a part of a joint research project with NHK broadcasting whose goal is the closed-captioning of TV programs. While some of the problems that we have investigated are Japanese-specific, others are language independent.

The broadcast news manuscripts used for constructing our language models were taken from NHK news broadcasts over a period between July 1992 and May 1996, and comprised roughly 500,000 sentences and 22 million words. To calculate word *n*-gram language models, we segmented the broadcast news manuscripts into words by using a morphological analyzer since Japanese sentences are written without spaces between words. A word-frequency list was derived for the news manuscripts, and the 20,000 most frequently used words were selected as vocabulary words. This 20,000-word vocabulary covered approximately 98% of the words in the broadcast news manuscripts. We calculated bigrams and trigrams and estimated unseen *n*-grams using Katz's back-off smoothing method.

The feature vector consisted of 16 cepstral coefficients, normalized logarithmic power, and their delta features (derivatives). The total number of parameters in each vector was 34. Cepstral coefficients were normalized by the cepstral mean subtraction (CMS) method.

The acoustic models were gender-dependent shared-state triphone hidden Markov models (HMMs) and were designed using tree-based clustering. They were trained using phonetically balanced sentences and dialogues read by 53 male speakers and 56 female speakers. The total number of training utterances was 13,270 for male and 13,367 for female, and the total length of the training data was approximately 20 hours for each gender. The total number of HMM states was approximately 2,000 for each gender, and the number of Gaussian mixture components per state was four.

News speech data, from TV broadcasts in July 1996, were divided into two parts, a clean part and a noisy part, and were separately evaluated. The clean part consisted of utterances with no background noise, and the noisy part consisted of utterances with background noise. The noisy part included spontaneous speech such as reports by correspondents. We extracted 50 male utterances and 50 female utterances for each part. Each set included utterances by five or six speakers. All utterances were manually segmented into sentences. Due to space limitations, we report only the results for the clean part here.

Reading-dependent language modeling. Japanese text includes a mixture of three kinds of characters: Chinese characters (Kanji) and two kinds of Japanese characters (Hira-gana and Kata-kana). Each Kanji has

We constructed a language model that depends on the readings of words in order to take into account the frequency and context-dependency of the readings.

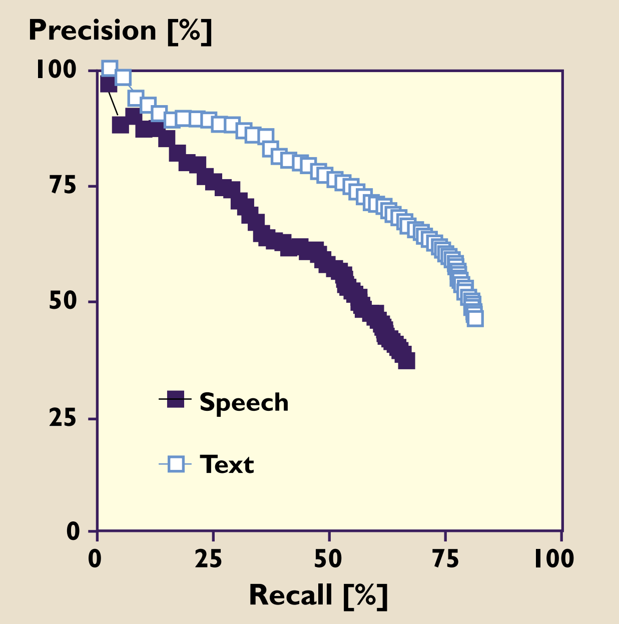
Figure 1. Determining the amount of information borne by the word w_i in a particular new article.

$$SgScore(W_i)=g_i \bullet \log\frac{G_A}{G_i}\,(i=1,2,\ldots,N)$$

Figure 2. Summation of word frequency.

$$G_A=\sum_i G_i$$

Figure 3. Topic-word extraction from broadcast news speech or news text.



multiple readings, and correct readings can only be decided according to context. Conventional language models are built using the written forms of words, and usually equal probability is assigned to all possible readings of each word. This method causes recognition errors in Japanese speech recognition because the assigned probability is sometimes very different from the true probability. We therefore constructed a language model that depends on the readings of words in order to take into account the frequency and context-dependency of the readings. This model reduced the word error rate by 4.4%, averaged over male and female, relative to the results of the conventional language model.

Filled-pause modeling. Broadcast news speech includes filled pauses at the beginning and in the middle of sentences, which cause recognition errors in our previous language models that use news manuscripts written prior to broadcasting. To cope with this problem, we introduced filled-pause modeling into the language model. This model reduced the word error rate by 7.9% relative to the results of the reading-dependent language model.

Online incremental speaker adaptation. Since each speaker in broadcast news utters several sentences in succession, the recognition error rate can be reduced by incrementally adapting the acoustic models within a segment that contains only one speaker. We applied online, unsupervised, instantaneous and incremental speaker adaptation combined with automatic detection of speaker changes. The maximum likelihood linear regression (MLLR) [1] and vector-field smoothing (VFS) [2] methods were instantaneously and incrementally carried out for each utterance.

We tried to determine the appropriate number of transformation matrices, or clusters, for MLLR. MLLR with seven clusters (silence, consonants, and five Japanese vowels) achieved the best performance in the experiment. This method reduced the word error rate by 11.8% relative to the results for the speaker-independent models.

By incorporating all the preceding methods, we achieved an 11.9% word error rate averaged over males and females for clean parts of broadcast news speech, which was a 25.1% reduction in word error rate over the baseline results.

Topic extraction. Transcribed broadcast news speech can also be used for automatic indexing or summarizing of the news. We have been investigating methods for automatically extracting a set of topic words expressing the content of each news broadcast [3]. Since most of the topic words were nouns, topic words are extracted from nouns in the speech recognition results on the basis of a significance measure for each word. Many of the measures that have been used in information retrieval from text databases were tried in a preliminary experiment, and the measure shown in Figure 1 was chosen. Figure 1 gives the amount of information borne by the word w_i in the particular news article. In Figure 1, N is the vocabulary size (nouns only), g_i is the frequency of word W_i in a news article, G_i is the frequency of word w_i in all news articles, and G_a is the summation of all G_i's (as shown in Figure 2).

The Nikkei newspaper articles over a five-year period were used for calculating G_i and G_a values. Twenty-nine broadcast news articles comprising 142 utterances by 15 male speakers (8 anchors and 7 others) were used for evaluation. Each news article had 2–14 utterances (5 utterances on the average). True topic-words were given by three subjects; 4–10 phrases were given for each news article by each subject. A true topic-word set was constructed for each article from topic-words given by at least one subject (35.7 words on average). Supplementary experiments were also conducted by giving correct texts instead of transcription results as input. Eighty-nine percent of the true topic-word set were nouns. Figure 3 shows the results averaged over the 29 news articles. It is observed that, if transcription results are used as input, precision as well as recall are reduced by roughly 10% in comparison with that obtained by using the correct texts as input. When five topic-words were chosen from speech recognition results (recall = 13%), 87% of them were correct on average.

References

1. Leggetter, C.J., et al. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Computer Speech and Language*, (1995), 171–185.
2. Ohkura, K., et al. Speaker adaptation based on transfer vector field smoothing with continuous mixture density HMMs. In *Proceedings of ICSLP'92*, (1992), 369–372.
3. Ohtsuki, K., et al. Improvements in Japanese broadcast news transcription. In *Proceedings of the DARPA Broadcast News Workshop*, (1999), 231–236.
4. Ohtsuki, et al. Recent advances in Japanese broadcast news transcription. In *Proceedings of Eurospeech'99*, (1999), 671–674.

Sadaoki Furui (furui@cs.titech.ac.jp) is a professor in the Department of Computer Science at Tokyo Institute of Technology.
Katsutoshi Ohtsuki (ohtsuki@nttspch.hil.ntt.co.jp) is a research engineer with the Media Processing Project at Cyber Space Laboratories in Kanagawa, Japan.
Zhi-Peng Zhang (zzp@cs.titech.ac.jp) is a doctoral candidate in the Department of Computer Science at Tokyo Institute of Technology.

This work is supported in part by the International Communications Foundation. The authors would like to thank NHK (Japan Broadcasting Corporation) for providing us with the broadcast news database and T. Fuchi at NTT Cyber Space Laboratories for letting us use their morphological analysis program. The authors are also grateful to Nihon Keizai Shimbun Incorporated for allowing us to use the newspaper text database (Nikkei CD-ROM 90-94) for our research.