論文 / 著書情報
Article / Book Information

題目(和文)	
Title(English)	A Study on Reliability Improvement of Speech Recognizer's Outputs for Spoken Document Processing
著者(和文)	浅見太一
Author(English)	Taichi Asami
出典(和文)	学位:博士(工学), 学位授与機関:東京工業大学, 報告番号:甲第11232号, 授与年月日:2019年6月30日, 学位の種別:課程博士, 審査員:篠田 浩一,徳永 健伸,村田 剛志,藤井 敦,下坂 正倫
Citation(English)	Degree:Doctor (Engineering), Conferring organization: Tokyo Institute of Technology, Report number:甲第11232号, Conferred date:2019/6/30, Degree Type:Course doctor, Examiner:,,,,
学位種別(和文)	博士論文
Type(English)	Doctoral Thesis

A Study on Reliability Improvement of Speech Recognizer's Outputs for Spoken Document Processing

Taichi Asami

January 2018

Abstract

This thesis presents novel methods for improving the reliability of speech recognizer outputs in spoken document processing (SDP) systems.

Chapter 1 describes the background of this study, the details and problems of SDP systems, and approaches investigated in this thesis. The two main approaches are "error rejection" and "feedback and reprocessing," and the sub-approaches of each approach are document-level/word-level confidence measure (CM) estimation and out-of-vocabulary (OOV) word detection. Furthermore, the main concept underlying all proposed methods is described: Global consistency information observed in multiple utterances or multiple spoken documents is essential in improving CM estimation and OOV word detection.

Chapter 2 first briefly explains of the general framework of ASR, and then summarizes conventional methods for CM estimation and OOV word detection and their issues in SDP systems.

Chapter 3 presents a novel document-level CM estimation method based on long-range contextual consistency information. The proposed method formulates contextual consistency by using context windows that cover several consecutive utterances; contextual consistency is the average point-wise mutual information (PMI) between word pairs in each window, and it is used as contextual CM values of the document. A smoothing method that deals with two PMI problems triggered by data sparseness is also proposed. Experiments show that the proposed document-level CM yield high correlation coefficients between CMs and true recognition rates, 0.721. It is also confirmed that the enhanced CMs actually increase the precision of keyword search on spoken documents.

In Chapter 4, an unsupervised word-level CM estimation method that fo-

cuses on consistency information observed in multiple documents is proposed. The issue that conventional methods cannot be applied to SDP systems in practice due to the cost of making human-labeled training data is addressed by a completely unsupervised framework that utilizes transcripts stored in deployed systems instead of human-labeled training data. In order to calibrate the CM of the target word by using consistency of word sequences; similar word sequences existing in the stored transcriptions are extracted as examples. The CM of the target word is updated using the similarity weighted average of the examples. Experiments show that the proposed word CM is superior to conventional word posterior probabilities in terms of rejecting incorrectly recognized words.

In Chapter 5, an OOV word detection method that uses the degree of consistency among multiple occurrences of the same phoneme sequence is proposed. The problem with conventional methods, they raise many false alarms due to disfluencies in spoken documents, is avoided by utilizing the consistency information to distinguish true OOV words from disfluencies. The proposed method first detects recurrent segments, segments that contain the same phoneme sequence in spoken documents by open vocabulary spoken term discovery using a phoneme recognizer. Then, the degree of consistency is measured by using the distribution (mean and variance) of features (DOF) derived from the recurrent segments; the DOF is used as an input for OOV word detection. Experiments illustrate that the proposed method can more robustly detect recurrent OOV words than the conventional method. It is also confirmed that detection performance improves with repetitions of the OOV words.

Reflecting the above work, this thesis makes the following contributions:

- Reliability of speech recognizer outputs for SDP systems is improved by the document-level/word-level CM estimation and the recurrent OOV word detection methods, both of which are founded on global consistency information.

- Both document-level/word-level CM estimation and OOV word detection are technologies that realize error aware systems. This thesis confirms that global consistency information improves CM estimation

and OOV word detection. This means that this thesis reveals key components of the information essential for achieving error awareness: The long-range contextual consistency information observed in multiple utterances for detecting global document recognition errors, the consistency information observed on multiple recognized word examples in multiple spoken documents for detecting incorrect recognition of words, and the consistency information observed in recurrently appearing phoneme sequences for detecting the presence of OOV words.

Acknowledgments

I would like to express my thanks and gratitude to my advisor Professor Koichi Shinoda, Tokyo Institute of Technology, for all of his support, encouragement, and guidance.

I have benefited greatly from my interaction with members of NTT Media Intelligence Laboratories, Nippon Telegraph and Telephone Corporation. There are too many people to mention individually, but I must thank Dr. Satoshi Takahashi, Dr. Yushi Aono, Dr. Hirokazu Masataki, Dr. Satoshi Kobashikawa, Dr. Ryo Masumura, Osamu Yoshioka, Yoshikazu Yamaguchi, and Narichika Nomoto. Without their help, I could not possibly have completed this work.

In addition, I would also like to thank Professor Sadaoki Furui of Toyota Technological Institute at Chicago, and Professor Koji Iwano of Tokyo City University. Without their help, I would not have started my research in the field of speech processing.

Finally, I would like to give my special thanks to my family for all their support over the years.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

The recent drastic progress of automatic speech recognition (ASR) technologies [42, 51] has enabled many kinds of speech processing services, such as voice search, personal voice assistants and automatic dictation. Especially for industrial use, spoken document processing (SDP) systems for call centers have been receiving a lot more attention [10, 46].

A spoken document is a long speech signal that contains particular topic(s). For example, speech signals in phone calls, lectures and consumer videos are spoken documents. SDP systems for call centers extract business intelligence, such as customer's needs, claims and interests, from the massive amount of phone calls being collected. The architecture of a typical SDP system is shown in Figure 1.1. The input spoken document (phone call) is automatically transcribed by ASR, and the transcript is stored in a database. Most call centers store over 10000 transcripts per day. The transcripts are analyzed by text processing functions, such as keyword search, frequent word extraction, or word co-occurrence analysis, for extracting important/useful information. For example, if the name of a particular product co-occurs frequently with the word "broken", the SDP system can immediately detect that the product is a concern.

The analyses of phone calls are conventionally performed by sampling surveys, i.e. human operators randomly pick up several phone calls and

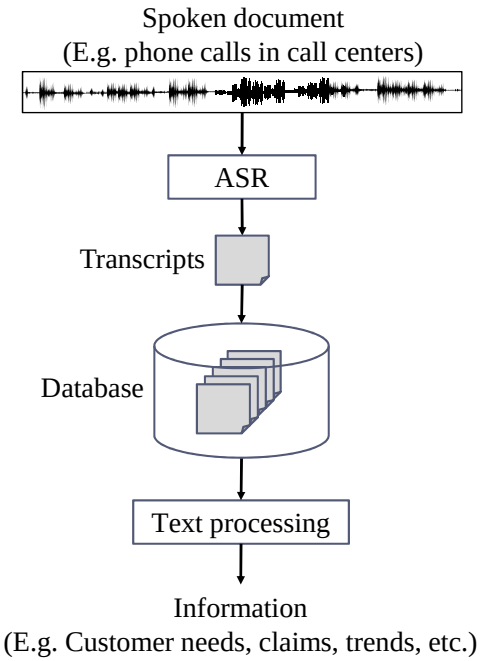

Figure 1.1: Spoken document processing system.

confirm their contents by hearing them. While surveys by human operators can detect precise information, they are too expensive and time consuming. SDP systems are expected to replace this manual process. In terms of SDP system output, high precision outweighs recall.

The problem with SDP systems is that they are prone to extracting *false information*. For example, when an SDP system raises the alert that a particular product name and "broken" co-occur frequently, the word "broken" may be an incorrectly recognized word. Such false alerts degrade the value of SDP systems. Speech recognizer outputs are not reliable enough for SDP systems due to recognition errors [3, 48].

Phone calls are plagued by several factors that can cause recognition errors. The acoustic environment such as noise, microphones and codecs vary widely for each phone call, and speaking style is spontaneous. The transcripts of very noisy phone calls contain many incorrectly recognized words. Even if noise is small, vocal irregularities such as repairs, hesitations and sloppy pronunciations frequently occur in spontaneous speech and cause recognition errors. Moreover, out-of-vocabulary (OOV) words, i.e. words not

contained in the vocabulary of the speech recognizer, also lead to recognition errors. OOV words are never correctly recognized. Unfortunately, important proper names such as the names of people/places/products and technical terms are likely to be OOV words in phone calls since it is impossible to create a vocabulary that covers all such proper names.

The objective of this thesis is to improve the reliability of speech recognizer output and thus achieve more practical SDP systems.

1.2 Approach and problem

Objective, approaches, problems and methods proposed in this thesis are summarized in Figure 1.2. This section describes the approaches, the problems, and provides brief explanations of the proposed methods.

For SDP systems, there are two approaches to deal with recognition errors in transcripts:

a). **Error rejection approach**: Detecting erroneous transcripts or misrecognized words and removing them in order to avoid the extraction of false information. This approach reduces the recall of information extraction but is still viable since precision is important for SDP systems.

b). **Feedback and reprocessing approach**: Adapting the ASR system to the target spoken document and reprocessing it by the adapted ASR system in order to reduce recognition errors directly.

The key technology for the error rejection approach is the confidence measure (CM) [18]; it quantifies the degree of correctness of the speech recognizer output. We refer to the CM of transcripts as document-level CM, and CM of words as word-level CM. If document-level CM accurately measures the recognition rates of each transcript and/or word-level CM accurately measures correctness of each word in transcripts, transcript reliability can be improved by rejecting transcripts/words with low CM values. Creating effective document-level/word-level CMs is the main problem in the error rejection approach.

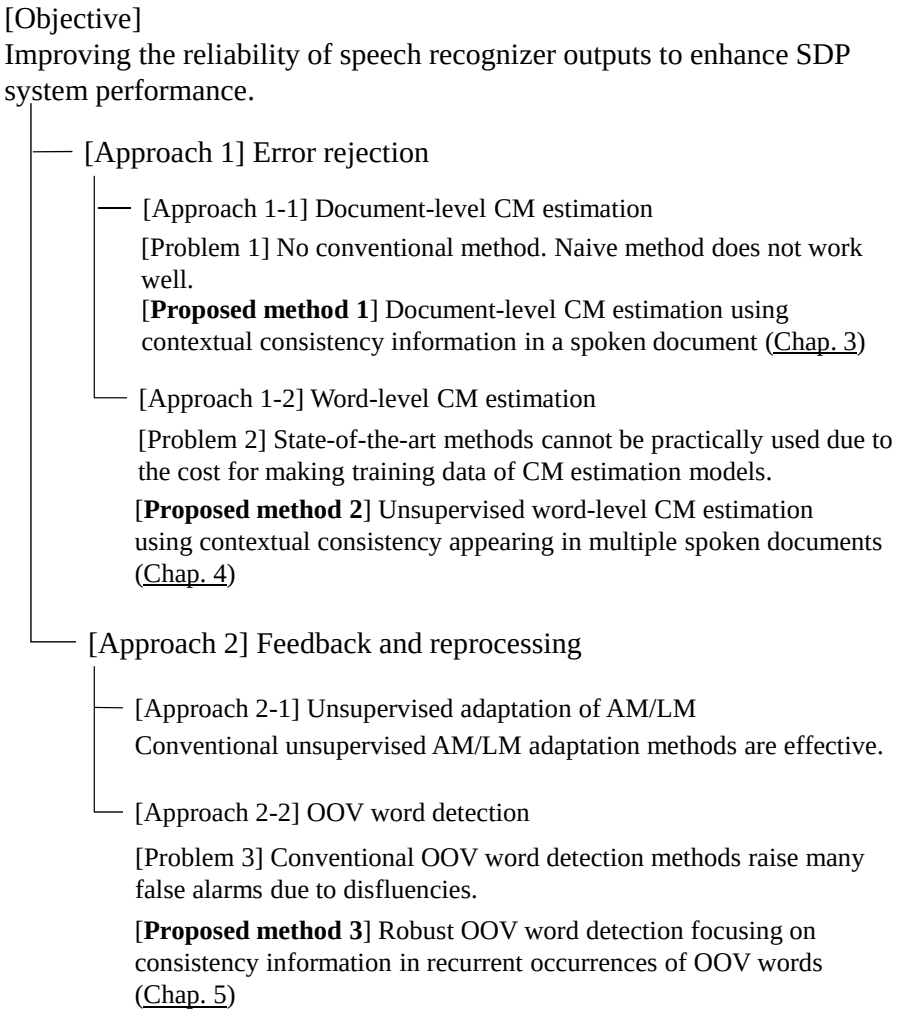

Figure 1.2: Structure of objective, approaches, problems and proposed methods.

The feedback and reprocessing approach first modifies the ASR system so as to reduce recognition errors and then uses the modified system to reprocesses the target spoken documents. A general ASR system has three components, an acoustic model (AM), a language model (LM) and a lexicon (see Section 2.1 for more details about ASR systems). AM and LM can be enhanced by unsupervised adaptation techniques [13, 47] that adapt AM/LM to the target spoken document by utilizing the first transcripts. The lexicon can be enhanced by OOV word detection methods [22, 41] that detect OOV

words uttered in spoken documents and report the presence of OOV words to a system operator. The operator uses the report when adding detected OOV words to the lexicon. AM/LM adaptation and OOV word detection are the main problems in the feedback and reprocessing approach.

While word-level CMs have been widely studied [18], document-level CMs have not been well studied since general ASR systems assume independency between utterances and process utterances on an utterance-by-utterance basis. However, since a spoken document consists of hundreds of utterances, we can utilize multiple utterances to develop a more effective document-level CM. We note that the state-of-the-art word-level CM estimation method uses supervised training as detailed in Section 2.2.2. Though supervised training definitely improves word-level CM performance, it raises the domain dependency problem of CM estimation models. In practice, making training data for CM estimation models for each domain is difficult due to the cost. Accurate word-level CMs that do not use supervised training are required. In this thesis, we propose two novel methods for: 1) document-level CM estimation, and 2) unsupervised word-level CM estimation.

Practical and effective AM/LM adaptation methods applicable to spoken document processing systems have already been proposed [13, 47]. However, existing OOV word detection methods have inadequate detection accuracy. Conventional OOV word detection methods detect sequence of acoustic units (e.g. phonemes) that do not match in-vocabulary (IV) words. However, spoken documents contain many disfluencies such as fillers, repairs, hesitations, and sloppy pronunciations. Since disfluencies often consist of irregular acoustic sequences, conventional OOV word detection methods tend to detect disfluencies as OOV words and so raise many false alarms. In this thesis, we propose a novel OOV word detection method that can robustly detect OOV words from spoken documents containing disfluencies.

1.3 Main idea

Document-level CM estimation, word-level CM estimation and OOV word detection improve the error-awareness. Our key concept undergirding all methods proposed in this thesis is that the consistency information observed

in long contexts of more than one utterance is essential for improving error-awareness. The consistency represents steadfast adherence to some principles or patterns. Our intuition is that humans must utilize similar kinds of consistency information to find errors.

For example, when we see a low quality transcript of a spoken document, we can easily discern that the transcript contains many incorrectly recognized words since the transcript does not make any sense due to contextually/semantically inconsistent word usage across multiple utterances. Contextual consistency information observed in multiple utterances must be considered as an important clue for enhancing document-level CM estimation.

Furthermore, if we know particular word sequences that consistently contain incorrectly recognized words in advance, for example "sun Francisco" ("sun" is incorrect), we can easily judge that another instance of "sun Francisco" also contains an incorrectly recognized word. This information can be obtained from other spoken documents that have already been processed and stored in the database. In this scenario, consistency information observed in multiple documents helps word-level CM estimation.

Finally, when we hear the same phoneme sequence multiple times in a spoken document and cannot understand it consistently, we become aware that the phoneme sequence is probably not contained in our vocabulary. Consistency information observed across a spoken document can be considered as a key clue for enhancing OOV word detection.

We refer to the consistency information observed in ranges of more than one utterance as global consistency information. All methods described in this thesis employ global consistency information to improve their performance.

As described in Section 2.1, the general ASR framework takes, as input, single utterances in isolation, i.e. utterance independency is assumed. However, utterances in spoken documents have dependency due to topics. Furthermore, each SDP system stores multiple documents in the database. Thus, SDP systems can utilize global consistency information obtained from longer range than conventional utterance-wise systems. The proposed methods leverage this property of SDP systems.

1.4 Contribution of Thesis

This thesis makes the following contributions:

- Reliability of speech recognizer outputs in SDP systems is improved by document-level/word-level CM estimation and recurrent OOV word detection methods, both employ global consistency information.
- Both document-level/word-level CM estimation and OOV word detection are technologies that can yield error aware systems. This thesis confirms that global consistency information improves CM estimation and OOV word detection. This means that this thesis reveals three key components of the information needed for achieving error awareness: The long-range contextual consistency information observed in multiple utterances for detecting global document recognition errors, the consistency information observed in multiple recognized word examples in multiple spoken documents for detecting incorrect recognition of words, and the consistency information observed in recurrently appearing phoneme sequences for detecting errors caused by OOV words.

1.5 Outline of Thesis

The remainder of the thesis is organized as follows. Chapter 2 summarizes previous studies of CM estimation and OOV word detection, and points out their issues. Chapter 3 describes the framework for estimating document-level CMs; it utilizes the contextual consistency observed in multiple utterances in a single spoken document. Chapter 4 describes a novel unsupervised word-level CM estimation method that leverages consistency information obtained from multiple spoken documents. Chapter 5 describes an OOV word detection method that utilizes consistency among recurrent appearance of OOV words in a spoken document. Finally conclusions and future work are described in Chapter 6.

Chapter 2

Previous Work

As described in Section 1.2, this thesis explores CM estimation and OOV word detection technologies.

This chapter starts with a brief explanation of the general framework of ASR. Then, representative methods of CM estimation and OOV word detection and their issues are summarized.

2.1 Automatic speech recognition

ASR transforms an input utterance into a word sequence. An utterance is a speech segment that contains human voice, as extracted from speech signals by voice activity detection [44,53]. A single spoken document contains dozens or hundreds of utterances. The ASR in SDP systems processes those utterances on an utterance-by-utterance basis.

The general ASR system formulates the recognition process as follows:

$$\hat{W} = \underset{W}{\operatorname{argmax}} P(W|X), \tag{2.1}$$

$$= \underset{W}{\operatorname{argmax}} P(X|W)P(W), \tag{2.2}$$

where $P(X|W)$ is the likelihood of input utterance X given word sequence W, $P(W)$ is the occurrence probability of word sequence W, and $\hat{W}$ is the recognizer output. $P(X|W)$ and $P(W)$ are called the acoustic likelihood and the language probability, respectively.

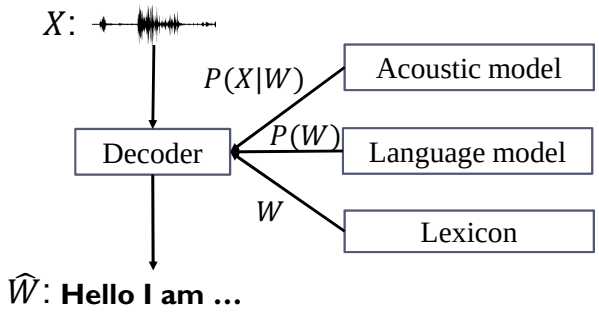

Figure 2.1: The general framework of an ASR system.

Figure 2.1 shows the general framework of an ASR system. The acoustic model (AM) computes the acoustic likelihood $P(X|W)$. Most ASR systems employ the hidden Markov model and deep neural network hybrid AMs to accurately determine the acoustic likelihood [6, 14]. AMs are preliminarily trained by using pairs of speech signals and their manual transcriptions.

The language model (LM) computes language probability $P(W)$. N-gram LM with backoff smoothing techniques [20,21] is almost always adopted. Recently, the recurrent neural network language model is additionally used for precise computation of the language probability [16, 32]. LMs are preliminarily trained on texts whose topics are similar to those of the input speech.

The lexicon is a predefined word set that is expected to be uttered by users. Words in the lexicon are called in-vocabulary (IV) words, and words not in the lexicon are called out-of-vocabulary (OOV) words. Since the decoder searches the word sequences consisting of IV words, OOV words are never correctly recognized.

The decoder performs recognition by using AM, LM and the lexicon. It generates word sequences as result hypotheses from the word set provided from the lexicon and evaluates the result hypotheses by computing $P(X|W)P(W)$. Finally, the decoder outputs the result hypothesis with the highest score as the recognizer output.

Result hypotheses generated in the decoder are represented by a directed acyclic graph, called word lattice. Figure 2.2 shows an example of a word lattice. Each arc has word hypothesis w_i, with acoustic likelihood $P_A(w_i)$ and language probability $P_L(w_i)$. Each node represents the start/end timestamp

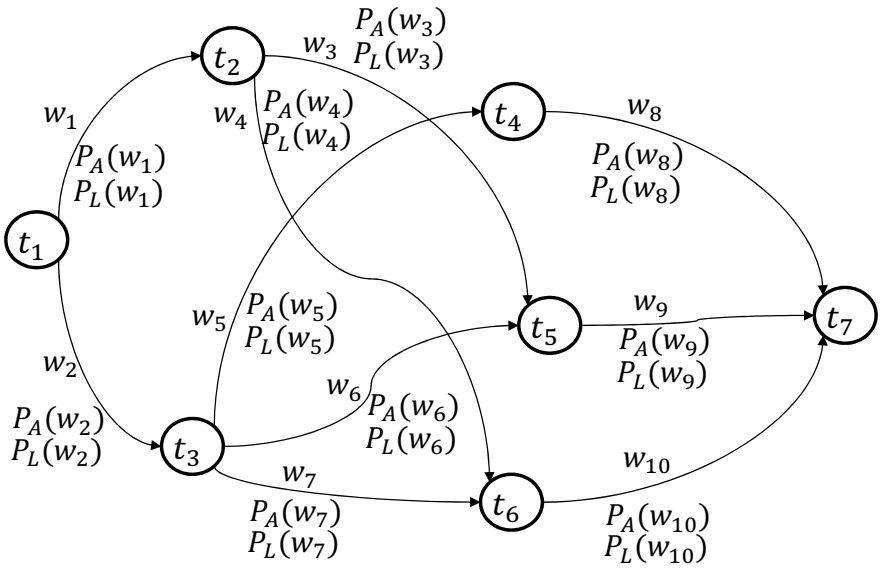

Figure 2.2: An example of a word lattice.

of words, t_j. Each complete path from the beginning node to the end node (t_1 and t_7 in Figure 2.2) represents a single result candidate.

2.2 CM estimation

CM estimation is the technology that estimates the degree of correctness of speech recognizer outputs. Word-level CM estimation predicts the CM of a target word in recognizer outputs; it represents the probability of the correctness of the target word.

Word-level CM estimation techniques have been studied for decades [2, 18, 52]. This section describes the most widely used approach called word posterior probability (WPP), improved CM estimation methods, and their issues that remain to be addressed.

2.2.1 Word posterior probability

WPP [49] is the method most widely usded to compute word-level CMs. WPP, which represents the degree of confusion of the speech recognizer, is computed from the word lattice and does not require any other information.

The posterior probability of a word hypothesis, i.e., arc a, in word lattice G is computed by the following equation:

$$P(a) = \frac{\sum_{C \in G, a \subset C} P(C|G)}{\sum_{C \in G} P(C|G)}, \tag{2.3}$$

where C is a complete path in G, $a \subset C$ denotes that complete path C passes through arc a. $P(C|G)$ is computed as follows:

$$P(C|G) = \prod_{w \subset C} P_A(w) \cdot P_L(w)^{\beta}, \tag{2.4}$$

where $w \subset C$ denotes that word w is included in complete path C; β is the scaling factor of the language score. WPP $P(a)$ can be efficiently computed by a forward-backward algorithm [49] and directly used as a word-level CM. $P(a)$ is also denoted by $P(w)$ where w is the word attached to arc a.

2.2.2 Improved CM estimation

WPP is a fundamental word-level CM and can be computed by using only word lattices, i.e. acoustic and language models. In order to improve CM quality, studies have investigated the use of richer information such as longer context information than N-gram language models. This section summarizes those studies.

Several studies have reported that contextual information is effective in estimating word-level CMs [12, 24]. These techniques are based on the idea that a word that appears to be contextually inconsistent in an utterance is likely to be wrong. The contextual consistency of an utterance hypothesis can be calculated by using word relatedness measures derived by latent semantic analysis [24] and point-wise mutual information (PMI) [12].

Another CM estimation approach uses a post-processing step to refine the raw WPPs obtained from the recognizer. This approach, called "confidence calibration" [52], is being progressively improved and results are promising [7, 8, 36, 50, 52]. These methods refine the WPPs by using discriminative models such as the maximum entropy classifier [50], conditional random field [7, 8, 36] and artificial neural networks [52]. Such models combine many features related to the calibration target word, and its context (i.e. words

around the target word). Given the availability of in-domain training data for the discriminative models, which require each word in the recognizer output to be labelled as either correct or incorrect, the confidence calibration approach can substantially improve the quality of confidence measures.

2.2.3 Issues

As regards error rejection in SDP systems, conventional CM estimation methods are plagued by several problems.

A spoken document consists of hundreds of utterances and includes thousands of words. Thus "document-level" CMs that estimate the recognition rate of the target spoken document can be effectively used for document-level error rejection. Unfortunately, conventional studies have focused on word-level CMs and document-level CMs have been ignored. The reason is that general ASR systems assume that utterances are independent and process utterances on an utterance-by-utterance basis. Since document-level CM estimation can use rich information such as longer range context information obtained from multiple utterances, specialized algorithms for document-level CM estimation are needed.

Word-level error rejection should also be used in combination with document-level rejection. The confidence calibration approach effectively yields high quality word-level CMs. However, the domain dependency of discriminative models is a serious problem in actual use cases. As described in Section 2.2.2, human-labeled in-domain training data is required for effective confidence calibration. In practice, such as SDP systems for call centers, it is impossible to create in-domain training data for each call center due to the cost. Unfortunately, a WPP is the only available word-level CM estimation method without supervised training, and its quality is insufficient.

Furthermore, the contextual information used in conventional methods [12,24] is limited to relatively short range context, i.e., at most one utterance, due to the utterance independency assumption. However, utterances in a spoken document exhibit strong dependency since each spoken document has a particular topic. Moreover, each system stores spoken documents that are very likely to have the same topic. The longer range context information

obtained from multiple utterances/documents has not been used for CM estimation in previous studies.

Issues of raised by CM estimation are summarized as follows:

- Document-level CM estimation is useful in SDP systems but has not been studied.

- State-of-the-art confidence calibration methods using supervised training cannot be applied in practical systems due to the high costs of making human-labeled training data.

- Contextual information used in the conventional methods are obtained from at most a single utterance.

In this thesis, we address the first issue by proposing a novel framework for document-level CM estimation in Chapter 3. We also propose a novel unsupervised confidence calibration method to solve of the second issue in Chapter 4. The proposed document-level CM estimation uses contextual information obtained from multiple utterances, and the proposed unsupervised confidence calibration method leverages consistency information found in multiple spoken documents.

2.3 OOV word detection

OOV words, i.e. words missing from the lexicon of the speech recognizer, never appear in the recognizer output even if they are actually present in the input speech. Thus some special technologies other than standard speech recognition systems are required for handling OOV words. OOV word detection is a technology that identifies speech segments where OOV words are uttered.

For OOV word detection, the word/fragment hybrid ASR-based approach is successful and widely used [23, 28, 36, 40, 41]. This section summarizes the approach and its issues to be addressed.

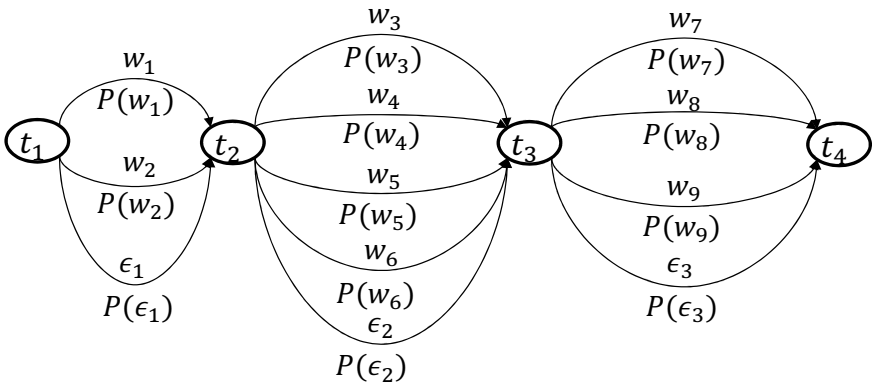

Figure 2.3: An example of a confusion network.

2.3.1 Confusion network

Word/fragment hybrid ASR-based OOV word detection leverages confusion networks (CNs) [26] generated by speech recognizers. An example of a CN is shown in Figure 2.3. The CN is a concise representation of recognition result hypotheses and can be obtained from a word lattice through the following steps [26]:

1. The WPPs of all words in the word lattice are computed as described in Section 2.2.1.

2. Intra-word clustering, which merges the arcs that correspond to the same word and overlap in time is applied. The resulting arc has the maximum WPP among all merged arcs. For example, w_3 and w_7 in Figure 2.2 overlap in time. If w_3 and w_7 are the same word and $P(w_3) > P(w_7)$, the arc of w_7 is deleted by intra-word clustering.

3. Inter-word clustering is applied, which gathers the arcs that overlap in time into a cluster, called a confusion set. Note that overlapping arcs always correspond to different words after the previous step. For example, w_3, w_4, w_5 and w_6 in Figure 2.2 are gathered into a confusion set. Note that w_7 was deleted in intra-word clustering.

4. The ϵ arc that represents a null-word is added to each confusion set. The WPP of the ϵ arc is the sum of the WPPs of arcs removed in the

second step.

5. The WPPs are normalized so that the sum of the WPPs among each confusion set becomes 1.

Each confusion set represents word hypotheses in a particular segment, called a *slot*. OOV word detection using CNs generated by a word/fragment hybrid ASR process is detailed in the next section.

2.3.2 Word/fragment hybrid ASR-based OOV word detection

A word/fragment hybrid ASR is an ASR system that uses a special lexicon and LM, called hybrid lexicon and hybrid LM, respectively; they include not only words but also *fragments*. Fragments are an important subset of all possible subword (phoneme) sequences [41]. Thus CNs generated from the hybrid ASR system contain both words and fragments. When OOV words exist in the input utterance, the confusion sets corresponding to the OOV word segments tend to contain fragments with high WPPs. OOV words can be detected by features that capture this tendency.

The fragments, i.e. a subset of all possible phoneme sequences, are selected from the LM training texts as follows [41]:

1. All words in the LM training texts are first converted into phoneme sequences by a grapheme-to-phoneme converter such as [9].

2. A phoneme 5-gram LM is trained using the converted texts, and entropy-based pruning [45] is applied to select important/informative phoneme N-grams.

3. All remaining phoneme N-grams (1, 2, 3, 4 and 5-grams) in the LM are extracted as fragments.

Since entropy-based pruning does not remove 1-grams, the fragments include all phoneme 1-grams. Thus all possible pronunciations can be represented as fragment sequences.

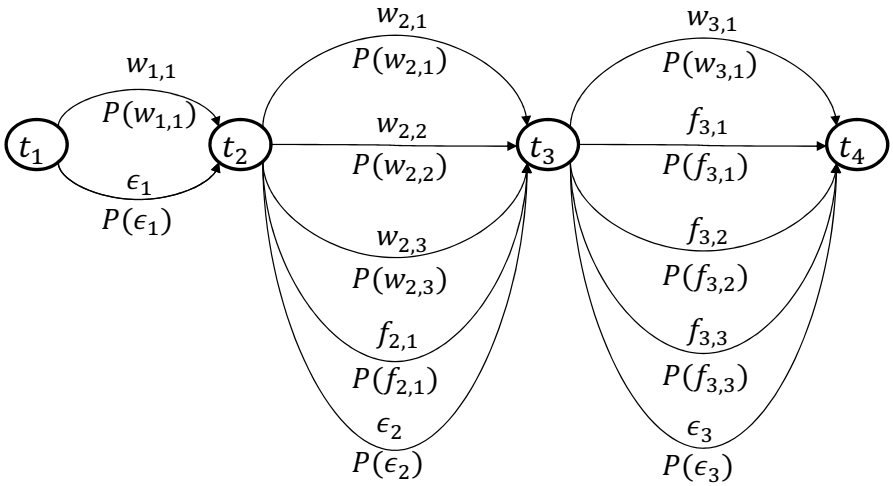

Figure 2.4: An example of a confusion network generated by a word/fragment hybrid ASR system.

Figure 2.4 shows an example of a CN generated from a hybrid ASR system. $w_{i,j}$ denotes words and $f_{i,j}$ denotes fragments where i is slot index and j is word/fragment index in the slot. Features for OOV detection are extracted from each slot. Representative features that capture the degree of match/mismatch to the IV words are as follows [36, 41]:

- **Fragment posterior**: Logarithm of the sum of WPPs of fragments in the target slot:

$$\text{FragmentPosterior} = \log \sum_{f \in F} P(f), \tag{2.5}$$

  where F denotes a set of fragments in the target slot.

- **Word entropy**: Entropy of WPPs of words in the target slot:

$$\text{WordEntropy} = -\sum_{w \in W} P(w) \log P(w), \tag{2.6}$$

  where W denotes a set of words in the target slot.

- **1-best posterior probability**: Maximum log WPP in the target slot.

- **LM score**: Log N-gram probability of the word/fragment that has the largest WPP in the target slot.

- **LM back-off order**: The back-off order of the 3-gram of word/fragment with the largest WPP in the previous 2 slots and the target slot.

These values are computed for each slot, and the values of surrounding slots are concatenated as context features. They are concatenated into one feature vector and input to an IV/OOV classifier, such as a maximum entropy classifier, a conditional random field, or an artificial neural network.

2.3.3 Issues

A big problem of conventional OOV word detection methods is that they trigger many false alarms due to disfluencies such as fillers, repairs, hesitations, or sloppy pronunciation. Word/fragment hybrid ASR-based methods detect the CN slot that does not match any IV word. Disfluencies are not OOV words, but do not match IV words since they consist of irregular phoneme sequences. Thus the conventional hybrid ASR-based method readily detects disfluencies as OOV words. Achieving disfluency-robustness is an important issue in OOV word detection.

In Chapter 5, we address this issue by proposing a novel OOV word detection framework that utilizes consistency information among recurrent appearance of OOV words in a spoken document in order to separate true OOV words and disfluencies.

Chapter 3

Spoken Document Confidence Estimation Using Contextual Consistency

3.1 Overview

As described in Section 1.2, error rejection is one of the two approaches to improve SDP system performance. In SDP systems, document-level rejection that removes poorly recognized spoken documents from the systems can be utilized as well as word-level rejection.

Document-level error rejection is based on a document-level CM estimation method that computes a CM that accurately predicts the recognition rates of each spoken document. However, as described in Section 2.2.3, document-level CM estimation has not been studied since the general assumption of ASR systems is utterance independency.

As described in Section 1.3, our key idea for document-level CM estimation is to leverage the contextual consistency information observed in multiple utterances. Several studies have reported that even if the range of context is limited to one utterance, contextual consistency is effective in estimating word-level CMs [12, 24]. In our task, estimating a document-level CM, contextual consistency is an especially powerful source of information, since a single spoken document usually consists of a few dozen to hundreds

あ/えっと/すいません/**駅**/ですね/**こそ**/**見た**/ん/です/けど/も

あの/**もっと**/**消防士**/いう/**やつ**/**ます**/けども/**含む**/**総合支援**/に/ついて/**用**/**会社**/です/けども

復元/あの/この/**食材**/**を**/**して**/**る**/**ん**/です/よね

これ/は/あの一/**導入**/**人**/から/えっと/**覚え**/**小学校**/**なる**/ん/でしょう/か

Figure 3.1: An example of a Japanese transcript with 45% recognition rate. Misrecognized words are shown in bold.

of utterances on the same topic. In other words, we can use recognition hypotheses with long ranges of up to several utterances to obtain contextual consistency. The transcripts of spoken documents that offer high recognition rates exhibit strong consistency over several consecutive utterances. In contrast, low recognition rate transcripts exhibit strong inconsistency over several utterances. In the example of Figure 3.1, the Japanese transcript with 45% recognition rate contains recognized words that do not make any sense at all. The difference in contextual consistency between high and low recognition rate transcripts is marked.

Based on this understanding, we propose a method that uses contextual consistency over several utterances for estimating a document-level CM. The proposed method sets windows covering several utterances in the transcript of a spoken document, and calculates CM values from the contextual consistency of the content words in each window.

Contextual consistency is formulated as the arithmetic mean of pointwise mutual information (PMI) in the window following [12]. However, data sparseness triggers two problems in PMI. The first one is that the PMI of word-pairs that are not present in the training sets cannot be calculated. The second one is that PMI values become too large when the occurrence frequencies of the words of a word-pair are quite low. Our solution is a PMI smoothing method that overcomes both problems.

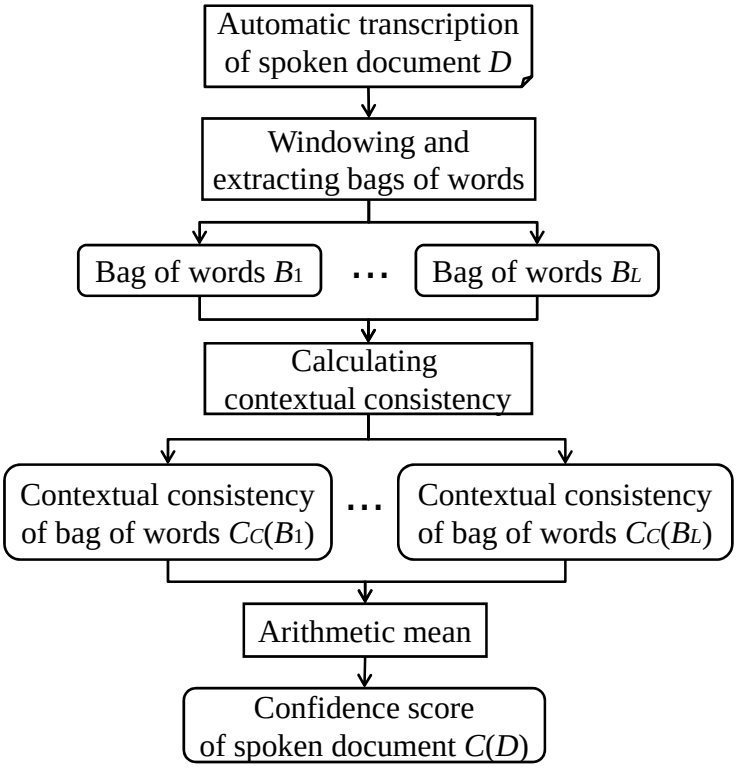

Figure 3.2: Flow chart of proposed spoken document CM estimation.

3.2 Spoken document CM estimation

The entire flow of the spoken document CM estimation proposal is shown in Fig. 3.2.

First, windowing is applied to the automatically generated transcript of spoken document D. The purpose of windowing is to obtain contextual consistency from the range that covers just one topic by adjusting the window length. We assume that topics can be switched within a spoken document, as in phone conversations. Strong contextual consistency is not guaranteed if multiple topics are covered by one window. Therefore, division of the document into segments that cover single topics, windowing, is required.

Window length is expressed as N content words and the amount of window shift is M content words ($N > 1$ and $1 \leq M \leq N$). Each window includes N content words and other words. The content words in each window are extracted and taken to be the bag of words for that window.

The spoken document is taken to be the set of all bags of content words, $D = \{B_1, B_2, \cdots, B_L\}$, where L denotes the number of extracted bags of words.

The contextual consistency of each bag of words is calculated. The proposed method formulates the contextual consistency based on the idea that contextual consistency is low if the relationships between the words in a bag are weak. The contextual consistency of bag of words B_l, which consists of N content words, $w_1, w_2, \cdots, w_N$, is calculated as

$$C_C(B_l) = \frac{1}{N} \sum_{i=1}^{N} \log \frac{P(w_i | B_l \setminus \{w_i\})}{P(w_i)}, \tag{3.1}$$

where $B_l \setminus \{w_i\}$ is the bag of words B_l from which w_i is omitted (i.e. the set of neighbor words of w_i), $P(w_i)$ is the probability that w_i occurs in B_l and $P(w_i|B_l \setminus \{w_i\})$ is the probability that w_i occurs in B_l when the neighbor words of w_i are given. $C_c(B_l)$ is the normalized log likelihood ratio and tests whether the occurrences of each word in B_l are due to a relationship with neighbor words or due entirely to chance. When $C_c(B_l)$ is small, the relationships of words in B_l are weak and the contextual consistency is low.

It is difficult to accurately estimate $P(w_i|B_l \setminus \{w_i\})$ since the number of combinations of words in B_l is enormous relative to the amount of data actually available. Consequently, the proposed method approximates $P(w_i|B_l \setminus \{w_i\})$ as the geometric mean of the probabilities that w_i occurs in B_l when *each* neighbor word of w_i is given. This is calculated as

$$P(w_i | B_l \setminus \{w_i\}) \approx \left(\prod_{j=1, j \neq i}^{N} P(w_i | w_j) \right)^{\frac{1}{N-1}}. \tag{3.2}$$

Substituting Eq. (3.2) into Eq. (3.1) yields the contextual consistency of bag

of words B_l as follows.

$$C_c(B_l) \approx \frac{1}{N}\sum_{i=1}^{N}\log\frac{\left(\prod_{j=1,j\neq i}^{N}P(w_i|w_j)\right)^{\frac{1}{N-1}}}{P(w_i)}, \tag{3.3}$$

$$= \frac{1}{N}\sum_{i=1}^{N}\log\left(\prod_{j=1,j\neq i}^{N}\frac{P(w_i,w_j)}{P(w_i)P(w_j)}\right)^{\frac{1}{N-1}}, \tag{3.4}$$

$$= \frac{1}{N(N-1)}\sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N}\log\frac{P(w_i,w_j)}{P(w_i)P(w_j)}, \tag{3.5}$$

where $P(w_i, w_j)$ is the probability that both w_i and w_j occur in B_l.

Eq. (3.3) is the arithmetic mean of PMI of all word-pairs in B_l. PMI indicates the strength of the relatedness of a word-pair [4]. When the contextual consistency of B_l is low (i.e. B_l includes many error words), more of the word-pairs in B_l are less likely to co-occur. Accordingly, the right side of Eq. (3.3) becomes small. Therefore, approximating $C_c(B_l)$ by Eq. (3.3) is appropriate as a measure of contextual consistency.

The context-based CM, $C_c(B_l)$, can be combined with the decoder-based CM, i.e. WPP [49]. The decoder-based CM of B_l is calculated as follows:

$$C_d(B_l) = \frac{1}{N}\sum_{i=1}^{N}\log P(w_i|\boldsymbol{O}), \tag{3.6}$$

where $\boldsymbol{O}$ is the input acoustic features, and $P(w_i|\boldsymbol{O})$ is the WPP obtained from the ASR decoder. The ASR decoder uses short-range linguistic information, i.e. the word trigram. In contrast, our context-based CM estimation uses longer-range linguistic information over more words, but does not use the word order information that the word trigram contains. It is expected that these two CMs complement each other. The combined CM is calculated as the linear interpolation of $C_c(B_l)$ and $C_d(B_l)$:

$$C_{cd}(B_l) = \lambda \cdot C_c(B_l) + (1-\lambda)\cdot C_d(B_l), \tag{3.7}$$

where λ is the interpolation weight ($0 \leq \lambda \leq 1$).

Finally, the arithmetic mean of $C_{cd}(B_l)$ is calculated as the CM of spoken

document D;

$$C(D) = \frac{1}{L}\sum_{l=1}^{L} C_{cd}(B_l). \tag{3.8}$$

3.3 Smoothing pointwise mutual information

The proposed method calculates the contextual consistency of a bag of words as the mean of PMI of all word-pairs in the bag by Eq. (3.3). Unfortunately, if data is sparse, PMI suffers two problems. Our PMI smoothing method can handle both problems.

3.3.1 Two problems of PMI

The PMI of two words, x and y, is expressed as follows [4].

$$\mathrm{PMI}(x, y) = \log \frac{P(x, y)}{P(x)P(y)} = \log \frac{f(x, y) \cdot K}{f(x)f(y)}, \tag{3.9}$$

where $f(x)$ is the occurrence frequency of x (the number of bags of words that include x), $f(x, y)$ is the co-occurrence frequency of x and y, and K is the total frequency (the number of all bags) in the training set. Positive PMI values indicate that x and y tend to co-occur more often than chance. Negative PMI values indicate that x and y tend not to co-occur more than chance, and PMI becomes 0 when x and y are independent.

The two problems are found in Eq. (3.9);

1. PMI cannot be calculated when $f(x, y) = 0$. In this case, PMI always becomes $-\infty$ according to the above definition.

2. PMI becomes too large when $f(x)$ and $f(y)$ are small. For example, when $f(x) = f(y) = f(x, y) = 1$, $\mathrm{PMI}(x, y)$ is $\log K$ (K exceeds 100,000 in practice). Meanwhile when $f(x) = f(y) = f(x, y) = 50$, x and y have a stronger relationship than is true in the former case. Nevertheless, $\mathrm{PMI}(x, y)$ is smaller, $\log(K/50)$.

Guo et al. proposed a smoothing technique in order to deal with the first problem [12]. This technique corrects the co-occurrence frequencies and

probabilities by adding a constant and interpolating as follows.

$$f(x,y) := f(x,y) + I, \tag{3.10}$$

$$P(x,y) := \frac{P(x,y) + \alpha P(x)P(y)}{1+\alpha}. \tag{3.11}$$

I and α are parameters optimized manually on a development set. All word-pairs have more than 0 frequency by this correction. However, this smoothing method does not deal with the second problem.

3.3.2 Proposed smoothing method

The proposed PMI smoothing method deals with both problems as follows.

Against the first problem, it corrects the frequency of unobserved word-pairs using the simple Turing estimator [11]. The co-occurrence frequencies are corrected as follows.

$$\hat{f}(x,y) = \begin{cases} f(x,y) & \text{if } f(x,y) > 0, \\ \dfrac{N_1}{N_0} & \text{otherwise,} \end{cases} \tag{3.12}$$

where N_1 is the number of word-pairs observed once in the training set and N_0 is the number of unobserved word-pairs. All word-pairs have non-zero frequency by this correction. This correction does not require manual parameter optimization.

Next, to counter the second problem, we introduce the idea that PMI should be 0 if $f(x)$ and $f(y)$ are too small to permit the relationship of the word-pair to be judged. The proposed method uses the t-test to examine whether $f(x)$ and $f(y)$ are large enough or not. The t-score, which tests whether the difference between $P(x,y)$ and $P(x)P(y)$ is significant or not, is calculated as follows [5].

$$\begin{aligned} t(x,y) &\approx \frac{|P(x,y) - P(x)P(y)|}{\sqrt{\frac{P(x,y)}{K}}}, \\ &= \frac{|\hat{f}(x,y) - \frac{f(x)f(y)}{K}|}{\sqrt{\hat{f}(x,y)}}. \end{aligned} \tag{3.13}$$

Table 3.1: Summary of evaluation task.

Size	782 phone calls (61 hours)
Utterance style	Spontaneous
Speakers	17 males / 31 females
Recording conditions	16 kHz / 16 bit
Acoustic model	Triphone HMMs
Language model	Word trigram
Vocabulary size	59,676 words
ASR decoder	VoiceRex [15, 29]

Finally, the smoothed PMI is obtained as follows.

$$\mathrm{PMI}(x,y) = \begin{cases} \log \dfrac{\hat{f}(x,y)\cdot K}{f(x)f(y)} & \text{if } t(x,y) > \theta, \\ 0 & \text{else,} \end{cases} \tag{3.14}$$

where θ is the threshold value of the t-test, which is determined from the significance level. Performing this t-test before PMI suppresses the second problem. For example, when the significance level is set to 5% ($\theta = 1.65$) and $f(x) = f(y) = f(x,y) = 1$, $t(x,y)$ becomes $1 - 1/K < 1.65$. Therefore, the proposed method can let $\mathrm{PMI}(x,y)$ be 0 by Eq. (3.14).

3.4 Experiments

3.4.1 Experimental conditions

Table 3.1 shows the evaluation task. Each phone call was a simulated Japanese call center dialogue. Two speakers, an operator and a customer, talked to each other as in call center dialogues, and the utterances of each speaker were recorded by separate microphones. 782 phone calls (391 operator channels and 391 customer channels) were used as the evaluation set. The lengths of the phone calls ranged from 2 to 17 minutes.

We treat a phone call as a spoken document. The acoustic model and the language model were trained by manual transcripts of 224 hours of call center

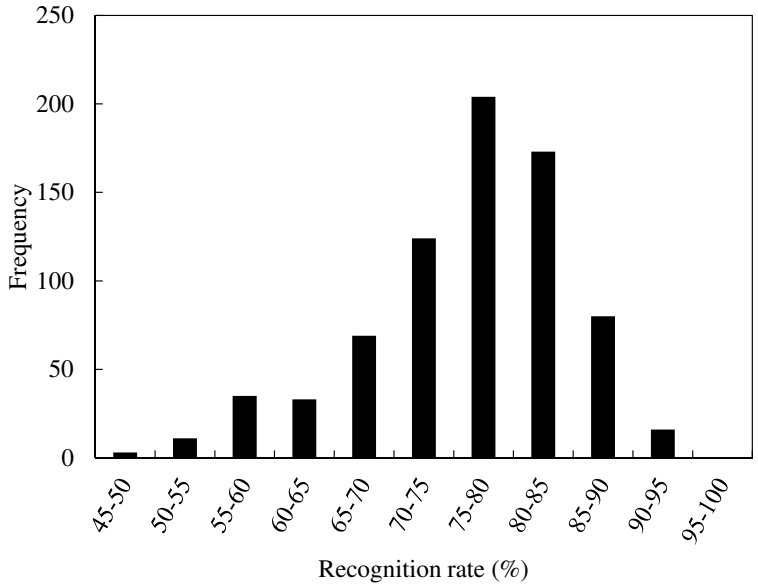

Figure 3.3: Distribution of recognition rates in the test set consisting of 782 calls.

recordings. Both training sets differed from the evaluation set. The average recognition rate (character correctness) of the evaluation set was 79.56%; the minimum was 48.95% and the maximum was 91.76%. In this study, we use characters instead of words as the units for computing the recognition rate to prevent the recognition rate from being varied by the word boundary ambiguity of Japanese. The entire distribution of the recognition rates is shown in Figure 3.3.

Window length N and window shift amount M in the windowing procedure, described in Section 3.2, were optimized so as to maximize the correlation coefficient between the true recognition rates and the document CMs on a development set consisting of 212 phone calls. As a result, $N = 20$ and $M = 10$. The development set differed from both the training and evaluation set. Nouns and verbs in the recognition vocabulary were used as content words. The word occurrence/co-occurrence frequencies used for calculating PMI were counted on the 113,079 bags of words ($K = 113079$), which were extracted from the training set of the language model by the windowing procedure.

3.4.2 Results

3.4.2.1 PMI smoothing

In order to evaluate the effect of our PMI smoothing procedure, the following 4 conditions of PMI smoothing were compared; "No smoothing": PMI was not smoothed at all (when $f(x, y) = 0$, PMI was large, negative, and constant), "Conventional": PMI was smoothed by the method described in [12], "Proposed (Turing)": PMI was smoothed by the method proposed in Section 3.3.2 using only the Turing estimator (without the t-test), "Proposed (Turing+t-test)": PMI was smoothed by the proposed method with both Turing estimator and the t-test. In the "Conventional" condition, smoothing parameters I and α (in Eq. (3.10) and Eq. (3.11)) were optimized on the development set and were, as a result, fixed to $I = 0.10$ and $\alpha = 0.15$, respectively. In the "Proposed" condition, the significance level of the t-test was set to 5% ($\theta = 1.65$) as recommended in [5]. In this experiment, the context-based CMs were not combined with the decoder-based CMs, i.e. λ was set to 1 in Eq. (3.7). The effectiveness was evaluated using the correlation coefficients between the CMs and recognition rates of each spoken document, which were calculated for the entire evaluation set.

Table 3.2 shows the results. "Conventional" and "Proposed (Turing)" offered higher correlation coefficients than "No smoothing," and the highest correlation, 0.614, was achieved by using "Proposed (Turing+t-test)."

"Conventional" addresses the fist problem described in Section 3.3.1, i.e. the zero-frequency problem, by Guo's PMI smoothing method [12]. Improvement from "No smoothing" to "Conventional" means the effectiveness of solving the first problem. "Proposed (Turing)" also addresses the first problem and offered slightly higher correlation than "Conventional." This means the proposed frequency correction based on the Turing estimator is effective since "Proposed (Turing)" requires no manual parameter tuning while "Conventional" has two parameters (I and α).

In addition to deal with the first problem, "Proposed (Turing+t-test)" addresses the second problems, i.e. too large PMI for low frequency word-pairs, by adopting the t-test. The improvement from "Proposed (Turing)" to "Proposed (Turing+t-test)" was achieved by dealing with both the first

Table 3.2: CM improvement by PMI smoothing.

PMI smoothing	Correlation coefficient
No smoothing	0.371
Conventional	0.485
Proposed (Turing)	0.514
Proposed (Turing+t-test)	**0.614**

and second problems.

3.4.2.2 Combination of context-based and decoder-based CMs

In order to evaluate the combination of the context-based CM and the decoder-based CM, we compared the following 3 conditions: "Context-based": Same condition as the "Proposed (Turing+t-test)" in Section 3.4.2.1 ($\lambda = 1$ in Eq. (3.7)), "Decoder-based": CMs are computed by Eq. (3.7) with $\lambda = 0$, "Context+decoder": CMs are computed by Eq. (3.7) with $\lambda = 0.1$. The interpolation weight *lambda* was optimized on the development set.

Table 3.3 shows the results. "Context+decoder" attained a higher correlation coefficient than either "Context-based" or "Decoder-based" conditions. The difference between "Decoder-based" and "Context+decoder" was statistically significant with $p < 10^{-6}$ by the Meng-Rosenthal-Rubin test [31]. This result confirmed that the long-range context information used by the context-based CM and the short-range linguistic information used by the decoder-based CM complement each other in generating a more effective spoken document CM as expected in Section 3.2.

The scatter plot of all spoken documents in "Context+decoder" condition is illustrated in Fig. 3.4. It is confirmed that the points demonstrate a linear arrangement and that our CMs can predict the recognition rates of each spoken document even in the evaluation set that includes both operator's and customer's speech.

Table 3.3: CM improvement by combining context-based and decoder-based CMs.

CM	Correlation coefficient
Context-based	0.614
Decoder-based	0.696
Context+decoder	**0.721**

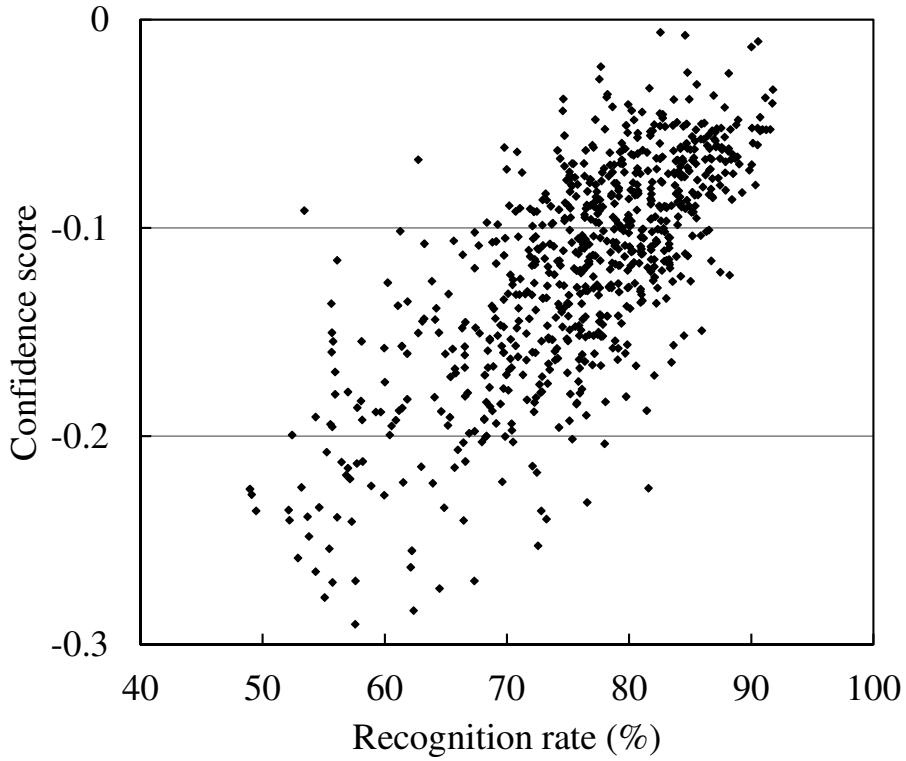

Figure 3.4: CMs and recognition rates of each spoken document in the condition of "Context+decoder" ($r = 0.721$).

3.4.2.3 Spoken document rejection

The purpose of document-level CM is to reject low quality transcripts of spoken documents for suppressing information retrieval errors in SDP systems. Specifically for keyword search systems of spoken documents, search precision is expected to be improved by rejecting transcripts with low CMs before keyword search. In order to evaluate this improvement, we conducted

spoken document search experiments.

First, a CM was created for each spoken document in the test set, then documents with low CM values were rejected by reference to a decision threshold. The set of accepted documents were treated as the target of keyword search; documents that include the query word were retrieved as relevant documents.

We changed the rejection rate by altering the decision threshold. For each rejection rate, keyword search was performed using 1714 nouns, all proper nouns in the test set, as search queries. The performance was evaluated by the average precision (AP), average precisions of all search queries. AP was calculated as follows:

$$AP = \frac{1}{Q} \sum_{i=1}^{Q} \frac{N_{iC}}{N_{iH}}, \tag{3.15}$$

where Q is the total number of queries, N_{iH} is the number of documents retrieved by the i-th query, and N_{iC} is the number of correctly retrieved documents. We counted a retrieved document as correct when the manual transcription of the document included the query word. AP decreased when the query was found to match misrecognized words.

We compared two document CMs, "decoder-based" and "Context+decoder" in Section 3.4.2.2. As a reference, we also investigated the ideal condition where true recognition rates were used instead of CMs ("Ideal").

Table 3.4 shows the results. At all rejection rates of 10~90%, "Context+decoder" yielded higher AP than "Decoder-based" condition. The difference between "Context+decoder" and "Decoder-based" was statistically significant with $p < 10^{-4}$ by the one-sided t-test. This result indicates that context-based CMs are effective in actual SDP systems. Even though actual rejection rate depends on the application and the number of stored spoken documents, the proposed method is effective in various practical situations since it improved search precision at all rejection rates.

Table 3.4: Improvement in average precision by spoken document rejection.

Rejection rate [%]	Average precision [%]		
	Decoder-based	Context+decoder	Ideal
0	69.83	69.83	69.83
10	70.58	**71.28**	71.80
20	72.26	**72.64**	73.18
30	73.25	**73.73**	74.95
40	74.06	**74.38**	76.14
50	75.20	**75.92**	78.08
60	76.56	**77.39**	80.06
70	79.08	**79.80**	82.92
80	81.91	**82.96**	85.80
90	86.84	**87.25**	89.02

3.5 Summary

This chapter presented a method that can, for each spoken document, estimate a document-level CM that accurately represents the recognition rate of each document. The proposed method uses word contextual consistency over several utterances for spoken document CM estimation. We also proposed a new smoothing method that deals with the two problems of PMI triggered by data sparseness.

Experiments were conducted to evaluate how accurately our spoken document CM estimated the recognition rate. The results showed that our document-level CM estimation framework with the PMI smoothing method yields higher correlation between CMs and true recognition rates than conventional PMI smoothing methods. It was also confirmed that combining the proposed context-based CMs and the conventional decoder-based CMs is effective in estimating spoken document CMs. Furthermore, keyword search experiments with document rejection using the document CM proposal showed that introducing contextual consistency information to CM estimation improves search precision and is beneficial for actual SDP systems.

The experiments detailed in this chapter confirmed that the long-range contextual consistency information effectively complements the short-range information in estimating document-level CMs.

Chapter 4

Unsupervised Word Confidence Calibration Using Examples of Recognized Words and Their Contexts

4.1 Overview

In Chapter 3, we proposed a method for estimating the document-level CM and demonstrated its effectiveness in precision-oriented retrieval for spoken documents. In this chapter, we focus on another approach for error rejection, i.e., word-level rejection that excludes incorrect words in transcripts. Even if poorly recognized spoken documents are removed by the document-level rejection, incorrect words still appear in remaining transcripts due to the difficulty of attaining perfect recognition. Of particular importance, incorrect content words cause false information retrieval in down-stream text processing tasks such as keyword search. Word-level rejection is also important for improving SDP system performance.

As described in Section 2.2, WPP [49] is widely used as a fundamental word-level CM, and confidence calibration methods using discriminative models trained by human-labeled training data [7, 8, 36, 50, 52] are currently successful approach. However, as pointed out in Section 2.2.3, it is actually

impractical to apply confidence calibration methods to SDP systems due to the cost of making the human-labeled training data needed for each domain. Moreover, the quality of calibrated CMs is seriously degraded by the domain mismatch problem.

In order to overcome the domain dependency problem and improve the quality of CMs without labeled training data, this chapter presents a novel framework for completely unsupervised confidence calibration. The key idea of the proposed framework is to utilize consistency information observed in multiple spoken documents. Specifically, the proposed method calibrates word CMs by using the CMs of identical words, called "examples," found in the recognition results stored in each deployed system instead of the discriminative models trained by labeled training data. Our proposal makes it possible to improve the CM quality against unknown domain data while totally eliminating the need and cost of creating human-labeled training data.

4.2 Example-based unsupervised confidence calibration

4.2.1 Basic idea

Our main idea is that the correct/incorrect decision of a target recognized word is made more reliable by using the CMs of identical words rather than using just the target's score. Generally, misrecognized words do not occur randomly and tend to have typical contexts (prior and post words) as do correct words. Therefore, identical words that have similar contexts tend to be correct (or incorrect) to the same degree. We focus on this characteristic and use it to improve the CMs.

We call words that are identical to the target word "examples." Examples can be extracted from recognition results stored in deployed systems. Among the many examples existing in the stored recognition results, those that have similar context to that of the target word, "similar examples," are most important. The importance of each example is determined by the context similarity between the target word and the example. Since similar

examples tend to be correct or incorrect to the same degree, the CMs of similar examples are biased high when the target word is correct, and low when the target is incorrect.

Using the mean of the CMs of similar examples improves the correct/incorrect decision. The variance of uncalibrated CMs is large, and that makes the correct/incorrect decision by a CM unstable. By averaging multiple CMs, which are biased high or low according to their correctness or incorrectness, the variance becomes small, and the decision should become stable.

Given a calibration target word, the proposed method first gathers words identical to the target, i.e. examples of the target, from the database of the SDP system. Then, in order to determine the importance of the examples, context similarity between the target word and each example is measured. The CM of the target word is calibrated to the similarity weighted mean of the CMs of the examples.

The implementation of this idea is detailed in the next section.

4.2.2 CM calibration using similar examples

Figure 4.1 shows the flow chart of the proposed method. When a recognition result is provided, the CMs in the result are calibrated by the procedure consisting of example and context extraction, context similarity calculation, and CM calibration. The input recognition result is passed through a part-of-speech filter which drops words other than content words (nouns, verbs and adjectives) since only content words are deemed to be important for information extraction. The calibration procedure for the CM, r_i, of the i-th word, w_i, is detailed below.

In the example and context extraction step, all K words that are identical to target word w_i are extracted from the recognition results already stored in the database of the SDP system. These K words are examples of w_i, where $w_i^{(k)}$ is the k-th example. Each example has its CM; $r_i^{(k)}$ represents the CM of $w_i^{(k)}$. Contexts of w_i and each example are concurrently obtained with context window width N, i.e. the set of N prior words and N post words form the context of the center word. Let c_i be the context of target

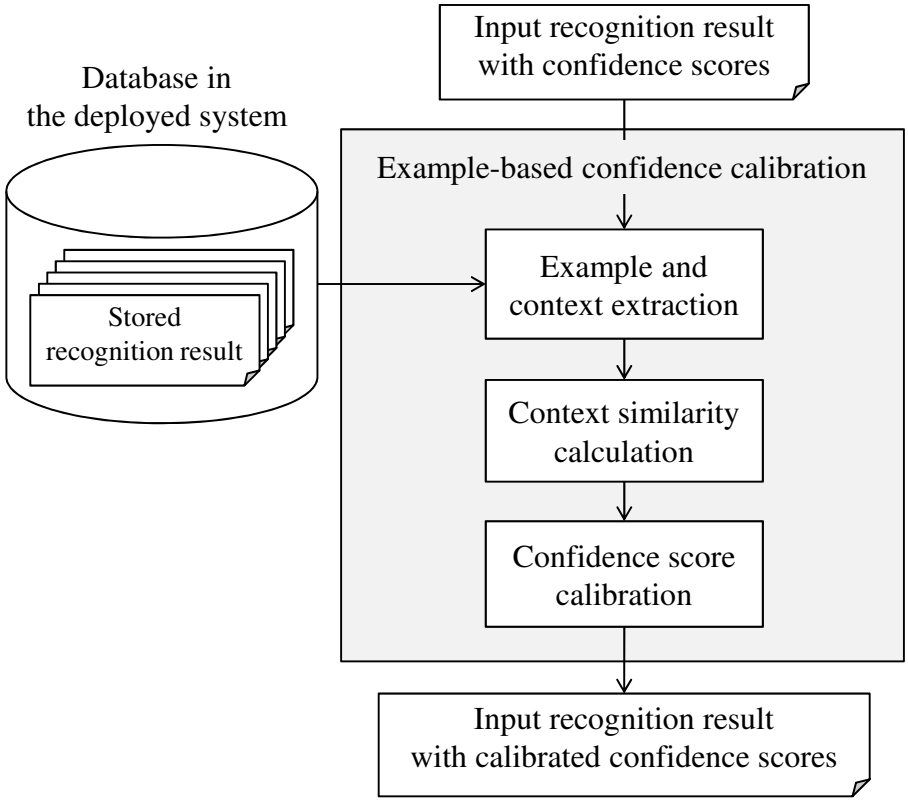

Figure 4.1: Flow chart of proposed confidence calibration.

word w_i, and $c_i^{(k)}$ be the context of example $w_i^{(k)}$. Figure 4.2 summarizes the notations and relationships of the target word, the examples and their CMs and contexts.

In the context similarity calculation step, context similarities between w_i and each example are calculated in order to determine the importance of each example. As mentioned in the previous section, the CMs of "similar" examples should be averaged to calibrate r_i and thus improve the reliability of the correct/incorrect decision. The proposed method uses the cosine similarity between the context of the target word, c_i, and the context of each example, $c_i^{(k)}$:

$$S(c_i, c_i^{(k)}) = \frac{|c_i \cap c_i^{(k)}|}{\sqrt{|c_i| \cdot |c_i^{(k)}|}}, \tag{4.1}$$

where $S(c_i, c_i^{(k)})$ is the similarity between c_i and $c_i^{(k)}$, $|c_i \cap c_i^{(k)}|$ is the number

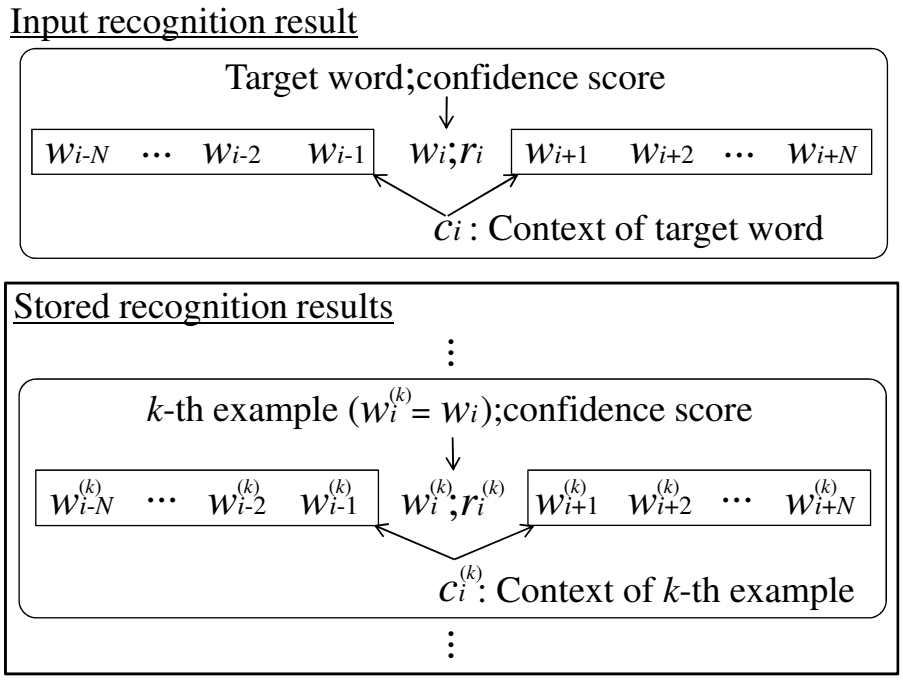

Figure 4.2: Notations and relationships of the target word, examples and their CMs and contexts.

of words that are commonly in c_i and $c_i^{(k)}$, and $|c_i|$ and $|c_i^{(k)}|$ are the number of words included in c_i and $c_i^{(k)}$, respectively. The maximum value of $S(c_i, c_i^{(k)})$ is 1, and the minimum value is 0 since neither the numerator nor the denominator are negative.

In the CM calibration step, CM r_i of target word w_i is calibrated by using the similarities and the CMs of the examples. The proposed method uses context similarity (i.e. importance) to weight each example and ignores dissimilar examples due to the assignment of low weights as follows:

$$\hat{r}_i = \frac{r_i + \sum_{k=1}^{K} S(c_i, c_i^{(k)}) r_i^{(k)}}{1 + \sum_{k=1}^{K} S(c_i, c_i^{(k)})}, \tag{4.2}$$

where $\hat{r}_i$ is the calibrated CM of the target word w_i. This is the similarity weighted mean of the CM of w_i and the examples. In determining the calibrated score, the maximum weight value of 1 is used for the uncalibrated (original) score of the target word, r_i. The calibrated CM $\hat{r}_i$ ranges from 0 to 1 since the proposed method assumes WPPs as uncalibrated CMs. In Eq.

(4.2), while $r_i^{(k)}$ can be replaced by calibrated score $\hat{r}_i^{(k)}$, we consider the case of uncalibrated scores.

The CMs of all words in the input recognition result are calibrated by the above procedure. This procedure is "confidence calibration" since it is a post-processing step that uses outputs of a general ASR engine (1-best hypotheses and CMs). Note that this method is completely unsupervised since this procedure only uses automatically generated information, i.e. CMs (r_i and $r_i^{(k)}$) and the words present in the recognition results (w_i, $w_i^{(k)}$, c_i and $c_i^{(k)}$), and is domain independent since it uses the data generated by deployed systems.

4.3 Experiments

In order to evaluate the effectiveness of our proposal, we conducted two experiments on a call center task as follows:

- **Experiment 1: CM distributions**
  The objective of this experiment is to validate the main idea in Section 4.2.1, that is the variance of CMs can be reduced by our proposal of using similarity weighted means of CMs of the examples.

- **Experiment 2: CM quality**
  This experiment evaluates the improvements in the quality of CMs yielded by the proposed calibration in terms of the performance of incorrect word detection.

The conditions and results of the experiments are described below.

4.3.1 Experimental setup

Phone calls recorded in an actual call center were used in the experiments. Table 4.1 shows utterance domains (topics) and data set size of the evaluation set.

Each phone call was transcribed by the WFST-based ASR decoder, VoiceRex [15, 29]. The acoustic model was speaker independent 3-state left-to-right triphone HMMs, which were discriminatively trained by the dMMI

Table 4.1: Data descriptions of the evaluation set.

Utterance domain	Contract of Internet services
# of calls for test	275 calls (39 hours)
Character error rate	19.6%
# of correct words	79,419
# of incorrect words	14,746

criterion [30] using a 224 hour training set. The language model was trained against a set consisting of manual transcripts of call center recordings, with a total of 1 million words. The vocabulary size was 59676 words. Both training sets differed from the evaluation set. Character error rate of the evaluation set was 19.6% and the numbers of correct and incorrect words after part-of-speech filtering, which passed only nouns, verbs and adjectives in the evaluation set, were 79,419 and 14,746, respectively (the 4th and 5th rows in Table 4.1).

WPPs [49] were given for each word as uncalibrated CMs. The examples used in the proposed method were extracted from the recognition results of calls using the leave-one-out approach, i.e. examples were extracted from 274 calls in the evaluation set other than the target test call.

The only parameter of the proposed method is context window width N. N was optimized on the development set, which differed from both the training and evaluation set, and was fixed to $N = 5$.

In Experiment 2, CM quality of the proposed method was compared to the quality of the uncalibrated WPP and that of the conventional supervised calibration method using discriminative models. Maximum entropy (MaxEnt) model was used as the discriminative model [50, 52]. 1-gram, 2-gram and 3-gram of word and part-of-speech tag, and WPP of the calibration target word and its prior and post 2 words were used as features, which can be extracted from the recognizer outputs in the post-processing step.

The training data for the MaxEnt model is the ASR transcript set; the words are manually labeled as either correct or incorrect. We assume the situation where the system is deployed to a new call center. Usually in this situation, the human-labeled training data is not available due to cost and

Table 4.2: Reduction in standard deviation of CM distributions from uncalibrated WPP to calibrated WPP by the proposed method.

	WPP		Calibrated WPP	
	Mean	Std. dev.	Mean	Std. dev.
Correct	0.75	0.24	0.80	0.16
Incorrect	0.47	0.26	0.52	0.22

time constraints. Using the MaxEnt model already trained by other call center data is a possible solution for the conventional supervised method. To simulate this out-of-domain condition, the MaxEnt model was trained against labeled recognition results of 782 calls (61 hours) recorded in a call center different from that of the evaluation set. The ideal (but not practical) situation for the conventional method is where the labeled training data of the target call center is available. We also simulated this in-domain condition by training the MaxEnt model using the labeled recognition results of a part of the evaluation set (4-fold cross validation on the evaluation set was conducted). The out-of-domain condition representing the situation of a new domain call center is our main target, and the ideal in-domain condition is merely a reference. In Section 4.3.3, "Conventional" denotes the out-of-domain condition, and "Ideal" denotes the in-domain condition.

4.3.2 Results of Experiment 1

Table 4.2 shows the means and standard deviations of both the uncalibrated and calibrated CMs of the correct and incorrect words.

The standard deviations of CMs of both correct and incorrect words are diminished after calibration, but the difference in the means between correct and incorrect words was not changed (ΔMean from uncalibrated to calibrated WPP were 0.05 in both conditions). The reduction in variances for both correct and incorrect words was statistically significant ($p < .01$ by the F-test).

These results validate our idea that calibration based on the similarity weighted mean of CMs of examples reduces the variance of CMs. This vari-

ance reduction should make the correct/incorrect decision more reliable.

4.3.3 Results of Experiment 2

CM quality was assessed by the normalized cross entropy (NCE). NCE can be calculated without using thresholds and is used for evaluating the overall CM quality [8, 52]. NCE is defined as follows:

$$\mathrm{NCE} = \frac{H_{\mathrm{base}} - H_{\mathrm{cond}}}{H_{\mathrm{base}}}, \tag{4.3}$$

$$H_{\mathrm{cond}} = -\sum_{i=1}^{M} \log\left[r_i \delta(y_i = 1) + (1 - r_i)\delta(y_i = 0)\right], \tag{4.4}$$

$$H_{\mathrm{base}} = -m \log\left(\frac{m}{M}\right) - (M - m) \log\left(1 - \frac{m}{M}\right), \tag{4.5}$$

where M is the number of total (correct and incorrect) words, m is the number of correct words and r_i is the CM of the i-th word. $y_i = 1$ if the i-th word is correct, $y_i = 0$ otherwise, and $\delta(x) = 1$ if x is true and $\delta(x) = 0$ otherwise. NCE becomes large when the CMs have good quality, i.e. the CMs of correct words are biased high and the scores of incorrect words are biased low.

As a preliminary experiment, we investigated the relationship between the number of examples used for calibration and the quality of calibrated CMs. We limited the maximum number of examples to $K_{\max}$, i.e. $K_{\max}$ examples that have higher context similarity were used for calibration by Eq. (4.2) when the number of extracted examples was more than $K_{\max}$. Table 4.3 shows the NCEs of calibrated CMs for several $K_{\max}$ values. $K_{\max} = 0$ means that the CMs were uncalibrated. $K_{\max} = \infty$ means the maximum number of examples was not limited. Basically the NCEs increased as $K_{\max}$ increased, but the improvement became small when $K_{\max} \geq 30$. The proposed method uses the weighted average CMs of the examples. The weighted average becomes stable when sufficient numbers of examples are used. This result confirmed that stable confidence calibration can be achieved by using 30 or 40 examples in our experiments.

We now compare the NCEs of uncalibrated and calibrated CMs for the evaluation task (see Section 4.3.1 for the details of each condition). $K_{\max}$

was set to 40 in the proposed method. The results are shown in Table 4.4. The CMs calibrated by the proposed method outperformed the uncalibrated WPPs and the conventional method in the out-of-domain condition. This indicates that the proposed unsupervised calibration is more effective than conventional supervised calibration if human-labeled in-domain training data is not available.

A general application of CMs is incorrect word detection. The performance of incorrect word detection by thresholding the CMs was evaluated by precision and recall. The precision is calculated as N_i/N_d, where N_d is the number of words whose CM is under the threshold and N_i is the number of incorrect words whose CM is under the threshold. The recall is calculated as N_i/N_I, where N_I is the total number of incorrect words in the evaluation set.

Figure 4.3 shows the precision-recall curve that plots the precision and recall values when the detecting threshold is altered in each condition. The conventional method could yield very high performance in the ideal condition if the in-domain training set was used ("Ideal"). However, in the out-of-domain condition where the training set of another call center was used, the performance of the conventional method fell dramatically ("Conventional"). The line (performance) of the proposed unsupervised calibration method is always above the lines of uncalibrated WPP and the conventional method in the out-of-domain condition. This confirms that the proposed calibration method can yield better CMs in terms of the performance of incorrect word detection than either WPP or the conventional method in the situation where the labeled data is not available.

The results described in this section confirm that the proposed unsupervised calibration method can yield CMs that have better quality than either WPP or the supervised calibration method if different domain training data is used. This means that the proposed method is valuable in practical situations where manually labeled in-domain data cannot be created, such as deploying the system to a new call center.

Table 4.3: Relationship between the maximum number of examples and CM quality.

$K_{\max}$	0	1	3	5	10	20	30	40	50	70	∞
NCE	0.015	0.100	0.144	0.160	0.173	0.181	0.184	0.186	0.187	0.188	0.190

Table 4.4: Improvements in NCE from uncalibrated WPPs to calibrated WPP achieved by conventional and proposed methods.

WPP	Calibrated WPP		
	Conventional	Proposed	Ideal
0.015	0.119	0.186	0.563

4.4 Summary

This chapter presented a novel unsupervised confidence calibration framework that uses examples of recognized words and their contexts present in the recognition results stored in deployed systems; it does not require any human-labeled training data at all. This framework makes it possible to improve the quality of word-level confidence measures in situations where in-domain labeled data is not available, such as the case of SDP systems newly deployed in a wide variety of call centers. The proposed method is based on the idea that the mean of confidence scores of the examples whose contexts are similar to the target word are more reliable than just the target's score. The confidence score of the target word is calibrated to the similarity weighted mean of the confidence scores of the examples found in the recognition results stored in the deployed system

Experiments showed that the calibration proposal stabilized correct/incorrect decision by reducing the variance of confidence scores and improved the performance of incorrect word detection on actual call center data. The results validated our idea and confirmed that the proposed method can yield greater improvements in the confidence measure quality and the accuracy of incorrect word detection than the conventional method in the practical situation where manually labeled in-domain data is not available.

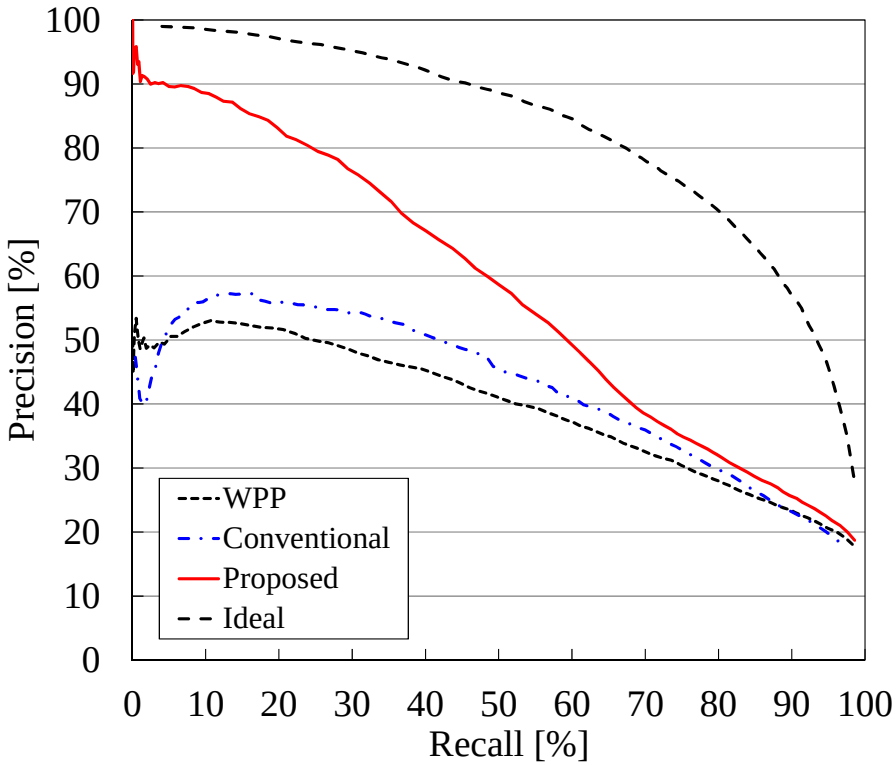

Figure 4.3: Improvements in incorrect word detection performance from uncalibrated WPPs and the conventional calibration method achieved by the proposed method on the unknown domain task.

The proposed method can yield more accurate word confidence measure than conventional methods in a completely unsupervised manner. The success of the proposed approach confirmed that the consistency information observed on multiple examples present in multiple spoken documents is essential for word-level confidence estimation.

Chapter 5

Recurrent Out-of-Vocabulary Word Detection Based on Distribution of Features

5.1 Overview

As described in Section 1.2, out-of-vocabulary (OOV) word detection is critical if we are to enhance the "feedback and reprocessing" approach. An ASR engine can correctly recognize only in-vocabulary (IV) words that are contained in the lexicon of the ASR engine, and OOV words are never correctly recognized. Furthermore, important keywords that are repeatedly uttered in a spoken document, e.g. names of people/places/products or technical terms, are likely to be OOV words since it is impossible to create a lexicon that covers all words possible. Even though the impact on the word error rate (WER) is small, such important OOV words likely to be content bearing and thus have a big impact on down-stream text processing that uses keywords in the recognizer outputs. Thus, detecting important OOV words and adding them to the lexicon of the ASR engine is essential for improving SDP systems.

As described in Section 2.3, word/fragment hybrid ASR-based OOV word detection is a successful approach. However, the big problem is its many false alarms due to disfluencies such as fillers, repairs, hesitation, or sloppy pro-

nunciation. This problem seriously hampers SDP system effectiveness since spoken documents are usually spontaneous speech containing more disfluencies than read speech.

The conventional hybrid ASR-based methods use features that represent OOV likelihood, which are extracted from confusion networks (CNs) generated by the hybrid ASR. To reduce false alarms caused by disfluencies we need some additional information that can separate OOV words and disfluencies. To this end, our key idea is to utilize the consistency of recurrent OOV words in a spoken document as mentioned in Section 1.3. True OOV words tend to have consistent syntactic and phonetic properties across multiple occurrences since they are words. On the other hand, disfluencies have weak consistency since they are not words. Extending the detection process to include the degree of consistency should improve the robustness of OOV word detection.

Based on this idea, this chapter presents a novel method that reduces false alarms by correctly detecting recurrent OOV words; for this we utilize their repeated appearance in spoken documents. The proposed method first detects recurrent segments, segments that contain the same word, in a spoken document by open vocabulary spoken term discovery using a phoneme recognizer [19, 35, 39]. The degree of consistency is then measured by using the distribution (mean and variance) of features (DOF) derived from the recurrent segments. When the same OOV word appears in multiple segments, the posterior probabilities of fragments in those segments become consistently high. This property can be captured by our DOF as large mean and small variance values. Finally, the DOF is used for robust IV/OOV classification.

Obviously this approach has a drawback in that singleton OOV words, i.e. OOV words that do not occur repeatedly, cannot be detected. However, we believe that detecting recurrent (important) OOV words with high precision is critical for practical lexicon maintenance operation. Actually, Our examination of academic lectures found that 66% of OOV words were uttered more than one time (see Section 5.3.1).

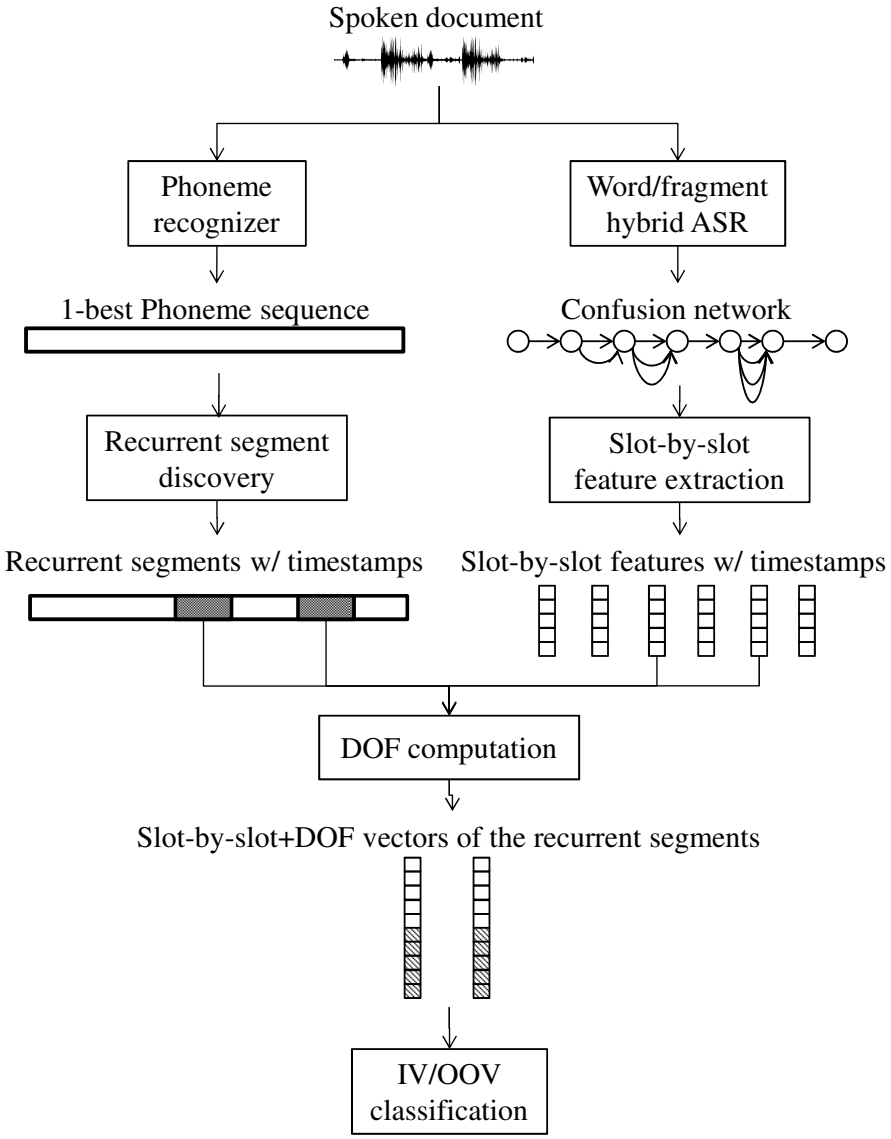

Figure 5.1: Recurrent OOV word detection using distribution of features.

5.2 Method for recurrent OOV word detection

The full procedure of our recurrent OOV word detector is illustrated in Figure 5.1.

The input spoken document is decoded by both a phoneme recognizer

and a word/fragment hybrid recognizer. From the output of the phoneme recognizer, recurrent segments, segments in which the same word is uttered, are detected by the recurrent segment discovery module. Standard slot-by-slot features are extracted from the CN yielded by the hybrid recognizer. DOFs are computed by using the slot-by-slot features that correspond to the recurrent segments. The slot-by-slot features and our DOF are concatenated and input to the IV/OOV classifier, and each recurrent segment is classified as either IV or OOV.

Note that Figure 5.1 shows the simplest case in which only one pair of recurrent segments are detected. Actually many recurrent segments (no overlaps) will be detected and the DOF computation and IV/OOV classification are applied to each recurrent segment. Details of each module are given below.

5.2.1 Recurrent segment detection based on phoneme recognition

In order to discover the phonetically consistent segments without the influence of OOV words, we use an ASR system without linguistic constraints, i.e. a phoneme recognizer. The input spoken document is converted into a 1-best phoneme sequence by phoneme recognition using a deep neural network-based triphone HMM (DNN-HMM) acoustic model and a phoneme 3-gram LM.

The objective of recurrent segment discovery is detecting segments wherein the same word is uttered. We borrow the idea of subword-based open vocabulary spoken term detection [19, 35, 39], and assume that similar sub-sequences appearing in a 1-best phoneme sequence can be treated as the same word. In the proposed method, similar phoneme sub-sequences are extracted by two steps: 1) Detecting sub-sequences whose frequency is at least N and length (number of phonemes) is at least L, and 2) clustering phonetically similar sub-sequences.

All sub-sequences that have at least frequency N and length L can be efficiently extracted by the PrefixSpan algorithm [37] which is widely used for frequent sequential pattern mining [34]. We set L to 5 since most OOV

words have at least 5 phonemes (approximately 3 Japanese moras), and N to 2 for extracting as many as possible recurrent segments (i.e. OOV word candidates). Sub-sequences are extracted with timestamps in the spoken document, and if detected sub-sequences overlap, they are merged into one longer sub-sequence.

Even if the same word is uttered, the decoded phoneme sequences are likely to be slightly different because of ambiguity in pronunciation or phoneme recognition errors. In order to deal with these small differences, we collect similar sub-sequences based on the edit distance between sub-sequences. The distance between two sub-sequences, s_1 and s_2, is calculated as the normalized edit distance:

$$D(s_1, s_2) = \frac{\text{edit}(s_1, s_2)}{\max(|s_1|, |s_2|)}, \tag{5.1}$$

where $\text{edit}(s_1, s_2)$ is the edit distance between s_1 and s_2, and $|s_1|$ and $|s_2|$ are the number of phonemes in s_1 and s_2, respectively. The edit distance is unweighted, i.e. insertion, deletion and substitution are treated equally. $D(s_1, s_2)$ becomes 0 when s_1 and s_2 are the same, and 1 when phonemes consisting of s_1 and those consisting of s_2 do not overlap at all.

Since the number of unique words in each spoken document is unknown, the number of clusters cannot be pre-determined. Thus, we employ a graph-based clustering method that detects the appropriate number of clusters automatically. A similarity graph of sub-sequences is constructed based on the normalized edit distance (similarity is $1 - D(s_1, s_2)$), and input to the graph-based clustering algorithm. In our experiments, the Chinese Whispers algorithm [1] is used as the graph-based clustering method, as it is parameter-free and has been reported to have good performance [27]. Sub-sequences in the same cluster are treated as recurrent segments. The use of edit distance-based clustering ensures that segments with phonetic consistency are extracted as recurrent segments.

Recurrent segments are candidates of recurrent OOV words, and each recurrent segment is an IV/OOV classification target. Note that the start/end timestamps of the segments do not necessarily match the start/end timestamps of actual words; multiple segments can overlap a word and multiple words can overlap a segment, which leads to ambiguity in counting correctly

classified segments. In this study we define a segment as corresponding to the word that has the longest overlap. Thus each segment always has one corresponding word while some words do not have any corresponding segments. In principle the start/end timestamps of detected OOV segments do not strictly match the timestamps of actual OOV words, however, in practice this subtle difference is not a problem since the human system operators usually check the detected segments and their surroundings by ear.

5.2.2 Slot-by-slot feature extraction using hybrid ASR

The input spoken document is also processed by the word/fragment hybrid ASR to extract slot-by-slot features.

First, fragments are selected from the LM training texts by the procedure described in Section 2.3.2. In our experiments, we adjusted the parameters of the entropy-based pruning so as to select 10K fragments as in [41].

In order to compare our method to the conventional method, the word/fragment hybrid lexicon and the hybrid 3-gram LM are also constructed in the same manner as [41]. The hybrid lexicon and 3-gram LM are constructed on the LM training texts in which words with frequency 1 are replaced by their fragment sequences. A fragment sequence of a word is determined by the leftmost longest match to the pronunciation (phoneme sequence) of the word obtained from the grapheme-to-phoneme converter. Note that the hybrid LM does not contain an UNK (unknown word) symbol since fragment sequences are used instead of the UNK symbol.

ASR using the word/fragment hybrid LM generates the CNs against the input spoken document. Features for OOV word detection are extracted from each slot. As the slot-by-slot features, we use the feature set described in Section 2.3.2, i.e. fragment posterior, word entropy, 1-best posterior probability, LM score, and LM back-off order. The effectiveness of these values was reported in previous studies [36, 41]. The five features are computed for each slot, and the features of surrounding 4 slots, i.e. the previous 2 and the post 2 slots, are used as the context. A concatenated 25 dimensional vector is used as a slot-by-slot feature of the target slot. We do not use the word itself as a feature since the raw lexical information is highly dependent on

the domain (topic) of the LM training texts.

5.2.3 DOF computation

In order to capture the consistency of the slot-by-slot features from the multiple appearances of the same word, distribution of features (DOF) are computed using the sub-sequence cluster (i.e. recurrent segments) obtained in Section 5.2.1.

Our DOF consists of the means and variances of slot-by-slot features. If recurrent segments in a cluster are recurrent OOV words, the segments are likely to have consistently OOV-like features, e.g. large fragment posteriors. This consistency is captured by taking the means and variances in the cluster, e.g. large mean and small variance of fragment posteriors strongly indicate that the recurrent segments in the cluster are recurrent OOV words. These statistics should be a more robust indicator of OOV than the individual slot-by-slot features.

A DOF is computed for each cluster as follows:

1. Slot-by-slot features corresponding to the cluster are selected based on timestamps. For each recurrent segment in the cluster, a slot-by-slot feature that has the longest overlap is selected as the corresponding feature.

2. The DOF of the cluster, $\boldsymbol{d}$, is computed as the element-wise means and variances of the selected slot-by-slot features:

$$\boldsymbol{\mu} = \frac{1}{M}\sum_{m=1}^{M}\boldsymbol{v}_m, \tag{5.2}$$

$$\boldsymbol{\sigma} = \mathrm{diag}\left\{\frac{1}{M}\sum_{m=1}^{M}(\boldsymbol{\mu}-\boldsymbol{v}_m)(\boldsymbol{\mu}-\boldsymbol{v}_m)^T\right\}, \tag{5.3}$$

$$\boldsymbol{d} = [\boldsymbol{\mu}^T\boldsymbol{\sigma}^T]^T, \tag{5.4}$$

where M denotes the number of recurrent segments in the cluster, and $\boldsymbol{v}_m$ denotes the corresponding slot-by-slot feature of the m-th recurrent segment in the cluster. T denotes vector transposition and diag represents the vector consisting of the diagonal elements of the matrix.

Table 5.1: Data set sizes.

Group	#lectures	Time length	Vocab. size
A	1351	266h	62741
B	1350	265h	62887

As a result, the m-th recurrent segment in the cluster has a 75 (25 slot-by-slot and 50 DOF) dimensional feature vector, $[\boldsymbol{v}_m^T \boldsymbol{d}^T]^T$, and this vector is used for IV/OOV classification. Note that recurrent segments in a cluster share the same DOF. By applying the above procedure to all clusters, all recurrent segments are assigned their own 75 dimensional feature vector with DOF.

5.2.4 IV/OOV classification

IV/OOV classification is based on the standard supervised training framework. A training set, a set of spoken documents in which true OOV segments are known, is used for training a classifier. The timestamps of the true OOV segments are obtained by forced alignment using manual transcriptions. The trained classifier is used for labeling recurrent segments in the test spoken documents either IV or OOV. Feature vectors with DOF described in Section 5.2.3 are used for classification.

Several binary classifiers can be used for IV/OOV classification. We use a multi-layer perceptron (MLP) for classification since the proposed DOFs are real values and an MLP can use real values as input without any quantization. Note that sequence classifiers such as the conditional random field or the recurrent neural network are not suitable since the classification targets (recurrent segments) do not necessarily form a sequence as shown in Figure 5.1.

Table 5.2: Word/phoneme error rates.

SNR	Group	%WER	%PER
Clean	A	22.9	10.2
Clean	B	23.0	10.3
10dB	A	31.5	17.3
10dB	B	32.1	17.6
5dB	A	44.7	28.1
5dB	B	45.5	28.9

Table 5.3: Data used in the experiments.

Group	Test 1	Test 2	Test 3	Test 4
A-1	ASRtrain	ASRtrain	OOVtrain	Test
A-2	ASRtrain	ASRtrain	Test	OOVtrain
B-1	OOVtrain	Test	ASRtrain	ASRtrain
B-2	Test	OOVtrain	ASRtrain	ASRtrain

5.3 Experiments

5.3.1 Data

The Corpus of Spontaneous Japanese (CSJ) [25] was used for OOV word detection experiments. It consists of 2701 Japanese academic lectures (531 hours, 7M words) with manual transcriptions. It includes various topics such as signal processing, Japanese history, and geography. Each lecture was treated as one spoken document.

The lectures were randomly split into two groups to make ASR training sets so that the amounts of the two groups were balanced. Table 5.1 shows the size of the groups. The DNN-HMM acoustic model, the hybrid lexicon and the hybrid 3-gram LM trained on Group A were used for recognizing Group B, and vice versa. The DNN of the acoustic model had 8 hidden layers with 2048 sigmoid units and a softmax output layer with 3072 units, which was initialized by discriminative pre-training [43] and fine-tuned by stochastic gradient descent (SGD) with momentum. 11 consecutive frames (center,

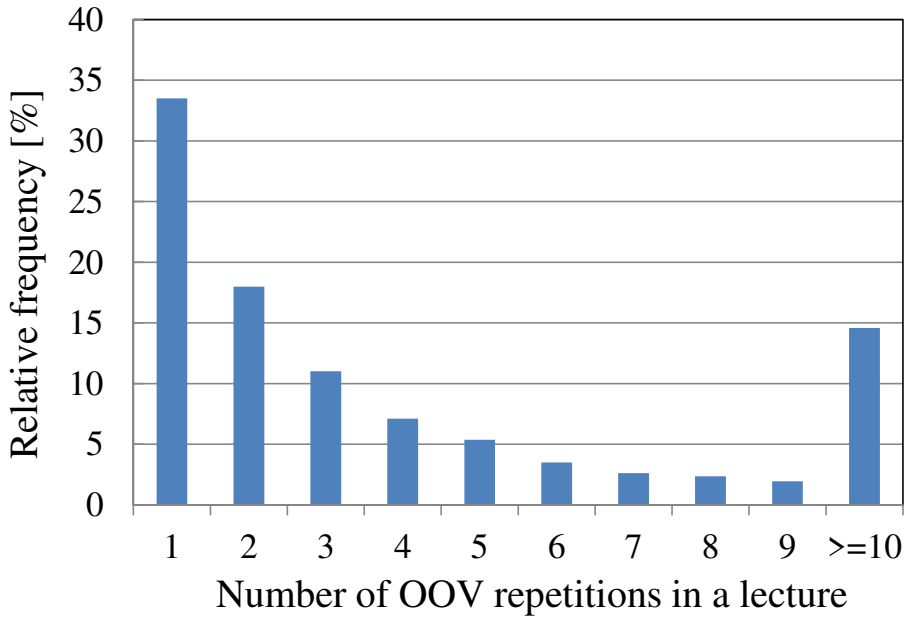

Figure 5.2: Histogram of number of OOV repetitions in a lecture.

previous 5 and post 5 frames) of 38 dimensional acoustic features (12MFCC, 12ΔMFCC, 12ΔΔMFCC, Δpower and ΔΔpower) were concatenated and input to the DNN. The same acoustic model is used in the hybrid ASR and the phoneme recognizer. JTAG [9] was used as the grapheme-to-phoneme converter in training the hybrid LM. Decoding was performed by the WFST-based decoder VoiceRex [15, 29]. In this setting, total number of OOV words in Group A and B was 79826, and the OOV rate was 1.1%. Figure 5.2 shows the histogram of the number of OOV word repetitions per lecture. According to the histogram, 66% of OOV words in a lecture appeared at least twice.

In order to reveal the impact of the error rates of the speech recognizers on OOV detection, we conducted experiments on two noisy conditions in addition to the clean condition. In the noisy conditions, all lectures were contaminated by white noise with signal to noise ratios (SNR) of 10dB and 5dB. Table 5.2 shows WERs of the hybrid ASR and phoneme error rates (PERs) of the phoneme recognizer of Group A and B in clean, 10dB, and 5dB conditions. In the experiments the decoding parameters were adjusted so as to minimize the error rates and thus obtain reasonably accurate start/end timestamps.

To make a training set for the OOV classifier separately from the ASR training set, we conducted two-fold cross validation. Table 5.3 shows the data used in our experiments. "ASRtrain" and "OOVtrain" represent the training sets of ASR and OOV classifier, respectively. All 2701 lectures were used as a test set through the four tests, and the overall results are reported in Section 5.3.3.

5.3.2 Experimental conditions

The parameters of recurrent segment discovery, hybrid ASR and slot-by-slot feature extraction are described in Sections 5.2.1 and 5.2.2. The MLP for IV/OOV classification has 2 hidden layers with 64 sigmoid units and a softmax output layer with 2 (IV or OOV) units. It was randomly initialized and trained by standard SGD with momentum. The momentum coefficient was set to 0.9. At the same time, 10% of samples were randomly selected from the training set of the OOV classifier and separated as a validation set. The learning rate was initialized to 0.08 and halved when classification accuracy on the validation set was decreased, and training was stopped when the learning rate fell under 0.0008. The model parameters that yielded the highest accuracy on the validation set were used in the test.

The true IV/OOV segments in the lectures were labeled by forced alignment using manual transcriptions. Recurrent segments and their feature vectors were extracted by the method described in Section 5.2. As described in Section 5.2.1, each recurrent segment corresponded to an actual word that had the longest overlap. In training, recurrent segments that corresponded to OOV words were treated as positive samples, and those that corresponded to IV words were treated as negative samples. In testing, the MLP gave OOV probabilities to recurrent segments, and the segments whose OOV probability exceeded a decision threshold were classified as OOV.

Note that recurrent segments are only a portion of each entire spoken document, i.e. there were the segments that were not extracted as the recurrent segments. Such segments were not classified as OOV. The segments corresponding to IV words and misclassified as OOV were counted as false alarms. The segments corresponding to OOV words and misclassified as IV and the

true OOV words that had no corresponding recurrent segments were counted as misses. When segments that overlap true OOV words were misclassified as IV, they were counted as misses even if other segments corresponding to the same OOV word were correctly classified as OOV.

The performance was evaluated by the receiver operating characteristic (ROC) curve, the contour of false alarm probabilities and OOV detection probabilities formed when the threshold is varied. The false alarm probability, $P(\mathrm{FA})$, and the OOV detection probability, $P(\mathrm{OOVdet})$, were computed as follows:

$$P(\mathrm{FA}) = \frac{N_{\mathrm{FA}}}{N_{\mathrm{Detect}}}, \tag{5.5}$$

$$P(\mathrm{OOVdet}) = 1 - \frac{N_{\mathrm{Miss}}}{N_{\mathrm{OOV}}}, \tag{5.6}$$

where N_{FA} is the number of false alarms, N_{Detect} is the number of recurrent segments classified as OOV, N_{Miss} is the number of misses, and N_{OOV} is the number of true OOV words.

In order to evaluate the effectiveness of DOF, we compared the following two conditions:

- **Baseline**: Classify recurrent segments using only slot-by-slot features described in Section 5.2.2.

- **Baseline+DOF**: Classify recurrent segments using the slot-by-slot features and DOF described in Section 5.2.3.

Moreover, the performance of DOF may be dependent on the number of OOV word repetitions since DOF represents the statistics of multiple features. Thus we compared the detection performance of OOV words repeated at least twice and that of OOV words repeated 5 or more times in a lecture. The true OOV segments appearing once in a lecture are ignored (i.e. not classified as OOV and not counted as misses) in "freq $\geq$ 2" condition, and those appearing 4 or fewer times in a lecture are ignored in "freq $\geq$ 5" condition.

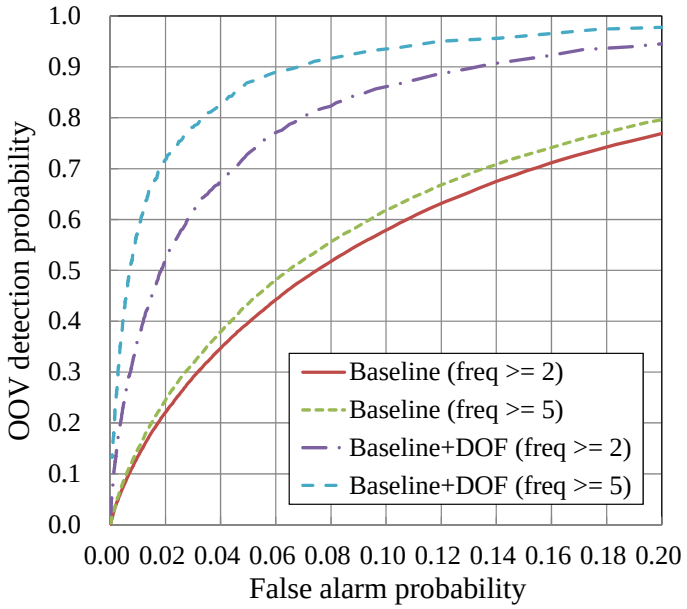

Figure 5.3: ROC curve of recurrent OOV word detection with/without DOF (Clean).

5.3.3 Results

5.3.3.1 Detection performance

The ROC curves are shown in Figures 5.3, 5.4 and 5.5. The two curves yielded with DOF lie above the curves created using only slot-by-slot features in all conditions. At any OOV detection probability, the use of DOF yielded an over 60% relative reduction in false alarms. This result confirms that the DOF extracted by the proposed framework dramatically reduces the detection errors of recurrent OOV words.

In both "Baseline" and "Baseline+DOF" conditions, detection error rate in "freq. $\geq$ 5" were lower than those in the "freq. $\geq$ 2" condition, but larger improvement was yielded when our DOF was used. This means that our framework effectively utilizes the repeated appearance of OOV words. While our DOF is effective in detecting OOV words repeated at least twice, it becomes more powerful as the number of OOV word repetitions increases.

When the decision threshold was set to yield the false alarm probability of 5% in the clean condition, the ratio of the number of false alarms in

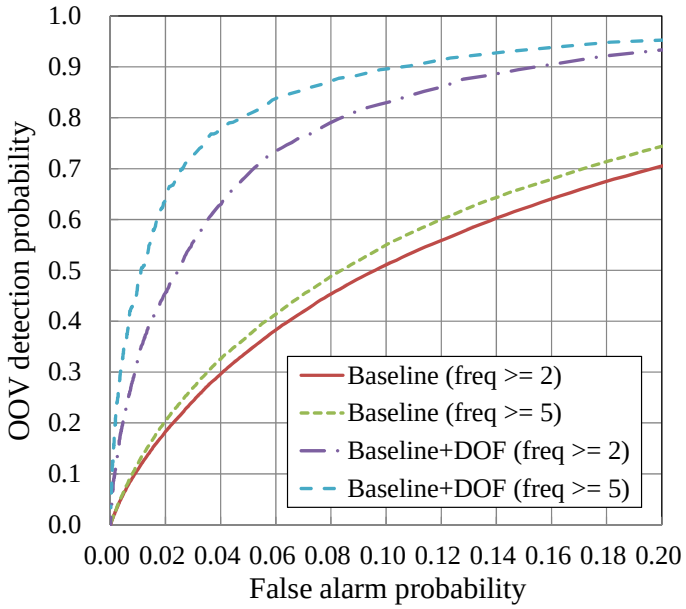

Figure 5.4: ROC curve of recurrent OOV word detection with/without DOF (10db).

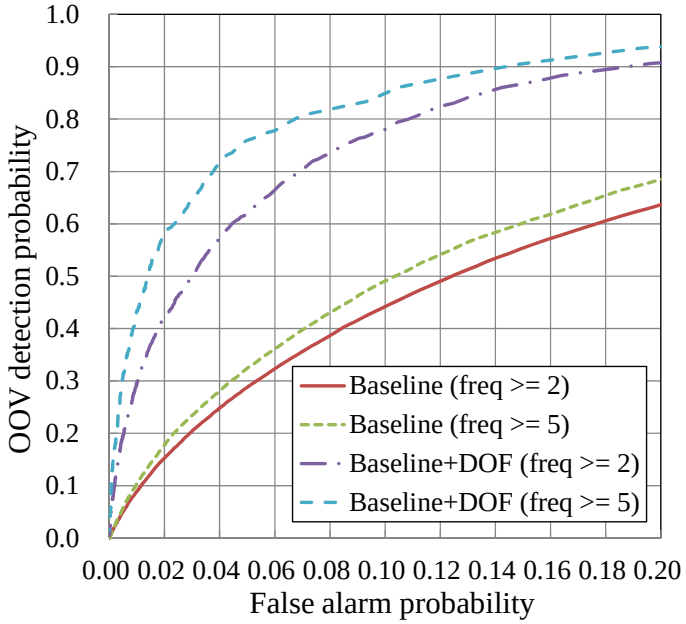

Figure 5.5: ROC curve of recurrent OOV word detection with/without DOF (5db).

Table 5.4: Examples of correct detection in "Baseline+DOF (freq $\geq$ 5)" condition.

Detected word	# of repetitions	Category
shinjiko ("Lake Shinji")	12	Name of place
matsumori-akiko ("Akiko Matsumori")	10	Name of person
juman ("JUMAN")	18	Name of system
enpo-tadamasa-bon ("Enpo Tadamasa Book")	6	Name of ancient text
tandemu ("tandem")	5	Technical term
nomoguramu ("nomogram")	7	Technical term
ten-yaku ("translation into Braille")	30	Technical term

Table 5.5: Examples of false alarms in "Baseline+DOF (freq $\geq$ 5)" condition.

Detected word	# of repetitions	Category
joutan-katan ("top and bottom end")	5	Rare phrase with IV words
tatami-komu ("convolute")	7	Rare phrase with IV words
ee-kono ("well this")	6	Disfluency with consistency

disfluency (filler) segments to the total number of false alarms in "Baseline (freq $\geq$ 2)" and "Baseline+DOF (freq $\geq$ 2)" conditions were 17.2% and 9.4%, respectively. This confirms that our DOF can effectively reduce false alarms created by vocal irregularities as expected.

Noise also causes false alarms. In noisy conditions (Figure 5.4 and 5.5), OOV detection performance was degraded from the clean condition. However, the degradation by noise was alleviated by using DOF as was seen for vocal irregularities. For example, at the OOV detection probability of 50%, the false alarm probability of "Baseline+DOF (freq $\geq$ 2)" was degraded 0.5% from clean to 10dB, whereas "Baseline (freq $\geq$ 2)" was degraded 2%.

5.3.3.2 Example analyses

In order to more fully understand the property of the proposed method, we analyzed individual detection results in several lectures. In this section, we pick up and discuss some important examples. Detected words and their

Table 5.6: Examples of misses in "Baseline+DOF (freq ≥ 5)" condition.

Missed word	# of repetitions	Category
shindoushi ("oscillator")	5	Existence of IV homonyms
kaisetsu ("diffraction")	7	Existence of IV homonyms
suikou ("revise")	15	Existence of IV homonyms

phonetic transcriptions shown in Tables 5.4 and 5.5 were manually extracted from correct transcriptions of the test sets according to the detected timestamps and true OOV timestamps obtained by forced alignment.

Table 5.4 shows examples of correctly detected OOV words in "Baseline+DOF (freq ≥ 5)" in clean speech condition. OOV words that have short and long length and were repeated a few and many times were detected. These words were proper names and technical terms related to the main theme or detailed technical descriptions. It is confirmed that repeatedly uttered OOV keywords could actually be detected by the proposed method.

Table 5.5 shows examples of false alarms raised in "Baseline+DOF (freq ≥ 5)" in clean speech condition. First and second examples, *joutan-katan* and *tatami-komu*, were phrases consisting of IV words (*joutan*, *katan*, *tatami* and *komu* were included in the vocabulary of the recognizer). However, the 2-grams, *joutan katan* and *tatami komu*, occurred only a few times in the training text of the LM. It can be considered that low LM scores are the cause of these false alarms. Though these examples were falsely detected as OOV words, the addition of these phrases as composite words to the recognizer's vocabulary would reduce LM mismatch. Thus these examples are considered as false but beneficial alarms.

The third example in Table 5.5 is obviously a false alarm caused by disfluency. Although the proposed method reduces the impact of disfluencies by utilizing the degree of consistency as DOF, the speaker of the lecture had tendency to insert *ee* ("well") before *kono* ("this"). This example broke our assumption that vocal irregularities have weak consistency. While it indicates the limitation of DOF-based OOV detection, most disfluencies actually have weak consistency and the overall number of disfluency-caused false alarms was reduced as described in Section 5.3.3.1.

Table 5.6 shows examples of missed OOV words in "Baseline+DOF (freq $\geq$ 5)" in clean speech condition. In our analyses, the existence of IV homonyms was found to be a typical cause of misses. For example, the missed OOV word, *kaisetsu* ("diffraction"), has a homonym, *kaisetsu* ("explanation"), in Japanese. The IV homonyms likely have high posterior probability in the CN slot corresponding to the OOV word. This makes the fragment posterior and the word entropy small, and classifying the slot as OOV becomes difficult. This is seen as another problem of the conventional word/fragment hybrid ASR-based OOV detection approach and is not addressed in this study.

5.4 Summary

This chapter presented a novel framework to extract effective features for detecting recurrent OOV words in a spoken document, such words seriously degrade the performance of speech recognizers. In order to deal with the sensitivity to disfluencies and improve the robustness of OOV word detection, we focused on the consistency of recurrent OOV words observed in a spoken document. The proposed method first discovers recurrent segments wherein the same word is uttered by using a phoneme recognizer, and uses the means and variances of slot-by-slot features corresponding to the recurrent segments as DOF for IV/OOV classification.

Experiments on 2701 academic lectures showed that the use of DOF achieves over 60% relative reduction of false alarms in both clean and noisy conditions. We also confirmed that our framework effectively reduces false alarms due to disfluencies and noise by utilizing the repetition of OOV words; our DOF becomes more effective as the number of repetitions increases. Detailed analyses of detection results revealed that the proposed method could actually detect OOV keywords in the lecture set.

The substantial improvement yielded by the proposed framework confirmed that the consistency information observed in a spoken document can greatly contribute to OOV word detection.

Chapter 6

Conclusions

6.1 Usage of the proposed methods in SDP systems

This thesis proposed document-level/word-level CM estimation and OOV word detection methods for improving SDP systems. This section qualitatively discusses the usage of the methods in SDP systems.

Both the document-level and the word-level CM estimation method take the error rejection approach, and are intended to be used simultaneously in SDP systems. The document-level rejection removes ill-recognized transcripts, while the word-level rejection further cleanses the remaining transcripts. Therefore, the decision threshold of document-level CM is usually set to a low value so as to remove poor quality transcripts. The decision threshold of the word-level CM should be adjusted according to the importance of precision.

Since the OOV word detection method belongs to the feedback approach, it is used independently of the CM estimation methods. Actually, OOV words such as the names of new products are generated day by day. Thus, OOV word detection should be periodically applied to keep the lexicon of the ASR system fresh so as to catch and use the new words. Note that some of these new words will have short lifetimes. In order to prevent unlimited expansion of the lexicon, it is effective to add only OOV words appearing in the latest spoken documents to the basic lexicon.

6.2 Summary of thesis

This thesis has presented novel methods for improving the reliability of speech recognizer outputs in SDP systems.

In Chapter 1, the background of this study, the details and problems of SDP systems, and approaches investigated in this thesis were described. The two main approaches described are "error rejection" and "feedback and reprocessing," and the sub-approaches for error rejection and feedback approaches were document-level/word-level CM estimation and OOV word detection, respectively. Furthermore, the main concept underlying all proposed methods was described; the use of global consistency information observed in multiple utterances or multiple spoken documents is essential to improve CM estimation and OOV word detection.

Chapter 2 first provided a brief explanation of the general framework of ASR, and then summarized conventional methods for CM estimation and OOV word detection and their issues in SDP systems.

Chapter 3 presented a novel document-level CM estimation method based on long-range contextual consistency information. The proposed method formulated contextual consistency in a context window that covers several consecutive utterances as an average PMI between word pairs in the window, and used it to generate contextual CMs of the document. A smoothing method that deals with two problems of PMI triggered by data sparseness was also proposed. Experiments showed that the proposed document-level CM yielded high correlation coefficients between CMs and true recognition rates, 0.721. It was also confirmed that the improvement of CMs actually increased the precision of keyword search on spoken documents. Chapter 3 confirmed that the long-range contextual information observed in several utterances effectively complements the short-range information obtained from one utterance.

In Chapter 4, an unsupervised word-level CM estimation method that focuses on consistency information observed in multiple documents was proposed. The issue that conventional methods cannot be applied to SDP systems in practice due to the cost of making human-labeled training data was addressed by a completely unsupervised framework that utilizes transcripts

stored in deployed systems; it dispenses with the need for human-labeled training data. In order to calibrate the CM of the target word by using consistency of word sequences; similar word sequences existing in stored transcriptions are extracted as examples. The CM of the target word is updated as similarity weighted average of the examples. Analysis of the standard deviation of the CMs revealed that this calibration stabilized the word CMs. Experiments showed that the proposed word CM significantly improved the performance of incorrect word rejection over conventional WPPs. Chapter 4 confirmed that the consistency information observed in multiple spoken documents is essential for word-level CM estimation.

In Chapter 5, an OOV word detection method that uses the degree of consistency among multiple occurrences of same phoneme sequence was proposed. The weakness of conventional methods, they raise many false alarms due to disfluencies in spoken documents was addressed by utilizing the consistency information to separate true OOV words from disfluencies. The proposed method first detects recurrent segments, segments that contain the same phoneme sequence in a spoken document by open vocabulary spoken term discovery using a phoneme recognizer. Then, the degree of consistency is measured by using the distribution (mean and variance) of features (DOF) derived from the recurrent segments, and use the DOF for IV/OOV classification. Experiments illustrated that the proposed method could more robustly detect recurrent OOV words than the conventional method. It was also confirmed that detection performance improved as the OOV words are repeated more often. Chapter 5 confirmed that the consistency information observed in a single spoken document offers a significant enhancement to OOV word detection.

Through the three studies, this thesis made following contributions:

- Reliability of speech recognizer outputs in SDP systems can be improved by using the document-level/word-level CM estimation and the recurrent OOV word detection methods that introduce global consistency information.

- Both document-level/word-level CM estimation and OOV word detection are technologies that realize error aware systems. This thesis

confirmed that global consistency information improves CM estimation and OOV word detection. This means that this thesis revealed key components of the information needed for achieving error awareness: The long-range contextual consistency information observed in multiple utterances for global document recognition errors, the consistency information observed on multiple recognized word examples in multiple spoken documents for incorrect recognition of words, and the consistency information observed in recurrently appearing phoneme sequences for errors caused by OOV words.

6.3 Future work

Future work will focus on the following issues: 1) Investigation of more sophisticated consistency representations, 2) reduction in training data requirements, and 3) achieving awareness of other types of errors.

This study used basic representations of consistency; average PMI in document-level CM estimation, weighted average of examples in word-level CM estimation, and means and variances of features in OOV word detection. We believe that there are more sophisticated representations that will yield even greater improvements in error awareness. For example, the neural network-based distributed representation of words [33,38] is promising for computing the contextual consistency and similarity between word sequences.

Training data required for computing the contextual consistency and constructing the IV/OOV classifier should be reduced to apply the proposed methods to many domains at reasonable cost. Domain adaptation techniques such as [17] will be necessary for the further growth of SDP systems.

This thesis addressed three types of errors: global document recognition errors, word recognition errors, and OOV words. However, there are other types of errors in speech processing such as missing truly uttered IV words, and voice activity detection errors. These were deemed beyond the scope of this thesis but should be tackled to achieve really intelligent speech processing systems.

Bibliography

[1] C. Biemann. Chinese whispers: An efficient graph clustering algorithm and its application to natural language processing problems. In *the first workshop on graph based methods for natural language processing*, pages 73–80, 2006.

[2] L. Burget, P. Schwarz, P. Matejka, M. Hannemann, A. Rastrow, C. White, S. Khudanpur, H. Hermansky, and J. Cernocky. Combination of strongly and weakly constrained recognizers for reliable detection of OOVs. In *IEEE ICASSP*, pages 4081–4084, 2008.

[3] N. Camelin, F. Bechet, G. Damnati, and R. D. Mori. Speech mining in noisy audio message corpus. In *INTERSPEECH*, pages 2401–2404, 2007.

[4] K. W. Church and P. Hanks. Word association norms, mutual information, and lexicography. In *ACL*, pages 76–83, 1989.

[5] K. W. Church and R. L. Mercer. Introduction to the special issue on computational linguistics using large corpora. *Computational Linguistics*, 19(1):1–24, 1993.

[6] G. E. Dahl, D. Yu, L. Deng, and A. Acero. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE/ACM Transaction on Audio, Speech, and Language Processing*, 20(1):30–42, 2012.

[7] R. Dufour, G. Damnati, and D. Charlet. Automatic error region detection and characterization in LVCSR transcriptions of TV news shows. In *IEEE ICASSP*, pages 4445–4448, 2012.

[8] J. Fayolle, F. Moreau, C. Raymond, G. Gravier, and P. Gros. CRF-based combination of contextual features to improve a posterior word-level confidence measures. In *INTERSPEECH*, pages 1942–1945, 2010.

[9] T. Fuchi and S. Takagi. Japanese morphological analyzer using word co-occurrence -JTAG-. In *COLING-ACL*, pages 409–413, 1998.

[10] T. Fukutomi, S. Kobashikawa, T. Asami, T. Shinozaki, H. Masataki, and S. Takahashi. Extracting call-reason segments from contact center dialogs by using automatically acquired boundary expressions. In *IEEE ICASSP*, pages 5584–5587, 2011.

[11] W. A. Gale. Good-turing smoothing without tears. *Quantitative Linguistics*, 2(3):217–237, 1995.

[12] G. Guo, C. Huang, H. Jiang, and R. H. Wang. A comparative study on various confidence measures in large vocabulary speech recognition. In *ISCSLP*, pages 9–12, 2004.

[13] M. A. Haidar and D. O'Shaughnessy. Unsupervised language model adaptation using LDA-based mixture models and latent semantic marginals. *Computer Speech and Language*, 29(1):20–31, 2015.

[14] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury. Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal Processing Magazine*, 29(6):82–97, 2012.

[15] T. Hori, C. Hori, Y. Minami, and A. Nakamura. Efficient WFST based one-pass decoding with on-the-fly hypothesis rescoring in extremely large vocabulary continuous speech recognition. *IEEE Transaction on Audio, Speech and Language Processing*, 15(4):1352–1365, 2007.

[16] T. Hori, Y. Kubo, and A. Nakamura. Real-time one-pass decoding with recurrent neural network language model for speech recognition. In *IEEE ICASSP*, pages 6364–6368, 2014.

[17] H. Daume III. Frustratingly easy domain adaptation. In *ACL*, pages 256–263, 2007.

[18] H. Jiang. Confidence measures for speech recognition: A survey. *Speech Communication*, 45:455–470, 2005.

[19] K. Katsurada, S. Sawada, S. Teshima, Y. Iribe, and T. Nitta. Evaluation of fast spoken term detection using a suffix array. In *INTERSPEECH*, pages 909–912, 2011.

[20] S. M. Katz. Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Transaction on Acoustic, Speech and Signal Processing*, 35(3):400–401, 1978.

[21] R. Kneser and H. Ney. Improved backing-off for M-gram language modeling. In *IEEE ICASSP*, pages 181–184, 1995.

[22] S. Kombrink, M. Hannemann, and L. Burget. *Out-of-vocabulary word detection and beyond*, pages 57–65. Springer, 2012.

[23] H. K. Kuo, E. E. Kislal, L. Mangu, H. Soltau, and T. Beran. Out-of-vocabulary word detection in a speech-to-speech translation system. In *IEEE ICASSP*, pages 7158–7162, 2014.

[24] B. Lecouteux, G. Linares, and B. Favre. Combined low level and high level features for out-of-vocabulary word detection. In *INTERSPEECH*, pages 1187–1190, 2009.

[25] K. Maekawa, H. Koiso, S. Furui, and H. Isahara. Spontaneous speech corpus of Japanese. In *LREC*, pages 947–952, 2000.

[26] L. Mangu, E. Brill, and A. Stolcke. Finding consensus in speech recognition: word error minimization and other applications of confusion networks. *Computer Speech and Language*, 14(4):373–400, 2000.

[27] A. D. Marcoa and R. Navigli. Clustering and diversifying web search results with graph-based word sense induction. *Computational Linguistics*, 39(3):709–754, 2013.

[28] A. Marin, T. Kwiatkowski, M. Ostendorf, and L. Zettlemoyer. Using syntactic and confusion network structure for out-of-vocabulary word detection. In *IEEE SLT*, pages 159–164, 2012.

[29] H. Masataki, D. Shibata, Y. Nakazawa, S. Kobashikawa, A. Ogawa, and K. Ohtsuki. VoiceRex – Spontaneous speech recognition technology for contact-center conversations. *NTT Technical Review*, 5(1):22–27, 2007.

[30] E. McDermott, S. Watanabe, and A. Nakamura. Discriminative training based on an integrated view of MPE and MMI in margin and error space. In *IEEE ICASSP*, pages 4894–4897, 2010.

[31] X.-L Meng, R. Rosenthal, and D. B. Rubin. Comparing correlated correlation coefficients. *Psychological Bulletin*, 111(1):172–175, 1992.

[32] T. Mikolov, M. Karafiat, J. Cemocky, and S. Khudanpur. Recurrent neural network based language model. In *INTERSPEECH*, pages 1045–1048, 2010.

[33] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, pages 3111–3119, 2013.

[34] C. H. Mooney and J. F. Roddick. Sequential pattern mining – Approaches and algorithms. *ACM Computing Surveys*, 45(2):19:1–19:39, 2013.

[35] H. Nishizaki, H. Furuya, S. Natori, and Y. Sekiguchi. Spoken term detection using multiple speech recognizers' outputs at ntcir-9 spokendoc std subtask. In *NTCIR-9 Workshop Meeting*, pages 236–241, 2011.

[36] C. Parada, M. Dredze D. Filimonov, and F. Jelinek. Contextual information improves OOV detection in speech. In *NAACL*, pages 216–224, 2010.

[37] J. Pei, J. Han, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal, and M. C. Hsu. PrefixSpan: Mining sequential patterns efficiently by prefix-projected pattern growth. In *ICDE*, pages 215–224, 2001.

[38] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543, 2014.

[39] J. Pinto, I. Szoke, S. Prasanna, and H. Hermansky. Fast approximate spoken term detection from sequence of phonemes. In *SSCS 2008: Speech search workshop at SIGIR*, pages 28–33, 2008.

[40] L. Qin and A. Rudnicky. Finding recurrent out-of-vocabulary words. In *INTERSPEECH*, pages 2242–2246, 2013.

[41] A. Rastrow, A. Sethy, and B. Ramabhadran. A new method for OOV detection using hybrid word/fragment system. In *IEEE ICASSP*, pages 3953–3956, 2009.

[42] G. Saon, T. Sercu, S. Rennie, and H.-K. J. Kuo. The IBM 2016 English conversational telephone speech recognition system. In *INTERSPEECH*, pages 7–11, 2016.

[43] F. Seide, G. Li, X. Chen, and D. Yu. Feature engineering in context-dependent deep neural networks for conversational speech transcription. In *IEEE ASRU*, pages 24–29, 2011.

[44] J. Sohn, N. S. Kim, and W. Sung. A statistical model-based voice activity detection. *IEEE Signal Processing Letters*, 6(1):1–3, 1999.

[45] A. Stolcke. Entropy-based pruning of backoff language models. In *DARPA Broadcast News Transcription and Understanding Workshop*, pages 270–274, 1998.

[46] L. V. Subramaniam, T. A. Faruquie, S. Ikbal, S. Godbole, and M. K. Mohania. Business intelligence from voice of customer. In *ICDE*, pages 1391–1402, 2009.

[47] P. Swietojanski, J. Li, and S. Renals. Learning hidden unit contribution for unsupervised acoustic model adaptation. *IEEE/ACM Transaction on Audio, Speech, and Language Processing*, 24(8):1450–1463, 2016.

[48] S. Tsakalidis, X. Zhuang, R. Hsiao, S. Wu, P. Natarajan, R. Prasad, and P. Natarajan. Robust event detection from spoken content in consumer domain videos. In *INTERSPEECH*, pages 2101–2104, 2012.

[49] F. Wessel, R. Schluter, K. Macherey, and H. Ney. Confidence measures for large vocabulary continuous speech recognition. *IEEE Transaction on Speech and Audio Processing*, 9(3):288–298, 2001.

[50] C. White, J. Droppo, A. Acero, and J. Odell. Maximum entropy confidence estimation for speech recognition. In *IEEE ICASSP*, pages 809–812, 2007.

[51] W. Xiong, J. Droppo, X. Huang, F. Seide, M. Seltzer, A. Stolcke, D. Yu, and G. Zweig. The Microsoft 2016 conversational speech recognition system. In *IEEE ICASSP*, pages 5255–5259, 2017.

[52] D. Yu, J. Li, and L. Deng. Calibration of confidence measures in speech recognition. *IEEE Transaction on Audio, Speech and Language Processing*, 19(8):2461–2473, 2011.

[53] X. L. Zhang and J. Wu. Deep belief networks based voice activity detection. *IEEE Transaction on Audio, Speech, and Language Processing*, 21(4):697–710, 2013.

List of Publications

Journal Paper

- Taichi Asami, Ryo Masumura, Yushi Aono and Koichi Shinoda, "Recurrent out-of-vocabulary word detection based on distribution of features," Computer Speech and Language, accepted, 2019.
- Taichi Asami, Narichika Nomoto, Satoshi Kobashikawa, Yoshikazu Yamaguchi, Hirokazu Masataki and Satoshi Takahashi, "Confidence estimation of spoken document recognition using word contextual coherence," Journal of the Acoustic Society of Japan, vol.68, no.7, pp.323–330, 2012 (in Japanese).

International Conference Paper (Refereed)

The presenter is underlined.

- Taichi Asami, Ryo Masumura, Yushi Aono and Koichi Shinoda, "Recurrent out-of-vocabulary word detection using distribution of features," INTERSPEECH, pp.1320-1324, 2016.
- Taichi Asami, Satoshi Kobashikawa, Hirokazu Masataki, Osamu Yoshioka and Satoshi Takahashi, "Unsupervised confidence calibration using examples of recognized words and their contexts," INTERSPEECH, pp.2217-2221, 2013.
- Taichi Asami, Narichika Nomoto, Satoshi Kobashikawa, Yoshikazu Yamaguchi, Hirokazu Masataki and Satoshi Takahashi, "Spoken document confidence estimation using contextual coherence," INTERSPEECH, pp.1961-1964, 2011.

Domestic Conference Paper (Non-refereed)

The presenter is underlined.

- Taichi Asami, Satoshi Kobashikawa, Hirokazu Masataki, Osamu Yoshioka and Satoshi Takahashi, "Evaluation of unsupervised example-based confidence calibration," ASJ Spring meeting, 2-9-5, pp.47–48, 2013 (in Japanese).
- Taichi Asami, Hirokazu Masataki, Osamu Yoshioka and Satoshi Takahashi, "Unsupervised case-based calibration of word confidence measures," ASJ Autumn meeting, 2-1-6, pp.69–70, 2012 (in Japanese).
- Taichi Asami, Narichika Nomoto, Satoshi Kobashikawa, Yoshikazu Yamaguchi, Hirokazu Masataki and Satoshi Takahashi, "Spoken document confidence estimation using smoothed pointwise mutual information," ASJ Autumn meeting, 2-10-3, pp.57–58, 2011 (in Japanese).
- Taichi Asami, Satoshi Kobashikawa, Yoshikazu Yamaguchi, Hirokazu Masataki and Satoshi Takahashi, "Evaluation of confidence estimation using word contextual coherence and acoustic likelihood," ASJ Spring meeting, 3-5-20, pp.133–136, 2011 (in Japanese).
- Taichi Asami, Satoshi Kobashikawa, Yoshikazu Yamaguchi, Hirokazu Masataki and Satoshi Takahashi, "Confidence estimation at the spoken document level using word contextual coherence and acoustic likelihood," IEICE Technical Report, vol.110, no.143, pp.43–48, 2010 (in Japanese).

Award

- The Awaya Prize Young Researcher Award, Acoustic Society of Japan, 2012.
- The Sato Prize Paper Award, Acoustic Society of Japan, 2014.
- IEICE ISS Young Researcher's Award in Speech Field, 2016.