Intelligent Transcription System based on Spontaneous Speech Processing

Tatsuya Kawahara
Academic Center for Computing and Media Studies, Kyoto University
Sakyo-ku, Kyoto 606-8501, Japan
kawahara@i.kyoto-u.ac.jp
http://www.ar.media.kyoto-u.ac.jp/

Abstract

With the improvement of the speech recognition technology, semi-automatic generation of transcripts or document records of lectures and meetings has become one of its promising applications. For this purpose, we need to take into account post-processing that includes cleaning of verbatim transcripts and segmentation into sentence and paragraph units. This article presents a novel statistical framework for an intelligent transcription system. The recent progress of automatic speech recognition of lectures and meetings is also reported. Then, several approaches to sentence unit detection and disfluency detection are described, as they are significant in the post-processing of transcripts generated by the speech recognizer.

1. Introduction

Speech has been one of the most fundamental means of communication by which human beings have exchanged knowledge and opinions. Even today, with the prevalence of e-mails and the internet, new ideas are discussed and important decisions are made primarily based on speech communications at seminars and meetings. Speech communication, however, is “volatile” in nature, and thus must be recorded. Recently, speech media can be stored digitally as is, but records are usually saved in the form of text for easy browsing and search.

Speech-to-text systems, or automatic speech recognition systems, have been investigated extensively for long years. Most of these studies, however, have focused on human-machine interfaces such as dictation systems. On the other hand, automatic transcription of spontaneous speech, such as human-to-human communication, is far more difficult because of the large variation in both acoustic and linguistic characteristics. Moreover, spontaneous speech processing requires the development of a different paradigm, in that faithful transcription is not necessarily useful because of the existence of disfluencies and the lack of sentence and paragraph markers. Utterances are made while thinking during interactions, and thus the disfluency is inevitable. Actually, cleaning of the transcript is performed by human stenographers in the making of records of lectures and meetings. This process involves the correction of colloquial expressions to document-style expressions. Speech is a time-dimensional signal, and so a transcript is simply a sequence of words, which corresponds to a text without punctuation or line-breaks. In spontaneous Japanese in particular, sentences are easily concatenated without explicit endings. Thus, the following issues must be addressed:

- deletion of disfluencies and redundant words
- correction of colloquial expressions and recovery of omitted particles
- segmentation of sentences and paragraphs

In making lecture notes and meeting minutes, it is also necessary to extract important sentences and compress them for a summary. Automatic speech summarization has also become an important research topic. However, this article focuses on the generation of speech transcripts that are both faithful and readable. Specifically, we have been investigating the development of automatic transcription systems for lectures and meetings, which can be used for the generation of records of lectures and meetings[1]. Applications also include the next-generation transcription system for the Japanese Diet (Congress).

The remainder of this article is organized as follows. The proposed intelligent transcription system is described in Section 2. Section 3 summarizes the recent progress of automatic speech recognition for lectures and meetings. Sections 4 and 5 describe approaches to sentence unit detection and disfluency detection, respectively. These issues are also addressed in the MetaData Extraction (MDE) task under the DARPA EARS project[2][3][4]. Finally, future areas for consideration are discussed in Section 6.

2. Intelligent Transcription System

2.1. System Overview

An overview of the proposed intelligent transcription system, which combines the cleaning post-process with the conventional automatic speech recognition, is illustrated in Figure 1. We adopt the framework of statistical machine translation (SMT) to transform a verbatim transcript V into a document-style text W. In a similar formulation to the automatic speech recognition, the proposed framework decomposes the posteriori probability $p(W|V)$ into two probabilities by the Bayes rule. The language model probability $p(W)$ can be reliably trained using an enormous number of documents, including newspaper articles and Web texts, whereas the translation probability $p(V|W)$ must be estimated with a parallel corpus of smaller size that aligns verbatim transcripts of utterances V and cleaned texts for documentation W.

We extend this framework to estimate the language model probability $p(V)$ for speech recognition, as below, by considering that the training text size for $p(W)$ is much larger than the size of verbatim transcripts needed for training $p(V)$.

$$p(V) = p(W) \cdot \frac{p(V|W)}{p(W|V)}$$

Here, $p(V|W)$ and $p(W|V)$ are estimated with the same parallel corpus. The probability $p(V|W)$ models the generation process of the spontaneous utterances, whereas the probability $p(W|V)$ is used for the cleaning process.

2.2. Analysis using the Diet Corpus

For a parallel corpus that aligns verbatim transcripts and document-style texts, we have compiled the "Diet Corpus" by transcribing the actual utterances made in the Lower House of the Japanese Diet (Congress) and aligning them with the official meeting records. The current size of the corpus is approximately two million words in text or 150-hour speech. For the morphological analysis, we used Chasen 2.2.3 with IPADIC 2.4.4. We investigated the difference in the transcripts of actual utterances V and the document-style texts W observed in this corpus. The transformation process from the former V to the latter W is classified into deletions, insertions and substitutions.

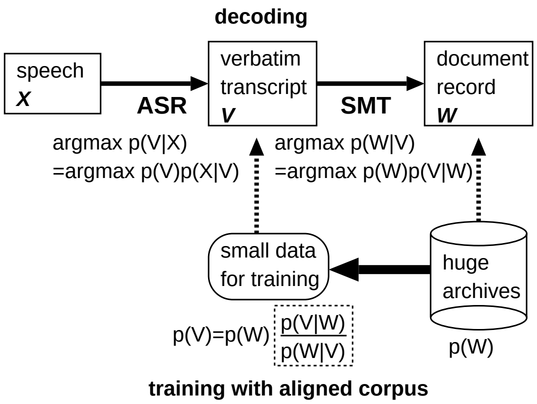

Figure 1. Overview of proposed transcription system

A summary of the statistics for the three categories and typical, or the most frequent, examples thereof are shown in Table 1. The differences in either category are observed for 11% of the words in total. The majority of these differences concern the deletion of redundant words. These words include not only fillers but also several end-of-sentence expressions, such as "*desune*" and "*to*". False starts and portions of self-repairs are also deleted, but their lexical patterns vary widely. On the other hand, the most insertions to documented texts are functional words and verb suffixes, '*i*', such as "*shi-te-(i)-ru*" and "*ki-te-(i)-ru*". The substituted words are related to colloquial expressions.

2.3. Style Conversion of Language Model

A direct approach to statistical modeling of these phenomena involves counting the frequencies of conversion patterns. However, in the cleaning process, redundant words, such as fillers, are always deleted, and colloquial expressions are always substituted, so that $p(W|V)$ for these patterns is set to be 1. The other patterns specified in Table 1 are to be estimated using the corpus. Since the translation probability apparently depends on neighboring words, the context-dependent modeling is desirable, but the data sparseness problem is encountered. Therefore, we introduce a Part-

Table 1. Major differences between spontaneous speech and document-style text

	$p(W\|V)$	$p(V\|W)$	frequency	examples
deletions of redundant words	1	estimate	8.5%	*ee, desune, ano, maa*
insertions of missing particles	estimate	estimate	1.0%	*i, wa, wo*
substitutions of colloquial expressions	1	estimate	1.8%	*toiu/teiu, keredomo/kedomo*

Of-Speech-based (POS-based) contextual model as a back-off model[5].

The size of the official records of the Diet is so large (for example 71M words for four years worth of documents) that the reliable statistics are estimated for the document-style model $p(W)$ and are then transformed to the verbatim model $p(V)$.

3. Spontaneous Speech Recognition

We have also been intensively studying automatic speech recognition (ASR) of lectures and meetings. The recent progress is described in this section. Here, we used the Corpus of Spontaneous Japanese (CSJ)[6][7] as a primary corpus as well as the Diet corpus.

3.1. Acoustic Modeling

Spontaneous speech has greater variation both in spectral and temporal structures than read-style speech. As the acoustic variation is largely dependent on speakers, feature normalization techniques such as vocal tract length normalization (VTLN)[8][9] are effective. Speaker adaptive training (SAT)[10][11] scheme has also been investigated.

3.2. Pronunciation Variation Modeling

The phonetic variation caused by spontaneous utterance can also be modeled in a pronunciation dictionary, in which a list of possible phone sequences for each word is defined. While the orthodox pronunciation forms are referred to as baseforms, the variants observed in spontaneous speech are called “surface forms”. These surface form entries are often derived from speech data by aligning them with phone models[12]. In the CSJ, actual phonetic (*kana*) transcription is given manually, so the set of surface forms is easily defined. However, the simple addition of surface form entries results in the side-effect of false matching. Thus, effective but constrained use of these surface forms are necessary.

An approach is statistical modeling, which is similar to language modeling. Namely, the unigram probability of each pronunciation form is assigned and is multiplied by the language model probability in decoding. In this case, the statistical framework of speech recognition is re-formulated as:

$$w' = \arg\max_{w,p} P(x|p)P(p|w)P(w)$$

Here, $P(p|w)$ is the pronunciation probability of surface form p for word w, while $P(x|p)$ and $P(w)$ represent conventional acoustic model probability and language model probability for input x, respectively. We investigated the comparison of statistical models using the CSJ, and concluded that cutoff of less frequent surface forms is crucial, and that the unigram model is effective, whereas the trigram model has a marginal gain[13].

When the surface form is derived for word units, it is dependent on the task and corpus, and is not necessarily applicable to different tasks. Phone-based modeling of pronunciation variation is more general and portable to various lexicons. Surface forms are obtained by applying such a model to phone sequences of baseforms. We proposed generalized modeling of subword-based mapping between baseforms and surface forms using variable-length phone context[14]. The variation patterns of phone sequences are automatically extracted together with their contexts of up to two preceding and following phones, which are decided by their occurrence statistics. A set of rewrite rules are then derived with their probabilities and variable-length phone contexts. The model effectively predicts pronunciation variations depending on the phone context using a back-off scheme. The model was applied and evaluated with two transcription tasks, domains of which are different from the training corpus (CSJ). The effects of the predicted surface form entries and their probabilities are individually evaluated, and each was found to have a similar impact on overall performance.

3.3. Language Model

Language model training of spontaneous speech is much more difficult than that for dictation systems, which can make use of huge language resources such as newspaper articles and Web pages. Most of the available language data are written text, and are mismatched with the spoken-style. For language modeling of spontaneous speech, a great deal of transcription is essential, but has a huge cost.

The most widely-used solution to enhance language model training data is to combine with or interpolate other existing text databases, which are not necessarily spontaneous speech corpora, but are related to the target task domain. These include proceedings of lectures, minutes of meetings, and closed captions for broadcast programs. Recently, the World Wide Web has become a major language resource, and not a few Web sites contain spoken-style documents, such as records of lectures and meetings. We also studied a method to filter spoken-style texts from the Web pages[15].

As described in Section 2, we propose a novel “translation” approach that estimates language model statistics (N-gram counts) of spontaneous speech from a document-style large corpus based on the framework of statistical machine translation (SMT)[5]. The translation is designed for modeling characteristic linguistic phenomena in spontaneous speech, such as insertion of fillers, and estimating their occurrence probabilities. These contextual patterns and probabilities are derived from a small parallel aligned corpus of faithful transcripts and their documented records. This method was successfully applied to the estimation of the language model for the Diet meetings from their minute archives.

3.4. Adaptation of Acoustic Model

Since the variation of acoustic features is very large in spontaneous speech, speaker adaptation of the acoustic model is effective and is almost essential. Although there are many factors affecting the acoustic characteristics of spontaneous speech such as speaking rate and speaking styles, speaker adaptation is a simple solution to handle all of these factors in an implicit manner. The acoustic model adaptation also involves channel adaptation, that is, the characteristics of rooms and microphones are also normalized.

In particular, in lectures and meetings, each speaker makes many utterances in the same session. Thus, a considerably large amount of data is available to conduct unsupervised adaptation in a batch mode, where the initial speech recognition result with the speaker-independent model is used for adaptation of the acoustic model, which is then used for rescoring or re-decoding. Standard adaptation techniques such as maximum likelihood linear regression (MLLR) are used, and filtering the reliable recognition hypotheses with confidence measures can also be incorporated.

3.5. Adaptation of Language Model

Adaptation of the language model is also important to deal with a variety of topics and speaking styles. In lectures and meetings, the topic is focused and consistent throughout the entire session. Therefore, language model adaptation is feasible even in an unsupervised or batch mode, as in the acoustic model adaptation, and computationally expensive methods can be allowed in off-line transcription tasks.

The simplest method is to construct an N-gram model from the initial speech recognition result and interpolate it with the baseline model. We also investigated methods to select the most relevant texts from the corpus based on the initial recognition result[13]. As a criterion for text selection, we used the tf-idf measure and perplexity by the N-gram model generated from the initial speech recognition result, and demonstrated that they have comparable and significant effects in improving speech recognition performance.

The cache model[16] and trigger model[17] weigh the probability of words recently used in the utterances or talk, or those directly related to the previous topic words. We also proposed a trigger-based language model adaptation method oriented to meeting transcription[18]. The initial speech recognition result is used to extract task-dependent trigger pairs and to estimate their statistics. This method achieved a remarkable perplexity reduction of 28%.

Recently, latent semantic analysis (LSA) which maps documents into implicit topic sub-spaces using the singular value decomposition (SVD) has been investigated extensively for language modeling[19]. A probabilistic formulation, PLSA[20], is powerful for characterizing the topics and documents in a probabilistic space and predicting word probabilities. We proposed an adaptation method based on two sub-spaces of topics and speaker characteristics[21]. Here, PLSA was performed on the initial speech recognition result to provide unigram probabilities conditioned on the input speech, and the baseline model is adapted by scaling N-gram probabilities with these unigram probabilities. The method was applied to automatic transcription of panel discussions and was shown to be effective.

3.6. Speech Recognition Performance

A summary of the current speech recognition performance for the CSJ (lectures) and the Diet (Budget Committee meeting) is given in Table 2. By combining

Table 2. Speech recognition performance for lectures & Diet meetings (word accuracy)

method	lecture	Diet
baseline	76.6	74.1
+ speaker normalization	78.3	76.7
+ acoustic model adaptation	81.2	80.5
+ language model adaptation	81.8	81.5

the adaptation of acoustic and language models using the initial speech recognition result, the word accuracy was improved to around 82% for both evaluation tasks.

4. Sentence Boundary Detection

Detection of the sentence unit is vital for linguistic processing of spontaneous speech, since most of the conventional natural language processing systems assume that the input is segmented by sentence units. Sentence segmentation is also an essential step to key sentence indexing and summary generation.

In spontaneous speech, especially in Japanese, in which subjects and verbs can be omitted, the unit of the sentence is not so evident. In the CSJ, therefore, the clause unit is first defined based on the morphological information of end-of-sentence or end-of-clause expressions. The sentence unit is then annotated by human judgment considering syntactic and semantic information.

Two approaches to automatic detection of sentence boundaries are described in the following subsections.

4.1. Statistical Language Model (SLM)

In fluent speech or read speech of well-formed sentences, it is possible to assume that long pauses can be interpreted as punctuation marks, and the insertion of periods (=sentence boundaries) or commas can be decided by the neighboring word contexts.

Thus, the baseline method makes use of an N-gram statistical language model (SLM) that is trained using a text with punctuation symbols, in order to determine a pause to be converted to a period. Specifically, for a word sequence around a pause, $X = (w_{-2}, w_{-1}, pause, w_1, w_2)$, a period is inserted at the place of the pause if $P(W_1) = P(w_{-2}, w_{-1}, period, w_1, w_2)$ is larger than $P(W_2) = P(w_{-2}, w_{-1}, w_1, w_2)$ by some margin. Actually, this decoding is formulated as the maximization of a likelihood $\log P(W) + \beta * |W|$, where $|W|$ denotes the number of words in W and β is the insertion penalty widely used in speech recognition.

In spontaneous speech, however, the approaches that rely heavily on the pauses are not successful. Speakers put pauses in places other than the ends of sentences for certain discourse effects, and disfluency causes irregular pauses (=interruption points), while consecutive sentences are often continuously uttered without a pause between them.

Therefore, we introduce a more elaborate model that selects possible sentence boundary candidates by considering contextual words (w_{-1} and w_1). Specifically, if these words match typical end-of-sentence expressions, a sentence boundary is hypothesized regardless of the existence of a pause, and if they match non-typical end-of-sentence expressions, a sentence boundary is hypothesized only if there is a pause. The hypotheses of sentence boundaries are verified using the N-gram model. The method is also formulated as a statistical machine translation framework[22] and is referred to as enhanced SLM.

4.2. Support Vector Machines (SVM)

A simpler but more general approach is to treat the pause duration as one of the features in addition to the lexical features, and feed them into a machine learning framework. We adopted support vector machines (SVM) because there are a wide variety of cue expressions suggesting sentence endings in Japanese. In this case, sentence boundary detection is regarded as a text chunking problem[23], and we adopt the IE labeling scheme, where I and E denote inside-chunk and end-of-chunk, respectively. For every input word, a feature vector is composed of the preceding and the following three words, together with their POS tags and the durations of the subsequent pauses, if any. The pause duration is normalized by the average in a turn or a talk, because it is affected by the speaking rate and significantly different between speakers. Dynamic features or estimated results of preceding input parts can also be fed into SVM. The SVM is considered to be powerful for handling a very large number of features and finding the critical features called “support vectors”.

4.3. Experimental Evaluations

Here, we present the results evaluated in the CSJ[24]. The test-set was that used for speech recognition evaluation and consists of 30 presentations or 71K words in total. Both SLM and SVM described in the previous subsections were trained with the Core 168 presentations of 424K words, excluding the test-set. In this experiment, we used automatic speech

Table 3. Results of sentence unit (boundary) detection in CSJ

	recall	precision	F-measure
SLM (text)	79.2	84.6	81.8
SLM (ASR)	70.2	71.6	70.9
SVM (text)	83.0	87.9	85.4
SVM (ASR)	73.9	81.7	77.6

recognition (ASR) results without conducting speaker adaptation and the word error rate was approximately 30%. The results are summarized in Table 3, where the recall, precision and F-measure are computed for sentence boundaries. Correct transcripts (text) are used for reference. SVM realizes significantly better performance than SLM, and is even more effective in the speech recognition case, demonstrating robustness for erroneous input. SVM is directly trained to classify boundaries, whereas SLM measures the linguistic likelihood of sentence boundaries. Moreover, features used in SVM are independent of each other and classification succeeds only if a key feature (support vector) is correctly detected, whereas a single error affects the likelihood of word sequences in the N-gram model. It is noteworthy that performance degradation by using speech recognition is much smaller than the word error rate.

4.4. Further Extension

Another approach for further improvement is to incorporate higher-level linguistic information, such as syntactic dependency and caseframe structures. We studied an interactive framework of parsing and sentence boundary detection, and showed that dependency structure analysis can help sentence boundary detection and vice versa[25]. However, the conventional dependency structure analysis does not necessarily work reliably and robustly for spontaneous speech having ill-formed sentences and disfluencies, especially for erroneous transcripts generated by speech recognition systems. Therefore, we propose a more robust method that is based on local syntactic dependency of adjacent words and phrases[26].

We also investigate the detection of quotations, which is similar to sentence boundary detection, but involves analysis of very complex sentence structures[27].

5. Disfluency Detection

Disfluency is another prominent characteristic of spontaneous speech. Disfluency is inevitable because humans make utterances while thinking about what to say, and the pipeline processing is often clogged. Thus, the detection of disfluencies may be useful for analyzing the discourse structure or speaker's mental status. However, disfluencies should be removed for improving readability and applying conventional natural language processing systems including machine translation and summarization.

Disfluency is classified into the following two broad categories:

- fillers (such as “um” and “uh”), including discourse markers (such as “well” and “you know”), with which speakers try to fill pauses while thinking or to attract the attention of listeners.

- repairs, including repetitions and re-starts, where speakers try to correct, modify, or abort earlier statements.

Note that fillers usually appear in the middle of repairs.

Lexical filler words are usually obtained as the output of the speech recognition system, and their recognition accuracy is much the same as that of ordinary words. However, there are a number of words that also functions as non-fillers such as “well” in English and “*ano*” in Japanese. For these distinctions, prosodic features will be useful since we can recognize fillers even for non-familiar languages. Previously, we investigated the difference in prosodic features in these words in Japanese[28].

On the other hand, the detection of self-repairs involves much more complex processes. The most conventional approach is to assume the repair interval model (RIM)[29], which consists of the following three parts in order:

`(RPD) ! (DF) (RP)`

(ex.) “I'm going {`RPD`: to Tokyo} ! {`DF`: no} {`RP`: to Kyoto}”

RPD (ReParanDum): portion to be repaired
DF (DisFluency): fillers or discourse markers
RP (RePair): portion to correct or replace RPD

The first step to the self-repair analysis based on this model is to detect DF or interruption points (IP), noted with ‘!’ in the above, which seem relatively easy to spot. DF usually consists of filler words, and IP detection can be formulated in much the same manner as the sentence boundary detection using neighboring lexical features together with prosodic features[30].

In the CSJ or Japanese monologue, however, we observe many cases that do not satisfy this assumption or

RIM. First, DF or filler words are often absent. Second, RPD and RP segments often have nothing in common on the surface level, although they may be semantically related, for example, "*ana* (hole) ! *mizo* (trench) *wa*..." This phenomenon makes it extremely difficult to perform machine learning using lexical features. Actually, using SVM, we obtained a detection accuracy (F-measure) of 77.1% for the cases in which RPD and RP segments have some words in common, but only 20.0% otherwise. This result suggests that high-level semantic information is necessary for further improvement.

6. Conclusions and Future Works

In this article, studies on automatics transcription of spontaneous human-to-human speech communications are described. Specifically, we focused on sentence segmentation and disfluency detection for improved readability of the transcripts. Since errors are inevitable in automatic speech recognition, manual post-processing including error correction is necessary. Therefore, we are developing an interface for human editors to correct and edit the transcripts generated by the automatic speech recognition system[31]. Appropriate segmentation of utterances is essential for the editing process. We are also investigating the degree of speech recognition accuracy that is needed to realize an efficient transcription system, including manual post-processing. The use of multiple candidates and confidence measures output by the speech recognizer should also be explored for this purpose.

One of the next targets will be the note-taking of lectures for handicapped people, which is real-time generation of closed-captions. The process should incorporate compaction or a kind of summarization technique to improve the readability, and the entire process must be performed with a small latency, although the transcripts may not be perfect as long as they are comprehensible.

In the future, integration with other media, such as video, should be studied more extensively. We are investigating this direction for lectures in our university classrooms. Furthermore, integration with knowledge processing must be explored, because speech is a media for exchanging knowledge and is thus a source of knowledge.

Acknowledgments: The work was conducted with the contribution of research associates and graduate students including Dr. H. Nanjo, Dr. Y. Akita, Mr. K. Takanashi, Mr. M. Mimura, Mr. K. Shitaoka and Mr. M. Saikou.

References

[1] T.Kawahara. Spoken language processing for audio archives of lectures and panel discussions. In *Proc. Int'l Conference on Informatics Research for Development of Knowledge Society Infrastructure (ICKS)*, pages 23–30, 2004.

[2] Y.Liu, E.Shriberg, A.Stolcke, B.Peskin, J.Ang, D.Hillard, M.Ostendorf, M.Tomalin, P.Woodland, and M.Harper. Structural metadata research in the EARS program. In *Proc. IEEE-ICASSP*, volume 5, pages 957–960, 2005.

[3] E.Shriberg. Spontaneous speech: How people really talk and why engineers should care. In *Proc. INTERSPEECH*, pages 1781–1784, 2005.

[4] Y.Liu, E.Shriberg, A.Stolcke, D.Hillard, M.Ostendorf, and M.Harper. Enriching speech recognition with automatic detection of sentence boundaries and disfluencies. *IEEE Trans. Audio, Speech & Language Process.*, 14(5):1526–1540, 2006.

[5] Y.Akita and T.Kawahara. Efficient estimation of language model statistics of spontaneous speech via statistical transformation model. In *Proc. IEEE-ICASSP*, volume 1, pages 1049–1052, 2006.

[6] S.Furui. Recent advances in spontaneous speech recognition and understanding. In *Proc. ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, pages 1–6, 2003.

[7] K.Maekawa. Corpus of Spontaneous Japanese: Its design and evaluation. In *Proc. ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, pages 7–12, 2003.

[8] L.Lee and R.C.Rose. Speaker normalization using efficient frequency warping procedures. In *Proc. IEEE-ICASSP*, pages 353–356, 1996.

[9] S.Wegmann, D.McAllaster, J.Orloff, and B.Peskin. Speaker normalization on conversational telephone speech. In *Proc. IEEE-ICASSP*, pages 339–342, 1996.

[10] J.W.McDonough, T.Anastasakos, G.Zavaliagkos, and H.Gish. Speaker-adapted training on the switchboard corpus. In *Proc. IEEE-ICASSP*, pages 1059–1062, 1997.

[11] D.Pye and P.C.Woodland. Experiments in speaker normalisation and adaptation for large vocabulary speech recognition. In *Proc. IEEE-ICASSP*, pages 1047–1050, 1997.

[12] E.Fosler et al. Automatic learning of word pronunciation from data. In *Proc. ICSLP*, 1996.

[13] H.Nanjo and T.Kawahara. Language model and speaking rate adaptation for spontaneous presentation speech recognition. *IEEE Trans. Speech & Audio Process.*, 12(4):391–400, 2004.

[14] Y.Akita and T.Kawahara. Generalized statistical modeling of pronunciation variations using variable-length phone context. In *Proc. IEEE-ICASSP*, volume 1, pages 689–692, 2005.

[15] T.Misu and T.Kawahara. A bootstrapping approach for developing language model of new spoken dialogue systems by selecting web texts. In *Proc. INTERSPEECH*, pages 9–12, 2006.

[16] R.Khun and R.De Mori. A cache-based natural language model for speech recognition. *IEEE Trans. Pattern Analysis & Machine Intelligence*, 12(6):570–583, 1990.

[17] R.Lau, R.Rosenfeld, and S.Roukos. Trigger-based language models: A maximum entropy approach. In *Proc. IEEE-ICASSP*, volume 2, pages 45–48, 1993.

[18] C.Troncoso and T.Kawahara. Trigger-based language model adaptation for automatic meeting transcription. In *Proc. INTERSPEECH*, pages 1297–1300, 2005.

[19] J.R.Bellegarda. A multispan language modeling framework for large vocabulary speech recognition. *IEEE Trans. Speech & Audio Process.*, 6(5):468–475, 1998.

[20] T.Hoffman. Probabilistic latent semantic indexing. In *Proc. SIG-IR*, 1999.

[21] Y.Akita and T.Kawahara. Language model adaptation based on PLSA of topics and speakers. In *Proc. ICSLP*, pages 1045–1048, 2004.

[22] T.Kawahara, M.Hasegawa, K.Shitaoka, T.Kitade, and H.Nanjo. Automatic indexing of lecture presentations using unsupervised learning of presumed discourse markers. *IEEE Trans. Speech & Audio Process.*, 12(4):409–419, 2004.

[23] T.Kudo and Y.Matsumoto. Chunking with support vector machines. In *Proc. NAACL*, 2001.

[24] Y.Akita, M.Saikou, H.Nanjo, and T.Kawahara. Sentence boundary detection of spontaneous Japanese using statistical language model and support vector machines. In *Proc. INTERSPEECH*, pages 1033–1036, 2006.

[25] K.Shitaoka, K.Uchimoto, T.Kawahara, and H.Isahara. Dependency structure analysis and sentence boundary detection in spontaneous Japanese. In *Proc. COLING*, pages 1107–1113, 2004.

[26] T.Kawahara, M.Saikou, and K.Takanashi. Automatic detection of sentence and clause units using local syntactic dependency. In *Proc. IEEE-ICASSP*, page (accepted for presentation), 2007.

[27] R.Hamabe, K.Uchimoto, T.Kawahara, and H.Isahara. Detection of quotations and inserted clauses and its application to dependency structure analysis in spontaneous Japanese. In *Proc. COLING-ACL*, volume Poster Sessions, pages 324–330, 2006.

[28] F.M.Quimbo, T.Kawahara, and S.Doshita. Prosodic analysis of fillers and self-repair in Japanese speech. In *Proc. ICSLP*, pages 3313–3316, 1998.

[29] C.Nakatani and J.Hirschberg. A speech first model for repair detectin and correction. In *Proc. ARPA Human Language Technology Workshop*, pages 329–334, 1993.

[30] Y.Liu, E.Shriberg, A.Stolcke, and M.Harper. Comparing HMM, maximum entropy and conditional random fields for disfluency detection. In *Proc. INTERSPEECH*, pages 3033–3036, 2005.

[31] H.Nanjo, Y.Akita, and T.Kawahara. Computer assisted speech transcription system for efficient speech archive. In *Proc. Western Pacific Acoustics Conference (WESPAC)*, 2006.