Noise-robust Open-vocabulary Information Retrieval in Large Spoken Document Collection

Von der Fakultät IV - Elektrotechnik und Informatik
der Technischen Universität Berlin
zur Verleihung des akademischen Grades
Doktor der Ingenieurwissenschaften
- Dr.-Ing. -
genehmigte Dissertation

vorgelegt von

Dipl.-Ing. Shan Jin

geb. in Shaanxi, VR China

Promotionsausschuss
Vorsitzender: Prof. Dr.-Ing. A. Raake
Gutachter: Prof. Dr.-Ing. T. Sikora
Gutachterin: Prof. Dr.-Ing. D. Kolossa (Ruhr-Universität Bochum)
Gutachter: Dr.-Ing. I. Keller

Tag der wissenschaftlichen Aussprache: 27.01.2015

Berlin 2015

D 83

To my parents, my husband Fei and my little daughter Christina

Acknowledgments

First of all I would like to thank my supervisor, Prof. Sikora, head of the Communication Group at the Technical University of Berlin, for his support and guidance throughout the past few years.

I would also like to express my gratitude to my second supervisor, Prof. Kolossa, for her useful advices.

My sincere thanks also goes to the rest of my thesis committee: Dr. Keller and Prof. Raake, for their insightful comments and encouragement.

I thank my fellow in Fachgebiet Nachrichtübertragung, for their excellent support.

Very special thanks goes to my parents, husband and daughter for their advice, support and love throughout.

Abstract

The amount of available spoken information is growing very fast. Consequently, there is an increasing need for effective and efficient approaches for the indexing and retrieval of spoken information.

Classical spoken document retrieval systems are often based on the word transcription provided by an automatic speech recognition system. A large vocabulary word recognizer will be used to transcribe spoken documents. If there are only few errors contained in the recognition transcription of spoken documents, this kind of spoken document retrieval approaches could achieve comparable performance to text-based information retrieval. However, the mismatch between training and application conditions will lead to a high rate of recognition errors. At the same time, the size of the vocabulary will grow with the size of data collection. The growing number of unforeseeable words that are not appearing in the recognizable vocabulary (out-of-vocabulary words) have become the main problem that word-based spoken document retrieval system has to deal with. This thesis focuses on the exploration of spoken document retrieval approaches dealing with misrecognition and the problems caused by out-of-vocabulary words.

We have collected our test data from the Wall Street Journal Corpus. It includes records made under variation in acoustic environment (background music or talking radio), records made from different channels, and records from different speakers. A $20k$ word recognizer has been built for transcribing speech into representations. This recognizer can achieve a word-error rate of 25% on our text collection. In this work, we will only consider the case of single-word queries. About 13% of queries are out-of-vocabulary words. Traditional word-based spoken document retrieval system is built as baseline system. It reaches a mean average value of 61% and a maximal recall rate of 78% on our test collection.

We first explore different word-based spoken document retrieval approaches dealing with misrecognition errors. Experiments with in-vocabulary queries show that enriching recognition transcription with multiple hypotheses is an effective way to compensate misrecognition errors. The maximal recall rate of 95% is yielded by a spoken document retrieval approach based on the word confusion network. The best mean average precision value of 86% is achieved when performing spoken document retrieval on the recognition transcription, including nine best hypotheses.

The experimental results also show that replacing an out-of-vocabulary word with an acoustically similar entry in the recognition vocabulary enables word-based spoken document retrieval systems to deal with out-of-vocabulary words, but with restricted performance. We then study another way to solve the out-of-vocabulary problem using subwords as indexing units. We investigate different indexing units and their ability to index and retrieve text information. The experimental results confirm that indexing spoken document with subword units could achieve acceptable retrieval performance. Nevertheless, we have to make a choice between information coverage and precision. Maximal information coverage could be achieved using phones as indexing units. Different spoken-document retrieval approaches based on phonetic recognition transcriptions are empirically explored in this work. We successfully integrate position information into term weight for phone 3-gram based spoken-document retrieval approaches. This weighting method shows its advantages in dealing with both in-vocabulary and out-of-vocabulary queries. The best performance for out-of-vocabulary queries retrieval is yielded by doing probabilistic string matching on mono-phonetic recognition transcription of spoken documents in the collection.

We propose a new hybrid approach to spoken document retrieval. This method achieves more robust retrieval by combining spoken document retrieval approaches based on the word confusion network and the monophone recognition transcription. The experimental results show that a mean average precision of 56.47% is reached. In comparison with the word confusion network, the mean average precision is improved by about 8.27%. The maximal retrieval recall with the novel hybrid SDR system reaches 91.08%. We also present a prototype with user interface for video retrieval by speech analysis. This prototype deals with queries in normal text form.

Zusammenfassung

Immer mehr Multimediadaten werden der Öffentlichkeit zugängig gemacht und die Menge der Daten nimmt dabei rasant zu. Der Großteil dieser multimedialen Dokumenten besteht zudem aus gesprochenen Informationen. Daher stehen die Anwendungen im Vordergrund, die ermöglichen, die gesprochenen Informationen in großen audiovisuellen Datenbeständen wiederzufinden. Der Retrieval-Ansatz von den gesprochenen Informationen (SDR) verläuft folgendermaßen. Die automatische Sprachererkennung (ASR) und Text Information Retrieval werden hintereinandergeschaltet. Das ASR-System transkribierte die gesprochenen Informationen im Text. Das Text Information Retrieval System dient dazu, die gewünschte Information in der ASR-Ausgabe zu finden. Die ASR-Ausgabe enthält Fehler, die häufig durch die Erkennung der out-of-vocabulary (OOV) Wörter, insbesondere bei Nebengeräuschen auf der Audioaufnahme, entstehen. Die Fehler in der ASR-Ausgabe führen zu einem Informationsverlust. Die robuste Informationswiedergewinnung in den fehlerhaften ASR-Ausgaben stellt eine große Herausforderung dar. Diese Dissertation konzentriert sich auf die Untersuchung von robusten SDR-Ansätzen, die mit den Erkennungsfehlern in der ASR-Ausgabe umgehen und die Probleme, die von OOV-Wrtern ausgelosten werden, vermindern können.

Die in Literatur beschriebenen SDR-Systeme werden nicht mit einheitlichen Datenbeständen evaluiert. Es fehlt ein gemeinsames Datenbestand, der den Leistungsvergleich zwischen verschiedenen SDR-Systemen ermöglicht. Aus diesem Grund, wird ein Testdatenbestand zusammengestellt. Die in dieser Arbeit verwendeten Testdaten stammen aus dem Wall Street Journal Corpus. Sie beinhalten gesprochene Informationen von unterschiedlichen Rednern, die unter verschiedenen akustischen Umgebungen mit abweichenden Aufzeichnungskanälen aufgenommen wurden. Zu den akustischen Umgebungen gehören z.B. Hintergrundmusik, Gespräche oder nebenläufige Sprachsendungen des Radios.

Ein automatisches Spracherkennungssystem (ASR) wurde aufgebaut, um eine Text-

Transkription der gesprochenen Informationen zu erstellen. Der Wortschatz des Spracheerkennungssystems umfasst dabei 20000 erkennbare Wörter. Auf unseren Testdaten erreicht dieses Spracherkennungssystem eine Erkennungsfehlerrate (WER) von bis zu 25%.

In dieser Arbeit, wird nur der Fall von Ein-Wort Anfragen betrachtet. 13% von den ausgewählten Anfragen sind OOV-Wörter. Der klassische SDR-Ansatz, der auf einer Wort-Transkription der gesprochenen Informationen basiert, wird als Vergleichsbasis (Baselinesystem) aufgebaut. Auf unseren Testdaten erhalt das Baselinesystem ein Mean Average Value (mAP) von 61% und eine maximal Recall Rate (max.RE) von 78%.

Zuerst untersuchen wir verschiedene Wort-basierte SDR-Ansätze, die mit Fehlern in der Text-Transkription gesprochener Informationen umgehen können. Die Erkennungsfehler in der ASR-Ausgabe können, durch Einbeziehung von mehreren ASR-Hypothesen, reduziert werden. Mehrere ASR-Hypothesen könnten als N-beste Wortliste, Wortgitter oder Word Confusion Networks (WCN) in der ASR-Ausgabe abgespeichert werden. Verschiedene SDR-Ansätze, die in einer ASR-Ausgabe, die mehrere ASR-Hypothesen beinhaltet, werden untersucht.

Die SDR-Ansätze, die auf der ASR-Ausgabe einschließlich der N-beste ASR-Hypothesen (N-best) basieren, wurden zuerst untersucht. Die Ergebnisse dieser Studie zeigen, dass die Anzahl der eingeschlossenen ASR-Hypothesen in der ASR-Ausgabe einen signifikanten Einfluss auf die Informationswiedergewinnungsleistung hat. Der max.RE steigt mit der zunehmenden ASR-Hypothesen (N). Die beste mAP (ca. $85,4\%$) beobachtet man bei $N = 9$. Wir vergleichen verschiedene Gewichtungsschemen wie z.B. tfidf-Gewichtung- und Wahrscheinlichkeit-Gewichtungsmethode. Die Ergebnisse zeigen, dass die Wahrscheinlichkeit-Gewichtungsmethode die mAP sich um weitere $0,7\%$ verbessern kann.

Die Untersuchung der auf Wortgitter basierenden SDR-Ansätze geht der Frage nach, wie man den Suchraum vernünftig reduzieren kann, sodass die Retrieval-Leistung gehalten wird. Der DNLLR-Wert wird für jede Verbindung in dem Wortgitter berechnet. Die Verbindung in dem Wortgitter, deren DNLLR-Wert unter eine Schwelle liegt, wird als ungültig erkannt und gelöscht. Die DNLLR-Schwelle ($[-118, -90]$) wird durch mehrere Versuche eingestellt. Die beste max.RE ist $94,5\%$ mit einer mAP von $76,2\%$.

WCN gilt als die kompakteste Form eines Wortgitters. Die Gesamtanzahl der Verbindungen in WCN hat sich im Vergleich zu einem normalen Wortgitter um $76,5\%$

reduziert. Eine max.RE von 95, 2% wird erzielt. Eine Vergleichsuntersuchung wird gemacht, um die Leistung verschiedener Gewichtschemen zu erkunden. Unsere Versuchen haben gezeigt, dass wenn die A-posteriori-Wahrscheinlichkeit der Verbindungen in dem WCN direkt als Term-Gewicht eingesetzt wird, dass dann die Genauigkeit niedriger Recall-Stufe verbessert werden kann. Aber das tfidf-Gewichtschema kann bessere mAP und max.RE erzielen. Ein neues Gewichtschema, das die A-posteriori-Wahrscheinlichkeit und tfidf-Gewicht fr Term-Gewicht miteinander kombiniert, wird vorgestellt. Mit dem neuem Gewichtschema, wird die Anzahl der Suchanfragen, die die richtige Antwort in ersten Rang in der Ergebnisliste stehen (E1), deutlich erhöht. Das neue Gewichtschema hat eine max.RE von 95, 23% und eine mAP von 63, 71% erzielt. Die WCN-basierten SDR-Ansätze erreichen die höchsten max.RE.

Die Bedienung der OOV-Suchanfrage von Wort-basierten SDR-Ansätze ist nur dann möglich, wenn mindestens eine von den Methoden (z.B. Suchanfrage-Erweiterung und Dokumente-Erweiterung), im Einsatz ist. Die von Moreau vorgestellte Suchanfrage-Erweiterungsmethode, die die originale Suchanfrage durch seine akustische ähnliche In-Vokabular Wort ersetzt, wird genau untersucht. Die experimentellen Ergebnisse zeigen auch: der Ersatz der OOV-Wörter mit ihrem akustisch ähnlich Eintrag in das Erkennungsvokabular ermöglicht, dass die Wort-basierten SDR-Ansätze die OOV-Anfrage behandeln können. Leider kann diese Lösung nur beschränkte Leistungen erbringen. Daher werden weitere Möglichkeiten untersucht, um die OOV-Probleme zu bekämpfen, wie zum Beispiel die auf Teilwort-Transkription (gesprochener Informationen) basierenden SDR-Ansätze. Das Verfahren der Umwandlung der gesprochenen Informationen in der Text-Transkription entspricht der Indizierungsphase eines Textinformation-Retrieval-Systems. Wir bezeichnen daher die Erkennungseinheit des Spracheerkennungssystems auch als Indizierungseinheit. In dieser Arbeit haben wir die Fähigkeiten verschiedener Teilwort-Indizierungseinheiten in Indizierung und Retrieval auf der Referenztext Transkription der gesprochenen Informationen untersucht. Die experimentellen Ergebnisse bestätigen, dass Teilwort-basierende gesprochene Dokumentabrufsysteme akzeptable Leistung erzielen können. Wir müssen dennoch die Auswahl zwischen Informationenerfassung und -genauigkeit treffen. Die maximale Informationsabdeckung wird mit dem Phon als Indizierungseinheit erreicht.

Die Phon-Transkription der gesprochenen Informationen kann man durch die Anwendung eines Phon-Erkennungssystem gewinnen. Mit Hilfe von einem Aussprache-Wörterbuch kann die Phon-Transkription der gesprochenen Informationen auch direkt von der ASR-Wortausgabe bereitgestellt werden. Experimentelle Ergebnisse weisen darauf hin, dass die Phon-Transkription, die durch Nachbearbeitung der ASR-Wortausgabe entstehen wird, weniger Fehler beinhaltet. Daher wird in weiteren Versuchen der monophon-basierten SDR-Ansätze eingesetzt. Phon-Transkription gewinnt mit zweiter Methode werden in folgenden Untersuchung eingesetzt.

Die SDR-Ansätze, die auf der Phon-3gram-Transkription der gesprochenen Informationen basiert ist, wurden genau untersucht. Die experimentellen Ergebnisse zeigen, dass die von Phon-3gram-basierte SDR-Ansätze erreichte max.RE generell höher als die von den Wort-basierte SDR-Ansätzen sind. Der SDR-Ansatz mit dem tfidf-Gewichtschema hat eine max.RE von $99,5\%$ und eine mAP von $65,2\%$ erreicht. Der SDR-Ansatz, der die Phon-Abwechslungswahrscheinlichkeit in Term-Gewicht integriert, hat keinen Gewinn in der Retrieval-Genauigkeit gebracht. Ein drastischer Verlust in mAP (ca. $25,3\%$) ist nicht zu vermeiden. Mit dem SDR-Ansatz, der die Positionsinformationen in Term-Gewicht integriert (Proximity), kann man eine mAP von $69,94\%$ erreichen. Leider kann der Ansatz, der die PSPL auf Phon-3gram erweitert, keinen Gewinn auf der Retrieval-Leistung bringen. Die auf Phon-3gram basierenden SDR-Ansätze können nicht wirklich mit der OOV-Suchanfrage umgehen. Dies wird auch durch Experimenten nachbewiesen. Es wird untersucht, ob die Abfragelänge eine Wirkung auf die Retrieval-Leistung hat. Bei einer langen Suchanfrage, übertrifft der Proximity-Ansatz alle anderen Phon-3gram basierenden SDR-Ansätze. Der Proximity-Ansatz bietet auch bessere max.RE bei kürzeren Suchanfrage an. Diese Aussagen werden mit zusätzlichen statistischen Signifikanz Tests verifiziert.

Eins von den Schwerpunkten der Untersuchungen von der Monophon-basierte SDR-Ansätze ist, die Ähnlichkeitsschätzungsmethode, die die Ähnlichkeit zwischen entdecktes Segment und der Suchanfrage bewertet, genau zu erforschen. Die INED-Methode nimmt die normalisierte Edit-Distanz als Ähnlichkeit-Score. Die SSPE-Methode integriert die Phon-Verwechslungswahrscheinlichkeit in die Bewertung der Ähnlichkeiten zwischen dem entdeckten Segment und der Suchanfrage. Die experimentellen Ergebnisse haben gezeigt, dass die INED-Methode bessere Retrieval-Leistung anbietet. Die beste max.RE wird von der INED-Methode erreicht. Die INED-Methode erzielt ähnliche mAP wie Phon-3gram basierter Proximity-Ansatz.

Vorherige Forschungsergebnisse haben gezeigt, dass die Wort-basierte SDR-Ansätze hohe mAP bei den in-Vokabular Suchanfragen erzielen können und Phon-basierte SDR-Ansätze ihre Vorteile im Umgang mit OOV-Suchanfragen haben. Basiert auf diese Forschungsergebnisse und die von Lee vorgestellte Information-Fusion Strategie, wird ein neuer Hybrid-Ansatz für den gesprochenen Dokumentenabruf entwickelt. Dieser An-

satz wirkt deutlich robuster im Fall von Erkennungsfehlern und vorkommenden OOV-Suchanfragen. Der neue gesprochene Dokumentabruf-Ansatz basiert auf einer mehrstufigen Transkription der gesprochenen Dokumente. Die mehrstufige Transkription beinhaltet Word-Confusion-Network und die Monophon-Darstellung eines gesprochenen Dokuments. Die experimentellen Ergebnisse zeigen, dass dieser Ansatz eine Mean-Average-Precision von $56,74\%$ erreicht. Im Vergleich zu den SDR-Ansätzen, die nur auf Word-Confusion-Network basieren, erhöht sich die Mean-Average-Precision-Rate um $8,27\%$. Die maximale Retrieval-Rate dieses Ansatzes erreicht bis zu $91,08\%$.

Als Letztes stellen wir ein Prototyp für das Video-Retrieval-System vor. Wir beschreiben die Hauptelemente von geeigneten Benutzerschnittstellen. Die Funktionsblöcke für die Auswahl von den verschiedenen Retrieval- und Fusionmodulen ermöglicht Benutzer den Systemkern zu konfigurieren. Jetzt befasst dieser Prototyp sich nur mit normalen Abfragen in Textform.

Contents

Chapter 1

Introduction

With increasingly powerful computers, data storage capacity and growing international information infrastructure, the amount of accessible data in different digital forms (text, image, audio, video etc.) is rising very fast. Consequently, the development of more effective, efficient methods to process, organize, analyze and use this data is of particular interest and has become a key issue.

We focus on the information retrieval (IR) domain, which is concerned with the representation, storage, organization and access of information items [98], and especially with the problem of selecting relevant items from a large collection of data, given a user's request. In recent years, a lot of research has been carried out in the field of text retrieval [43], and the retrieval of relevant information from other media, such as audio [89], video [65] and speech [26] [44]. Significant progress has been made in text information retrieval. However, the research on effective and efficient methods for information retrieval in other forms, especially in speech, is relatively new.

Speech, as one of the primary means of human communication, is contained in data like broadcasting news, speech recordings, interviews, conference talks, lectures given in universities, etc. A lot of usable semantic information, the so-called spoken content, which consists of the actual spoken words, is enclosed in the speech segments. The retrieval of information in speech is clearly becoming ever more significant.

The following sections provide an introduction to information retrieval, describe information retrieval in spoken documents, define objectives and summarize the scientific contributions. The overall structure of this thesis will be described in the final section of this chapter.

1.1 Information Retrieval

The task of information retrieval is to identify information items within a large collection, that are relevant to a request formed by the user. Generally, a information retrieval system consists of three basic components: indexing, query formation and retrieval. Indexing is a process used to prepare the document representation for use by the information retrieval system. Some factors that need to be considered at this stage: accurate representation of meanings (semantics), exhaustiveness whether all content has been covered, and the facility for a computer to manipulate. Query formation generates the representation for the request of a user, the so-called query. The query must be in the same form as the representation of a spoken document to be retrieved, to enable the comparison between them. The comparison between the query and document representation is realized by the retrieval system. A matching function returns a score representing the similarity between the query and a document. This score indicates how relevant the document is, according to the given query. According to this score, a ranked list of relevant documents will be returned by the retrieval system.

The history of information retrieval systems is presented in Figure 1.1. It can be seen from the Figure 1.1 that the first information retrieval system was proposed by Luhn in 1957 [68]. This system used words as indexing units. Luhn measured the word overlap between query and document and used this information as a criterion for retrieval. He evaluated his system on a small-scale document collection, including 1200 technical reports. The first large-scale information retrieval systems were developed in the 1970's. The most famous one is MEDLINE (Medical Literature Analysis and Retrieval System online) [113]. MEDLINE is an online computerized biomedical bibliographic retrieval system that was launched by the National Library of Medicine in 1971. Since the 1990's, information retrieval techniques have been widely used for searching all documents on the Internet provided by the FTP server, searching the World Wide Web (WWW), building recommender systems and automated text categorization/clustering. In the 21st century, the research on information retrieval systems focused on the development of multimedia information retrieval systems and techniques for cross-language information retrieval and document summarization.

The development of information retrieval systems, especially text information retrieval systems has been conducted over five decades. Significant progress has been made in this area. These automatic methods, however, are still far from perfect. The

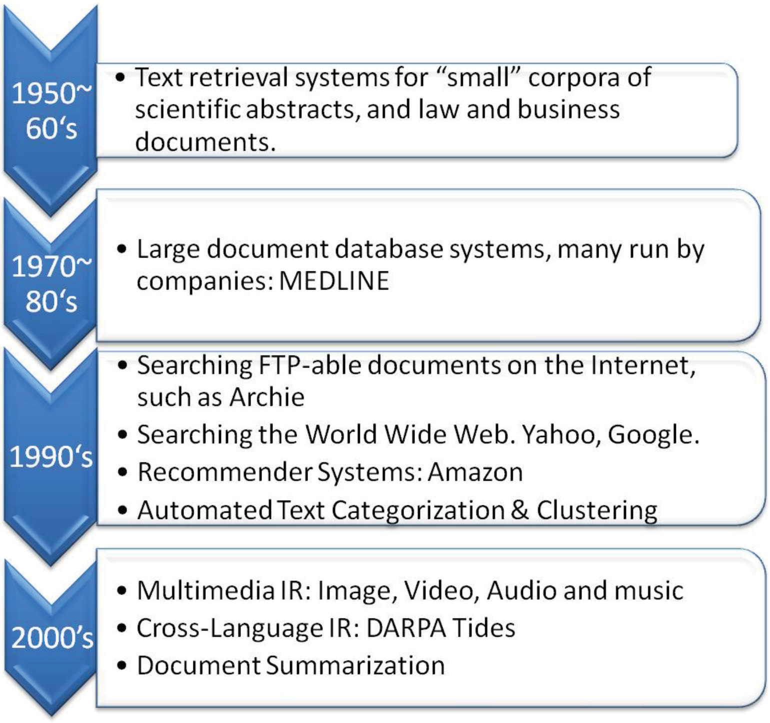

1950~ 60‘s
• Text retrieval systems for “small” corpora of scientific abstracts, and law and business documents.
1970~ 80‘s
• Large document database systems, many run by companies: MEDLINE
1990‘s
• Searching FTP-able documents on the Internet, such as Archie
• Searching the World Wide Web. Yahoo, Google.
• Recommender Systems: Amazon
• Automated Text Categorization & Clustering
2000‘s
• Multimedia IR: Image, Video, Audio and music
• Cross-Language IR: DARPA Tides
• Document Summarization

Figure 1.1: History of information retrieval.

task of automatically indexing, organizing and retrieving collections of information, especially information contained in multimedia data, are still open research problems. This thesis will focus on the issue of automatic methods for information retrieval in spoken document collections.

1.2 Information Retrieval in Spoken Documents

Spoken document retrieval (SDR) is a subdomain of Information Retrieval (IR). It is concerned with retrieving spoken documents in response to a written or spoken query [20]. Speech and text, however, are quite different media. Two main issues need to be addressed for the development of an effective and efficient spoken information retrieval system. The first is how to extract and represent the spoken content accurately in a form that can be efficiently stored and searched. The second one is the issue of how to deal with noise or errors in the transcription.

A text-query-driven spoken document retrieval system is shown in Figure 1.2. Current spoken document retrieval systems can work with written or spoken queries and

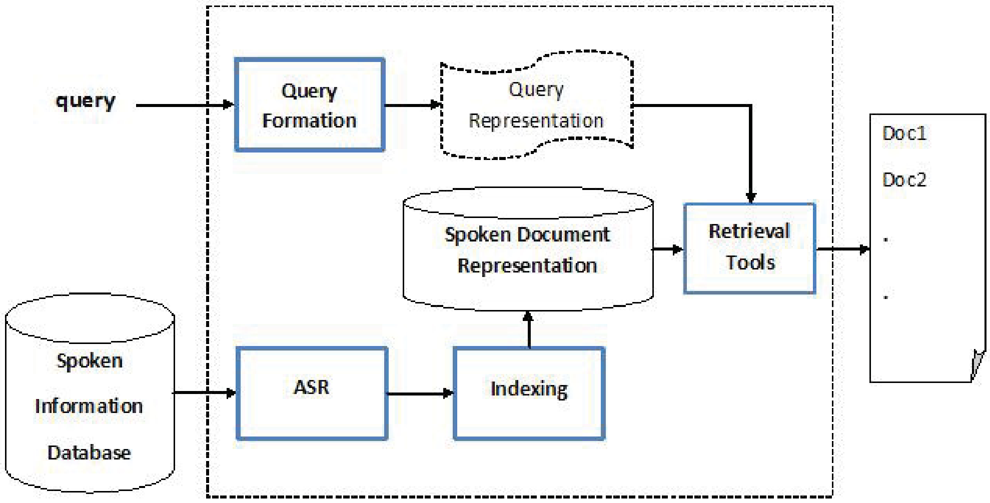


Figure 1.2: Schematic view of a spoken document retrieval system.

will return a ranked list of spoken documents. Generally, a spoken document retrieval system consists of four main components: automatic speech recognition (ASR), indexing, query formation, and retrieval. The automatic speech recognition system transcribes speech into a sequence of predefined units which could be words, syllables or phonemes. The transcription of the spoken documents will then be prepared for the information retrieval system during the indexing stage. After the indexing process, only important information will be kept in the document representation. This process enables the information retrieval system to handle large document collections efficiently. The representation of a query is generated by query formation. In the retrieval stage the representation of the query is compared with the representation of each document in the collection and a similarity score, which will be used to construct a ranked list of documents, is estimated.

In accordance with the type of speech recognizer applied, current spoken document retrieval systems can be classified in four categories: word-based, subword-based, phone-based and combined approaches. **Word-based spoken document retrieval systems** transcribe spoken content in a sequence of words. Text information retrieval methods can then be directly used to find the relevant information for the given queries. The British and North American broadcast news retrieval system presented by Renals [95], and the on-line spoken document retrieval system SpeechFind presented by Zhou [123], are both word-based spoken document retrieval approaches. In the speech track of the TREC [43] conference, Harman has reported that a large vocabulary word recognition system (LVCSR)-based spoken document retrieval system may achieve remarkable results in terms of the retrieval effectiveness.

Subword-based spoken document retrieval systems will transcribe the spoken content into a sequence of predefined subword units, such as vowel-consonant-vowel (VCV) features, presented by Luhn [68]. Glavitsch [38] selected about 1000 VCV units directly from the English text and built a vowel-consonant-vowel feature recognizer for transcribing the spoken content into a sequence of VCVs for information retrieval. **Phone-based approaches** perform information retrieval on the phone transcription of the spoken documents provided by an automatic phone recognizer [31], [82]. **Combined systems** use more than one automatic speech recognizer for transcribing the spoken content into multi-sequences of different units. One representative example is an application of the traditional information retrieval techniques to spoken documents presented by James [51].

1.3 Motivation

As mentioned before, a spoken document retrieval system uses an automatic speech recognizer to transcribe the digital spoken signal into a sequence of predefined terms, which could be further processed by the retrieval procedure. Hence the quality of the ASR system has a significant influence on the effectiveness of the SDR. The quality of an ASR system is evaluated by counting misrecognized terms in the output transcription, the so-called term error rate. The misrecognized terms will cause confusion during the retrieval process. The quality of an ASR system is usually affected by the following factors:

- Recording environment : background noise, microphone directions;
- Speech variability : the temporal and acoustic properties of the same spoken text vary from one recording to another;
- Speech type : continuous speech or isolated spoken word;
- Number of distinct units to be recognized : a large set of recognizable units increases the computation complexity;

- Amount and quality of training data needed to train both acoustic models and language models.

- Number and gender of different speakers. Every speaker has individual speaking speed and pronunciation. Speaker-independent recognition is more difficult than the speaker-dependent recognition.

In order to reduce errors in the transcription, a huge set of training data is required for building reliable acoustic models of the recognizer. However, the size of the recognizer vocabulary is limited and fixed, which leads to a restriction of the query vocabulary. Hence the classical word-based spoken document retrieval system has to deal with the so-called **out-of-vocabulary (OOV)** problem. In order to make the retrieval system operate efficiently and accurately in a large domain, new words have to be integrated into the recognizer vocabulary. Thus the recognition vocabulary increases in size. The entire spoken information collection will have to be re-indexed when the recognizer vocabulary changes. Computationally, this is a very intensive process.

A way to deal with the problem of growing vocabulary without updating the automatic speech recognition system is to use subwords instead of complete words as the recognition units. The use of subword units can also reduce the number of recognition units without losing their descriptive power. However, the recognition performance degrades for shorter units. More misrecognized terms will appear in document representations. If the subword units are directly extracted from the text without considering their acoustic properties, then the retrieval system may not detect a discriminated unit exactly, when there are other units with similar acoustic properties. The phone, as basic sound unit, could also be used to analyze spoken information. The phone-based spoken document retrieval approaches are dependent only on a number of phones which remains fixed with growing vocabulary. The number of unique phones in one language is less than 100. Therefore the construction of a phone recognizer requires less training data than the word-based approaches. At the same time, the phone recognition can be processed considerably faster this way than using a large vocabulary continuous word recognizer. The key problem of phone-based spoken document retrieval systems is the high error rate (phone-error-rate = 50%) in the output transcription. This is due to the fact that a phone recognizer is very sensitive to the mismatch between training and test data. The change of the environment, speech style, etc. could cause a lot of errors in the output transcription.

Using different recognition systems to generate multi-scale representations for the spoken information retrieval task could improve the retrieval effectiveness of a spoken document retrieval system. However, more training data is needed to build reliable recognizers and the indexing will become a very time-consuming task.

1.4 Purpose and Contributions

The main goal of this thesis is to develop effective open-vocabulary spoken document retrieval techniques for information retrieval in spoken document collections containing multiple speakers, different recording environments (with music and talking radio in the background), varying channels (with varying distance between microphone and speaker) and varying speaking styles (clean and spontaneous speech). We are going to tackle two main problems of a spoken document retrieval system: the **out-of-vocabulary** and the **mis-recognition problem**. The following issues are addressed in this work:

- How much improvement can we achieve by using word transcription with multiple hypotheses for information retrieval?
- How many kinds of index units can be used for the spoken document retrieval task and how well will they perform?
- Exploration of the behaviour of the subword units dealing with recognition errors contained in spoken information transcription.
- How can we achieve higher retrieval performance via modifying indexing and retrieval components?

A $20k$ word recognizer has been built for transcribing spoken documents into word representations. It is evaluated with records made under different acoustic situations, including variable speaking style, different kinds of environment noise and variation in channels. Errors in the transcription and their potential effect on retrieval performance are researched in the next stage. In order to reduce information loss due to the misrecognition problem, a transcription including multiple recognition hypotheses is prepared. A number of word-based retrieval methods that work with multiple recognition hypotheses are investigated. Some of the errors in ASR transcription are caused by the out-of-vocabulary words. These kind of errors could be solved by using subwords as indexing units. A number of subword units have been used for the spoken document retrieval task. We thoroughly investigate a set of indexing units and their ability to effectively index and retrieve information. Using phones as indexing units yields the highest information coverage. Therefore, an exploration of the phone-based spoken document retrieval is the other main focus of this work. A new spoken document retrieval method is presented. This method works on the multi-level representation of a spoken document, consisting of the word confusion network and the monophone sequence. The experimental results have verified the improvement achieved with this method and its power of dealing with recognition errors and out-of-vocabulary queries.

The following scientific contributions are made:

- The impact of mismatching between training and application scenario and the out-of-vocabulary words on the recognition quality are researched.
- Word-based SDR methods apply text information retrieval methods on the recognized representation of the spoken documents. The retrieval performance depends on the quality of the recognition output. As a result of the recognition errors in the document representation, some relevant information may be missing during the recognition process. If the recognition errors are caused by the mismatching between the training and application scenario, the enrichment of the document representation with multiple recognition hypotheses can reduce the impact of these kind of recognition errors on the retrieval performance. Multiple recognition hypotheses in the spoken document representation can be presented in the form of the Nbest list, lattice and word confusion networks. The spoken document retrieval methods that work on the Nbest, lattice and word confusion networks are thoroughly investigated.
  - Different document term weighting methods are evaluated for the N-best list-based spoken document retrieval.
  - We research the lattice rescoring method which is based on the durationally-normalized log likelihood ratio ($DNLLR$) and its impact on retrieval performance. We also evaluate the retrieval performance achieved by using the $DNLLR$-value for document term weighting.
  - Word confusion network as more compact lattice is constructed based on the posterior probability of each link in ASR output lattice. We present a novel

  document term weighting method that integrates the posterior probability into the classic term-frequency weighting scheme. This novel weighting method yields the highest retrieval performance.

  - Out-of-vocabulary words are common nouns and names that bear important information. Updating a word recognizer with new words is a method to deal with the out-of-vocabulary words. However, it is a very time-consuming task. Another way to solve the out-of-vocabulary words problem is to expand the queries or document representations. The experimental results have verified that the query and document representation expansion method enables the word-based spoken document retrieval approaches to deal with out-of-vocabulary words, however with poor performance.

- Different subword units are presented in the current literatures. We perform an empirical study on different index units and their behaviour in information retrieval. The highest information coverage is yielded by using phones as indexing units. Therefore we have thoroughly investigated phone-based spoken document retrieval methods.

  - Different weighting schemes for phone-3gram based spoken document retrieval are evaluated. We present a new weighting method which integrates the position information of the phone-3grams into document term weighting and relevance scoring. The experimental results have verified that this novel weighting method could benefit the spoken document retrieval system. An improvement in the mean average precision and the recall rate can be observed.
  - Context independent monophone-based approaches with different scoring methods are also empirically explored.
  - We also investigate the impact of the query-length on the performance of the phone-based spoken document retrieval approaches. Additional statistical significance tests have been made here to confirm the conclusions made about the effect of the query-length on the retrieval performance.

- We present a new spoken document retrieval approach based on multi-level transcription. In our experiment, this new spoken document retrieval method has shown its advantages in handling recognition errors and the out-of-vocabulary words. Based on the exploration results, potential fusion is indicated. Different fusion strategies are also evaluated.

- Based on the findings of those studies, we built a demo for video indexing and retrieving via spoken content analysis. This demo enables the user to configure the spoken document retrieval system by himself via the user interface.

1.5 Thesis Structure

This thesis is structured as follows:

- Chapter 2 provides an introduction into the classical spoken document retrieval system.
- Background information on experiments presented in this thesis is introduced in Chapter 3.
- Chapter 4 investigates several variants of robust word-based retrieval methods working with the transcription containing multiple hypotheses, provided by a automatic speech recognizer.
- Different indexing units are presented and discussed in Chapter 5. In this chapter, several subword-based spoken document retrieval methods are explored and discussed.
- We describe the development of a new hybrid spoken document retrieval method in Chapter 6 and evaluate its performance using spoken document collection with variable acoustic conditions and out-of-vocabulary words.
- We present our demo for speech driven video indexing and retrieving in Chapter 7.
- The findings of this thesis and possible future directions are presented in Chapter 8.

Chapter 2

Fundamentals of Spoken Document Indexing and Retrieval

As a special domain of modern information retrieval, spoken document retrieval consists of spoken document processing, indexing, query formation and retrieval. This chapter describes those three main components in detail: spoken document processing, indexing and retrieving. In the spoken document processing and indexing stage, the digital spoken document will be transcribed in a representation. It contains a sequence of pre-defined index units. This representation can be generated manually or automatically. Nowadays, with growing amount of accessible spoken documents, we prefer the automatic transcribing tools, namely the automatic speech recognition (ASR) . The retrieval element will guide the user to the relevant information corresponding to a given query. This chapter starts with a short introduction to automatic speech recognition systems in the Section 2.1. Detailed descriptions of spoken content indexing and retrieval are provided in Sections 2.2 and 2.3.

2.1 Spoken Document Processing - ASR

The task of the automatic speech recognition system is to transcribe an acoustic signal into a string of predefined units. The predefined units could be phones, syllables, morphs or complete words. The speech recognition task could be formalized as finding the most likely unit string W with $argmax_W P(W|O)$ for feature vectors

$O = o_1, o_2, ...o_T$. This probability is computed according to Bayes' rule:

$$P(W|O) = \frac{P(O|W)P(W)}{P(O)} \tag{2.1}$$

where $P(W)$ is the probability of W; $P(O|W)$ is the acoustic evidence that W is spoken; P(O) is the average probability that feature vector O will be observed. The most probable unit string W could be identified as the one with maximal product $P(O|W)P(W)$.

$$W = argmax_W P(O|W)P(W) \tag{2.2}$$

From the isolated word recognizer reported by Davis et al. [27] to a current speaker-independent large vocabulary continuous speech recognizer, the research in automatic speech recognition by machine has been performed for more than five decades. The existing ASR systems could be classified into four categories: isolated word recognizer, connected word recognizer, keyword spotting system and large vocabulary continuous speech recognizer (LVCSR). Keyword spotting and LVCSR approaches are usually applied in a spoken document retrieval approach for transcribing digital speech record into text representations. The progress made in recognition performance of automatic speech recognition systems can be viewed in Figure 2.1. In Figure 2.1 the recognition performance of an automatic speech recognition system is denoted with the word error

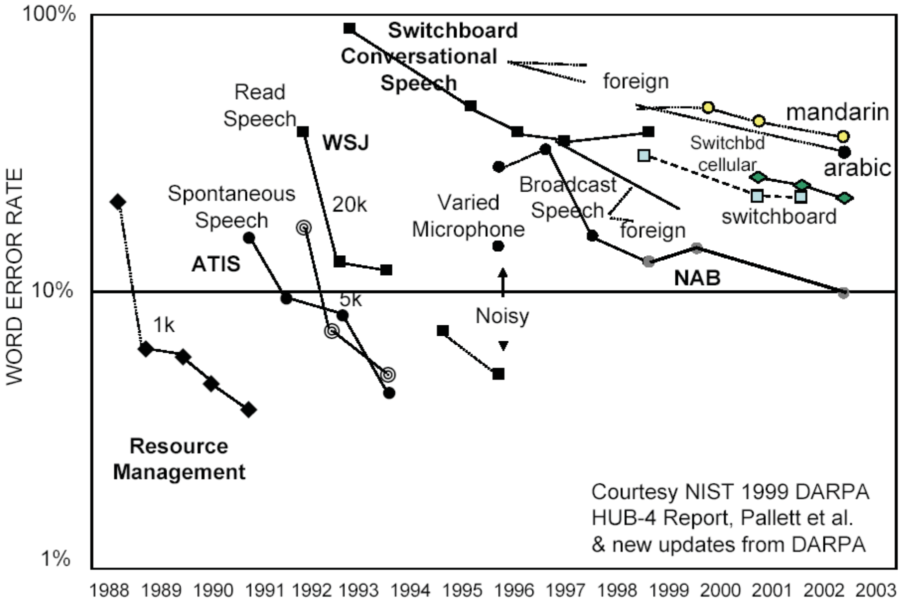


Figure 2.1: Darpa Speech Recognition Benchmark Test [57]

rate defined in Equation 2.3,

$$WER = \frac{\mathrm{S{+}D{+}I}}{\mathrm{N}} \cdot 100\% \tag{2.3}$$

where S represents the number of substitution errors (spoken word w_1 recognized as w_2); D indicates the number of deletion errors (spoken word w_1 missing) and I represents the number of insertion errors (word w_1 is added to the transcription). N denotes the total number of words in the reference transcription.

Two other measures are also very popular in denoting the performance of an automatic speech recognition system. They are the percentage of correctly recognized labels defined in Equation 2.4 and Accuracy defined in Equation 2.5,

$$Correct = \frac{C}{N} \cdot 100\% \tag{2.4}$$

$$Accuracy = \frac{C - I}{\mathrm{N}} \cdot 100\% \tag{2.5}$$

where C represents the number of correct labels; I denotes the number of insertion errors and N is the total number of words in the reference transcription.

The basic structure of an automatic speech recognition system is shown in Figure 2.2.

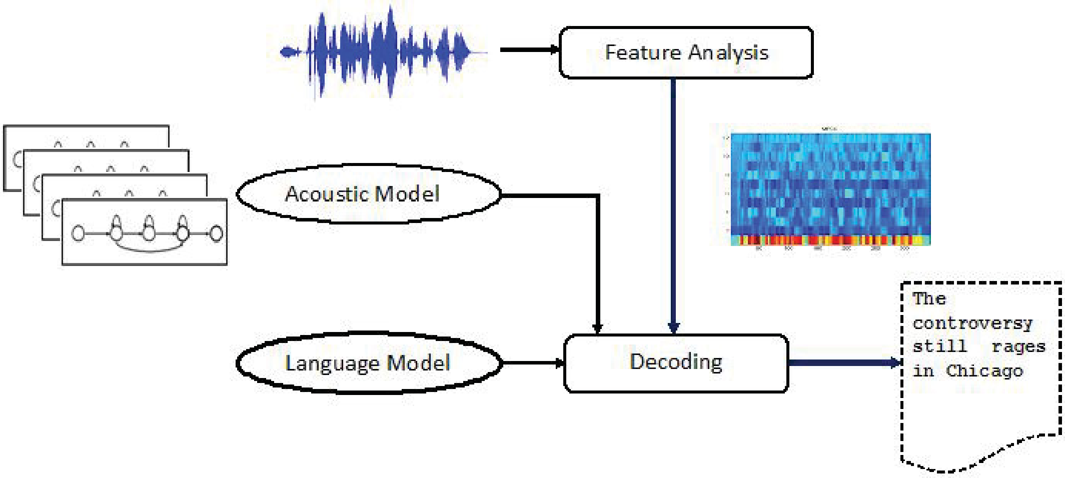


Figure 2.2: ASR basic structure

Generally, an automatic speech recognition system consists of two main components: feature analysis and decoding [50]. The input speech signal is first converted into a sequence of acoustic feature vectors via front-end feature analysis. As core component, the decoder then plays the role of mapping, matching and decision-making for a speech signal input, given the acoustic and language model. The acoustic model is the representation of knowledge about acoustics, phonetics, environment variability and gender. The language model represents the knowledge of what constitutes a possible word, what words could co-occur, and in what sequence. A more detailed description of each component mentioned before can be found in the following sections.

2.1.1 Feature Analysis

The task of feature analysis is to convert the input speech signal into acoustic feature vectors. The performance of an automatic speech recognition system relies heavily on this issue. Perfect features should capture the salient aspects of the speech signal while not be affected by the distortions caused by environmental aspects etc. Significant efforts have been devoted to this issue. Many appropriate acoustic features have been presented in the current literature, such as linear prediction coefficients (LPC), linear predictive cepstral coefficients (LPCC) [72] [93], mel-frequency cepstral coefficients (MFCC) [91] and perceptual linear predictive (PLP) coefficients [46]. Among them LPCC, MFCC and PLP features are widely used for the automatic speech recognition task. It can be seen in the Figure 2.3 that all LPCC, MFCC and PLP analysis algorithms start with a frame blocking to split the digitized speech signal into small intervals in which the speech signal may be treated as stationary. Then Hamming windowing is performed on each frame to minimize the signal discontinuities at the beginning and the end of each frame. Following this step, LPCC, MFCC and PLP will diverge from each other.

LPCC

Linear predictive cepstral coefficients are computed from LPC parameters which are obtained with autocorrelation analysis and subsegment calculation of the prediction coefficients. The m-th cepstral coefficient $c(m)$ is recursively derived from the LPC

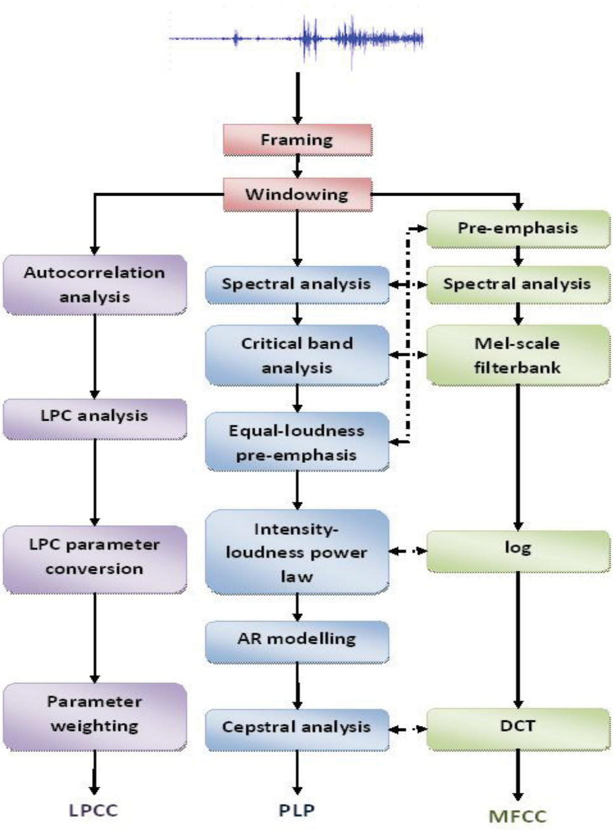


Figure 2.3: Extraction of audio features

parameters as defined in the following equations:

$$c_0 = ln\ \sigma^2 \tag{2.6}$$

$$c_m = a_m + \sum_{k=1}^{m-1} (\frac{k}{m}) c_k a_{m-k},\ 1 \leq m \leq p \tag{2.7}$$

$$c_m = \sum_{k=1}^{m-1} (\frac{k}{m}) c_k a_{m-k},\ m > p \tag{2.8}$$

where σ^2 is the gain term in the LPC model; p is the order of the predictor and a_m are the LPC coefficients.

MFCC

Short-term spectral-based MFCCs represent the speech amplitude spectrum in a compact form [66]. The method was proposed in the early 80's. The creation of MFCCs is motivated by perceptual consideration. The nonlinear relation between perceived frequency and actual frequency was discovered in perception experiments. A couple of perception based frequency scales like the Bark [126] and Mel scale [86] were proposed to describe this nonlinear relation. The Mel scale, as defined in Equation 2.9, is based on the subjective perception of pitch. The signal frequency will be transformed corresponding to this nonlinear human perception of frequency as shown in Figure 2.4.

$$mel = 2595 \log_{10}(\frac{f}{700} + 1) \tag{2.9}$$

where mel means Mel frequency and f is frequency in Hz.

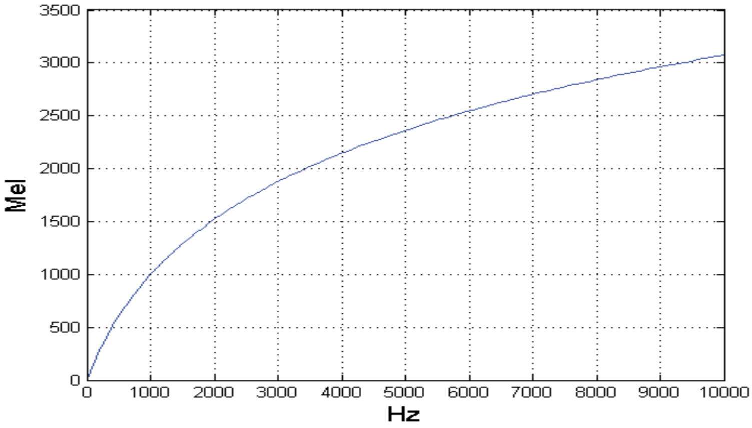


Figure 2.4: Mel Scale

In the MFCC method, the input signal is converted to its auditory representation via Mel-scale frequency warping. As shown in the Figure 2.3, MFCCs are extracted from input speech signal in the following steps:

- **Preemphasis filtering**. A preemphasis filter amplifies the spectrum above $1kHz$, because human hearing is more sensitive in this frequency range. Applying preemphasis filtering assists the spectral analysis algorithm in modelling the perceptually most important aspects of the speech spectrum.

- **Spectral analysis**. The short-term power spectrum is obtained by applying a Fourier transformation to a frame of Hamming windowed speech. The duration of a window is typically 20 to $30ms$.

- **Mel-scaled filterbank** groups neighbouring frequency bins into overlapping triangular banks with equal bandwidth according to the Mel scale, and the contents of each band will be summed.

- **Log**. Calculation of the logarithm for each sum.
- **DCT**. Discrete cosine transformation is used to compute cepstral coefficients from the log mel-filterbank.

PLP

The perceptual linear prediction (PLP) speech analysis technique was proposed by Hermansky [46]. This technique is also based on short term spectra. PLP features are extracted as shown in the middle of the Figure 2.3. Broken arrows link the similar processing stages of PLP and MFCC analysis. Diverging from MFCC analysis, a **bark-scaled filterbank** is applied after spectral analyses. The Bark scale is defined as:

$$bark = 13\arctan(0.00076f) + 3.5\arctan((f/7599)^2) \tag{2.10}$$

where f represent the actual frequency in Hz. Then an equal-loudness preemphasis is performed to compensate for the unequal sensitivity of human hearing across frequency.

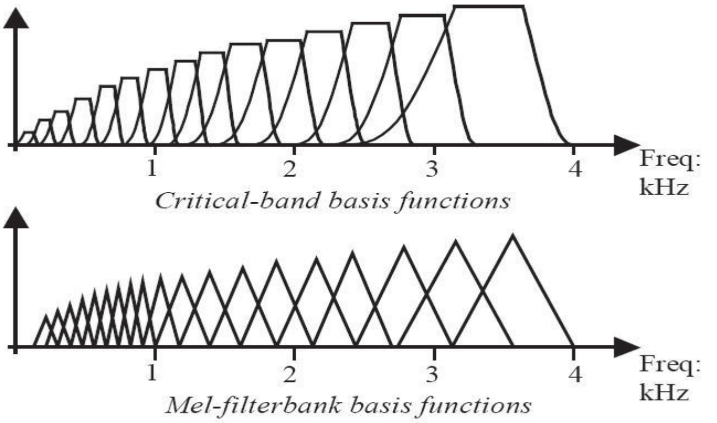


Figure 2.5: Critical band analysis vs mel-filterbank analysis [75]

The similarity of the critical-band (Bark filterbank) analysis and the mel-filterbank analysis can be seen in the Figure 2.5. More filters are allocated to the lower frequencies where hearing is more sensitive to differences in frequency. In contrast to the MFCC analysis, the amplitude responses of the critical-band filters are flat-topped and non-symmetric. **The intensity-loudness power law** models the non-linear relation between the intensity of sound and its perceived loudness (by taking the cubic root of the intensity). In the **AR modelling and Cepstral Analysis** stage, the auto-correlation coefficients are estimated by applying inverse DFT after the application of intensity-loudness law. These autocorrelation coefficients will be transformed into autoregressive coefficients. Cepstral coefficients can then be recursively computed from the autoregressive coefficients.

Acoustic Feature Selection in this Work

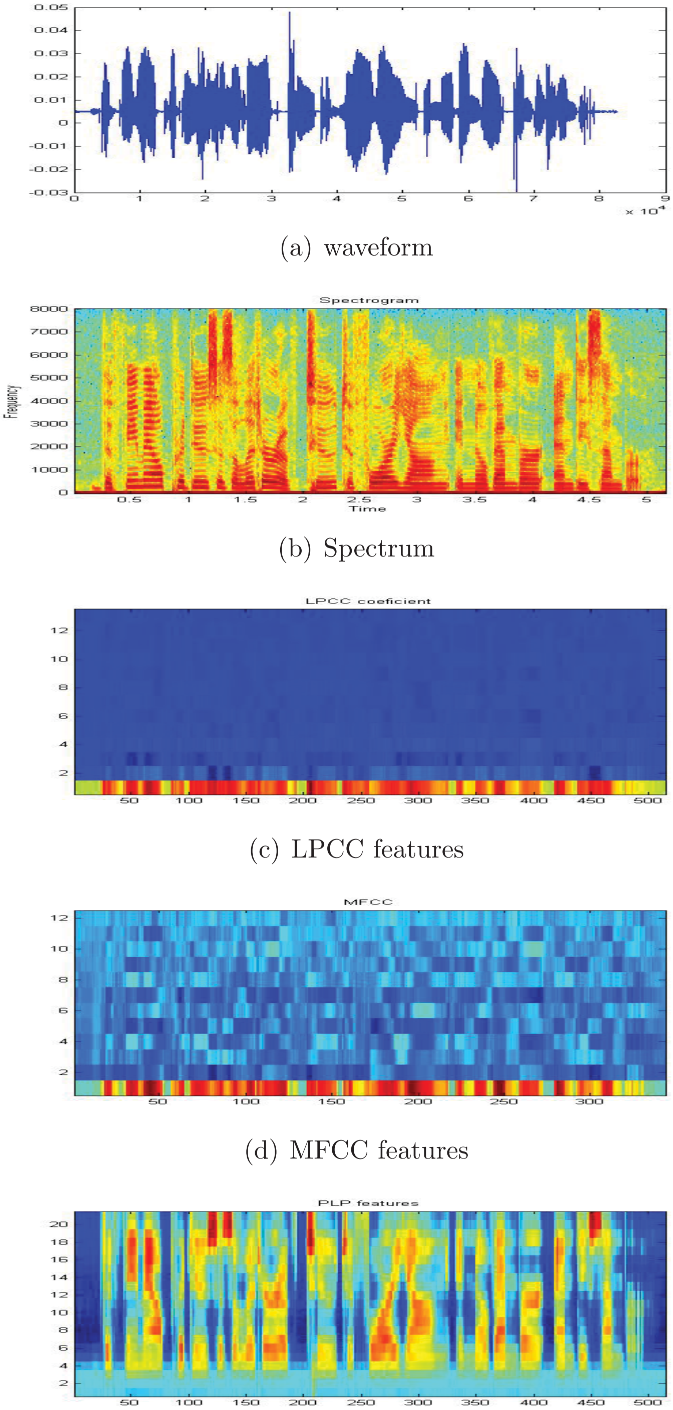
The Figure 2.6 provides an overview of announced acoustic features extracted from the file 4p0a0101 in the WSJ corpus [90]. The performance of an automatic speech recognition system relies heavily on the speech parametrization. Davis [28] compared the MFCCs with LP techniques on a syllable-oriented speaker dependent clean speech recognition task. He verified that the MFCC outperforms the LP methods and achieves the best performance. With its robust and cost-effective computation, MFCC has become a standard choice in the automatic speech recognition applications. Hermansky and Mporas et al. have experimentally verified that only under specific conditions can PLP features outperform MFCC [46] [78]. MFCCs have been selected in this work for the speech parametrization task based on these previous studies.

2.1.2 Decoding

The Equation 2.2 is the mathematic description of a speech recognition task. Given acoustic models $P(O|W)$ and language models $P(W)$, the task of speech recognition is to find a unit string with the maximum $P(W|O)$. The key challenge has become building accurate acoustic models and language models to meet the needs of real applications.

Acoustic Modelling with Hidden Markov Models (HMMs)

After feature extraction, we have a sequence of acoustic feature vectors. The probability of these acoustic feature vectors has to be estimated given acoustic models $P(O|W)$. Acoustic modelling converts acoustic features into probabilities of particular labels. Since the 1980's, HMM has become the most preferred method for ASR acoustic modelling. The intrinsic variability of the speech signal and the structure of spoken language can be modelled with stochastic processes in an integrated and consistent statistical modelling framework [94]. HMM is specified with an output alphabet, a set of states, initial state probability distribution, transition probability matrix and



(a) waveform

(b) Spectrum

(c) LPCC features

(d) MFCC features

(e) PLP features

Figure 2.6: Overview of audio features extracted from WSJ file 4p0a0101

emission probability distribution. Figure 2.7 shows an HMM-based phone model with transition probability parameters $\{a_{ij}\}$ and emission probabilities $\{b_j\}$.

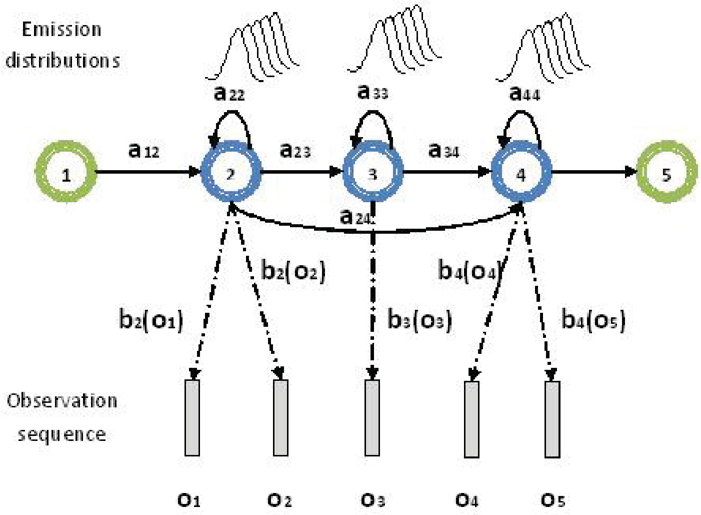


Figure 2.7: HMM with 3 emitting states

In the speech recognition task the linguistic structure is represented by a Markov chain. The variability in the acoustic realization of the sounds in speech is represented by a set of probability distributions. Figure 2.7 shows an example of a typical 5 states left-to-right HMM for a speech recognition task. Observations o_1 to o_6 can be generated by a HMM M by moving through the state sequence $X = 1; 2; 2; 3; 4; 4; 5$. The joint probability $P(O, X|M)$ can then be expressed as:

$$P(O, X|M) = a_{12}b_2(o_1)a_{22}b_2(o_2)a_{23}b_3(o_3)a_{34}b_4(o_4)a_{44}b_4(o_5) \tag{2.11}$$

The HMM parameters can be estimated via training with the Baum-Welch algorithm [8]. Gaussian mixture densities are often used to represent the output distribution $b_i(o_j)$. By speech recognition task, only the observation sequence O is known and the underlying state sequence X is hidden. Therefore the required likelihood representing that the observation sequence O is generated by HMM M is estimated by summing up the probability of all possible state sequences $X = x(1), ...x(T)$, which is computed as follows:

$$P(O|M) = \sum_X a_{x(0),x_(1)} \prod_{t=1}^{T} b_{x(t)}(o_t) a_{x(t)x(t+1)} \tag{2.12}$$

where $x(0)$ is the model entry state and $x(T+1)$ represents the model exit state.

Language Modelling

The language model (LM) represents the knowledge of what constitutes a possible word, what words could co-occur, and in what sequence. The language model is a probability distribution capturing the statistics of natural language. N-gram language models that represent the probability of a sequence with N words are generally used in ASR task [53]. Given the word string $W = w_1, w_2, ...w_n$, the probability $P(W)$ is computed as follows:

$$P(W) = \prod_{i=1}^{n} P(w_i|w_{i-N+1}, ..., w_{i-1}) \tag{2.13}$$

Pronunciation Dictionary

A pronunciation dictionary maps recognizable units like words, syllables, etc. into a sequence of phones (or names representing basic acoustic models). The international phonetic alphabet (IPA) is the most common phone code. However, IPA contains some special letters that are not available on computers. Some other computer friendly phone codes have been proposed, such as ARPABET [104], TIMIT Alpahbet [125], etc.

Viterbi Decoding

Viterbi decoding [33] is performed on the recognition network constructed from a language model, a pronunciation dictionary and a set of HMMs. The recognition network consists of nodes (HMM model instance or word end) and arcs. As shown in the Figure 2.8, the word network is first transformed into a phone-level recognition network with the help of a pronunciation dictionary, and then extended to HMM state level. The nodes of the final recognition network are the HMM states connected by transitions.

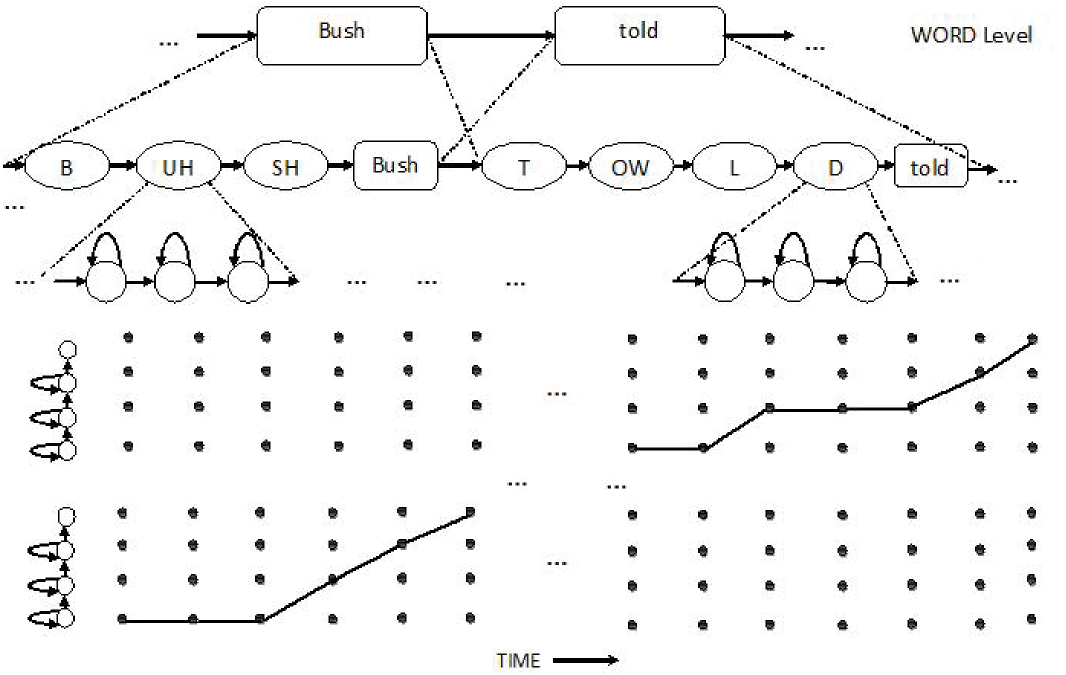


Figure 2.8: Recognition Network and HMM trellis for continuous speech recognition example.

The Viterbi algorithm produces a match between the acoustic observation and a path in the recognition network to find the optimal word sequence, which corresponds to the most suitable state sequence.

As mentioned before, the performance of a spoken document retrieval system depends heavily on the quality of the applied ASR-systems. Many research groups have already made technological improvements in all areas of ASR-system, for example, discriminative acoustic model training techniques for ASR [45]. Discriminative training takes the competing classes into account to optimize the parameters to be estimated. In contrast to discriminative training, the classic maximal likelihood (ML) criterion is optimized with the iterative expectation-maximization (EM) algorithm [30]. Heigold [45] provided a review about discriminative training techniques and presented the result of his experimental research about different discriminative training criteria, such as the maximum mutual information (MMI) [118], the minimum classification error (MCE) [69] and so on. Discriminative training is expected to outperform ML in case of imperfect model assumptions based on limited data [80]. Hinton [47] et al. confirmed with their experimental results that further improvement in ASR could be achieved by using the deep neural networks (DNNs) [122] instead of the Gaussian mixture models (GMMs) for acoustic modelling. More advanced and focused techniques in different areas of ASR, especially LVCSR, were summarized in [100]. The improvement on spoken information retrieval performance based on the technological improvements in ASR-system were reported in [25], [49] and [58]. How to improve the ASR-performance is a widely studied topics. However, it is out of the scope of this thesis.

2.2 Spoken Document Indexing

The aim of the spoken document indexing is to find important meanings and create an internal representation. The selection of accurate index units or terms plays an important role in information retrieval. The indexing term should be useful in distinguishing documents from one another. The following criteria need to be considered:

- The number of different indexing features must be small
- The collection frequency of the indexing term must be high enough

By the end of the indexing stage, every document will be represented by a set of weighted keywords or terms in the form $D_1 \rightarrow \{(t_1, w_1), (t_2, w_2), ...\}$. An inverted file composed of a vocabulary of terms and a list of occurrences will be simultaneously constructed.

Performing document indexing in advance can avoid linearly scanning documents in large collections for finding each query-word. One of the main issues in information retrieval is how to find the best index unit to precisely and exhaustively represent the content in the document. In current spoken document retrieval systems, automatic speech recognizers are usually used to transcribe the digital speech signal into a sequence of selected index units like phone, syllable, morphs and word.

Phoneme, Phone, Allphone and Phone N-Sequence

Phoneme denotes any of the minimal units of speech sound in a language that can serve to distinguish one word from another. The International Phonetic Alphabet (IPA) is normally used to represent phonemes consistently. A **phone** is the acoustic realization of a phoneme. For example the pronunciation of letter *d* in the English words *dog* and *drive* are different, as the position of the tongue is slightly different. Therefore the phoneme /*d*/ corresponds to at least two phones. The set of phones corresponding to the same phoneme is called **allphone**. Ng [81] verified that overlapping, fixed-length phonetic sequence (**Phone N-sequences**) could also be applied as an indexing unit. In Ng's work, phone N-sequences are extracted from the phonetic transcription provided by a phone recognizer.

Broad Phonetic Class Sequence

The idea of broad phonetic classes is to use more general phonetic classes for the spoken document indexing task [84]. Broad phonetic classes capture a lot of phonological constraints. Halberstadt [42] has verified that recognition errors often occur among phones within a same broad phonetic class. As shown in the Figure 2.9, the 41 phones in the TIMIT corpus [36] are hierarchically clustered. The clustering is based on acoustic measurements of the phones. This measuring of phone acoustics is based on

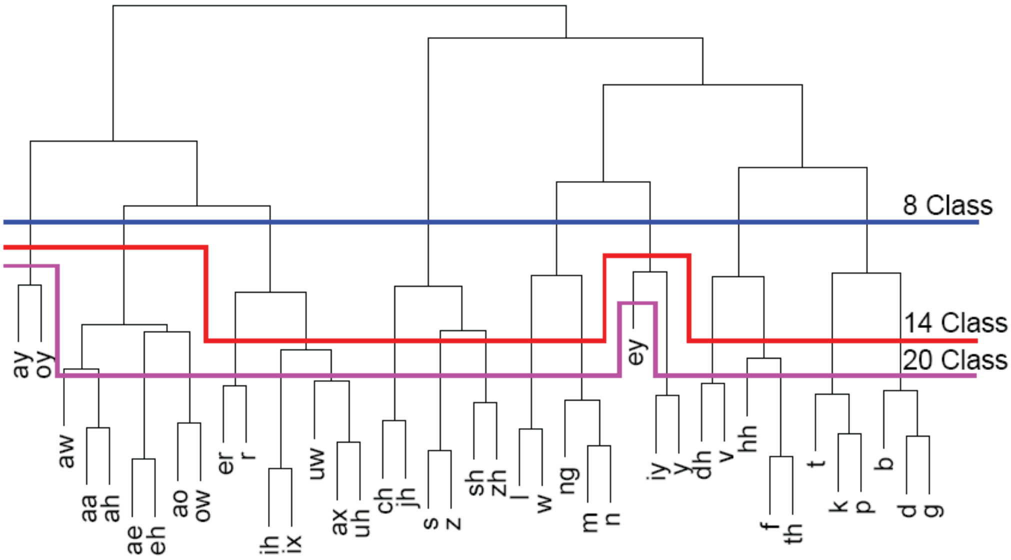


Figure 2.9: The hierarchical clustering tree (dendrogram) used to generate phonetic broad classes was presented in Ng's work [84].

61-dimensional feature vectors. A phone can be acoustically split into three phases. The 61 dimensional feature vector of a phone consists of:

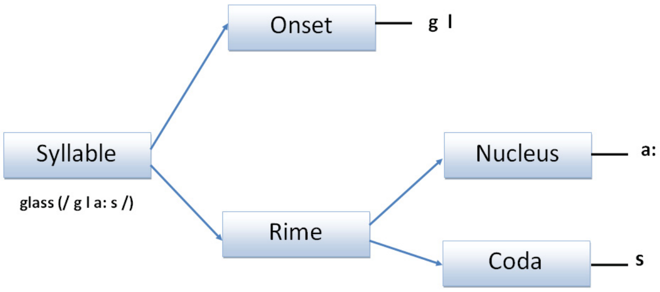
- 12 MFCCs and their derivations for the beginning and end phase
- 12 MFCCs for the middle phase

- The Log duration

The final feature vector used for the phone clustering task is estimated by averaging the feature vectors derived from all occurrences of a phone in the TIMIT training set. The Euclidean distance between acoustic feature vectors is used to measure the similarity between corresponding phone classes. In each clustering step, two phone classes with minimal Euclidean distance will be merged. This process will continue to aggregate groups together until there is just one big group. As shown in the dendrogram (Figure 2.9, with the cuts in the dendrogram, three sets which include a different number of broad phone classes can be derived.

Syllable

Syllables that center around vowels in English are intermediate units between the phone and words level. The internal structure of a syllable is shown in the Figure 2.10. Generally, a syllable consists of an *onset* (initial consonants if any) and *rime*. The *rime* consists of the *nucleus* (a vowel peak) and the *coda* (consonants following the *nucleus*). Alternatively, the syllable could be abstracted as the $vowel - consonant - vowel$ (VCV) feature. Most languages allow 'open' syllables like *coda*-less syllables with the structure V, CV, CCV. As proposed by Glavitsch and Schauble, syllable-like



;

Figure 2.10: General internal structure of a syllable

units (VCV-features) are derived from analyzing text documents for spoken document retrieval [39] [102]. The letters A, E, I, O, U and Y are counted as *vowels*. By building a syllable recognizer, words in the recognition vocabulary are split into a sequence of syllables (vowel-consonant-vowel features). When a vowel is placed in the middle of a word, a cut will be made in the middle of this vowel. Each half will be assigned to a feature. As an example, the word *President* can be split into a sequence of syllables: *pre*, *esi*, *ide* and *ent*. Because that the vowel-consonant-vowel features are extracted from the text, not all vowel-consonant-vowel features are suitable for building speech recognition. Two factors must be considered whenchoosing vowel-consonant-vowel indexing features:

- It should occur enough times to build a robust acoustic model (collection frequency)
- It should allow for discriminationbetween different messages

A threshold will be assigned to the collection frequency of the vowel-consonant-vowel features. However, a syllable-based spoken document retrieval approach may not be very effective, as the vowel-consonant-vowel features are directly extracted from text regardless of their acoustic characteristics. The syllable representation of a spoken document to be retrieved may include a lot of recognition errors, which would lead to a drop in retrieval performance.

Morpheme, Morph and Allomorphs

Another important kind of indexing unit is the *morpheme*. A *morpheme* is defined as the smallest linguistic unit that has semantic meaning. Some morphemes can stand alone as a word but many cannot. A *morph* is the phonetic representation of a morpheme. *Allomorphs* are a set of *morphs* corresponding to the same *morpheme*. For example, the English word 'unbelievable' consists of three morphemes 'un-', 'believ' and '-able'.

Morphemes could be directly extracted from a text collection. Under the assumption that words are formed by a concatenation of morphemes, a number of unsupervised methods have been developed for automatic morpheme analysis ([11], [21]). Automatic morpheme analysis can be achieved by word decomposition using generative models or by applying minimum description length (MDL) for the optimal text segmentation ([22], [29] and[40]).

The sequential segmentation method based on the MDL cost was proposed by Creutz [22]. This method could be considered as constructing a model of the data (text sequence). This model consists of a vocabulary of morphs (codebook). A concise set of morphs which gives a concise representation of a text is detected based on the MDL-cost C defined in Equation 2.15. Given a source text D that consists of n morph-tokens and a codebook that includes $N_{codebook}$ morphs, the MDL-cost C consists of the cost of the source text in this model and the cost of the codebook,

$$\begin{aligned} C &= Cost(source\_text) + Cost(Codebook) && (2.14) \\ &= \sum_{i=1}^{n} -\log p_D(m_i) + \sum_{j=1}^{N_{codebook}} k * l(m_j) && (2.15) \end{aligned}$$

where the cost of the source text is the negative log-likelihood of the morph token $p_D(m_i)$, summed over all the morph tokens that comprise the source text. $p_D(m_i)$ is estimated as:

$$p_D(m_i) = \frac{N_{m_i}}{N_{\text{source text}}} \tag{2.16}$$

where N_{m_i} denotes the number of morph tokens m_i in the source text and the total number of morph tokens in the source text is represented by $N_{\text{source text}}$.

The cost of the codebook $Cost(Codebook)$ is defined as the length in bits needed to represent each morph separately as a string of characters, summed over the morphs in the codebook. k indicates the number of bits needed to code a character. $l(m_j)$ is the length of the morpheme m_j in number of characters.

With this method, a word will be recursively decomposed. When a new word is read from the text, it will be considered as a morph and added to the codebook. Then every "cut" that segments the word into two parts is evaluated and the split with minimum cost is selected. The processing of the word finish once there are no possible splits left. Then, a new word will be read from the input text.

Morphs have been confirmed to be very useful in retrieving a highly inflectional language like Finnish [112]. Because the words of a highly inflectional language commonly consist of many inflections and compounds, lexical and language modelling becomes to be the main difficulty in using the standard large vocabulary word recognition technology. Huge corpora is required for training the models of sufficient coverage of the language. Even though, those models are infeasible to process in speech recognition. There are many errors hiding in the speech transcripts, so the performance of spoken document retrieval drops as well. In this case, the application of morph recognizer reduce the errors in speech transcripts and consequently benefits the performance of retrieving spoken documents. The work presented by Kurimo [59] pointed out that average morph-based document retrieval could achieve a precision comparable to the one obtained from human reference transcripts.

The performance of this kind of retrieving spoken documents depends heavily on the quality of morph-based speech transcripts that is in practice often filled with errors caused by non-ideal segmentation of inflected word forms. This problem is very similar to *understemming* and *overstemming* problems and may alleviated by the use of additional query expansion [60] or latent semantic indexing [111].

Word

A word is the single unit in the language. It could also be defined as a lexical item with an agreed-upon meaning in a given speech community. Words have been widely used for spoken document indexing tasks. However, we need to solve the problem of detection of word-boundaries and out-of-vocabulary words in spoken language.

2.3 Spoken Information Retrieval

Spoken content retrieval deals mainly with search, scoring and ranking of documents in a collection, with respect to the given queries. The general retrieval model consists of representations of documents D; a representation of a query Q; a modelling framework for D, Q and their relationships; and a ranking or similarity function $Sim(q, d_i)$, based on which retrieved documents will be ranked. In recent studies, classic text retrieval models like the boolean model, the vector space model, etc. are applied for information retrieval in spoken document word representation provided by an automatic speech recognition system.

2.3.1 Boolean Model

The boolean retrieval model looks for exact matches of query terms in documents. In order to avoid linearly scanning the texts for each query, the documents in the archive are indexed in advance. A binary term-document incidence matrix is then constructed. If there are four documents D in the archive as shown in Table 2.1, the corresponding binary term-document incidence matrix is shown in Table 2.2. The answer to a query

Doc_id	Terms in Doc
D_1	$t_1\ t_2\ t_3\ t_4$
D_2	$t_2\ t_4\ t_5\ t_6$
D_3	$t_1\ t_3\ t_5\ t_7$
D_4	$t_2\ t_3\ t_4\ t_7$

Table 2.1: Documents in the archive

	D_1	D_2	D_3	D_4
t_1	1	0	1	0
t_2	1	1	0	1
t_3	1	0	1	1
t_4	1	1	0	1
t_5	0	1	1	0
t_6	0	1	0	0
t_7	0	0	1	1

Table 2.2: Example of binary term-document incidence matrices.

consisting of terms t_1, t_3 and t_7 is formed as bitwise AND among document vectors for t_1, t_3 and t_7:

$$1010 \wedge 1011 \wedge 0011 = 0010 \tag{2.17}$$

The result indicates that document D_3 is most relevant to the given query. Two other operators OR and NOT could also be used to form queries. The main drawback is that this retrieval model is based on a binary decision criterion without any notion of grading. All terms are assumed to be equally important. This kind of exact match may lead to either too little or too many documents being retrieved. However, the boolean retrieval model with its clean formalism and simplicity is still one of the dominant retrieval models.

2.3.2 Vector Space Model

Vector space models [92] were developed to attack the problems of the boolean model, caused by binary decision and equal term weights. Non-binary weights will be assigned to index terms in the documents ($\overrightarrow{d_i} = (w_{i,1}, w_{i,2}, ..., w_{i,t})$) and in the query ($\overrightarrow{q} = (w_{q,1}, w_{q,2}, ..., w_{q,t})$), which enables partial matching between documents and query. Similarity between document and query will then be estimated based on term weights. Two main issues need to be discussed more in detail: term weighting [35] and scoring schemes (how to estimate the similarity score).

Term Weighting

Weight will be assigned to each term in the documents and the query. This weight indicates how important a term is with respect to a query or document. In order to enhance the retrieval effectiveness, higher weights should be given to important terms. We can imagine that if a query term t (nonstop words like: a, and, any, etc.) occurs many times in document d, then document d should be relevant to the given query. The recall rate described in the Section 3.4 of an information retrieval system relies on terms with high occurrence frequencies. In this work, tf denotes the term frequency. It is one of term frequency factor in SMART notation and is labelled with n as showed in Table 2.3. Assigning a term weight directly to its **term frequency** is the most straightforward way to integrate the number of term occurrences into the ranking score. This weighting method has been integrated in the vector space model since the 1960s.

However, a weight based on the term frequency alone cannot ensure acceptable retrieval performance. High frequency terms may be prevalent in the whole archive. High value should be given to terms concentrated in a few documents of a collection. A **collection-dependent factor** must also be considered. One of the most famous collection-dependent factors is the **inverse document frequency (idf)**. It is one of document frequency factor in SMART notation and is labelled with t in Table 2.3. As defined in Equation 2.18, high frequency terms will get low idf value while the idf of a rare term is high,

$$idf_t = log\frac{N}{df_t} \tag{2.18}$$

where N indicates the number of documents in the collection and df_t is the number of the documents in collection that contain term t (**document frequency**). The idf factor is used to scale down the weight of the terms that occur too often in the collection. The useful terms should be those which occur frequently and which are prevalent in a restricted number of documents as well. They should be able to be used to distinguish certain documents from the rest of the archive. The $Tf - idf$ **term**

weighting scheme defined in Equation (2.19) combines term frequency and *idf* into one composite weight which will be assigned to terms in the documents.

$$W_{tfidf_{t,d}} = tf_{t,d} \cdot idf_t. \tag{2.19}$$

Documents in a collection have different lengths. Longer documents tend to be represented by large term vectors. Hence, longer documents have a better chance of being retrieved than short ones. Therefore, in case of retrieval from a collection consisting of documents with widely varying length, a **normalization factor** should be incorporated into the term weight to equalize the influence caused by varying document lengths. Table 2.3 [103] presents the important factors used to form the term weight in the SMART notation.

The combination of term weighting is denoted within SMART-Notation *qqq.ddd*. The first triple represents the term-weight combination of the query terms and second triple represents the term-weight combination of the document terms. For example, *ntc.atn* means that the query term will be weighted as:

$$\frac{tf_{q,t}.log\frac{N}{df_t}}{\sqrt{\sum_{i \in q}(tf_{q,i} \cdot log\frac{N}{df_i})^2}} \tag{2.20}$$

and the weight of the terms in document will be formed as:

$$(0.5 + \frac{0.5tf_{d,t}}{\max_{j \in d}(tf_{d,j})}) \cdot log\frac{N}{df_t} \tag{2.21}$$

Scoring Scheme

The similarity between document and query vectors can be estimated as the normalized inner dot product between document and query vectors (cosine similarity function) [99]. If $\vec{d}$ represents the document vector and $\vec{q}$ denotes the query vector, the similarity of document d to query q can be expressed as:

$$Sim(\vec{d}, \vec{q}) = \frac{\vec{q} \cdot \vec{d}}{|\vec{q}||\vec{d}|} = \sum_{t \in q} \frac{q[t]}{|\vec{q}|}\frac{d[t]}{|\vec{d}|} \tag{2.22}$$

Term Frequency Factor

Symbol	Value	Description
b	$(0, 1)$	binary weight factor [35]. It equals to 1 when the term is included in a vector. Otherwise the value will be 0
n	$tf_{t,d}$	natural term frequency (tf) factor indicates how many times a term occurs in a document or query text
a	$0.5 + \frac{0.5 \times tf_{t,d}}{max_t(tf_{t,d})}$	natural term frequency factor normalized by maximum tf in the vector (augmented normalized term frequency)
l	$1 + log(tf_{t,d})$	log-weighted term frequency
L	$\frac{1+log(tf_{t,d})}{1+log(ave_{t \in d(tf_{t,d})})}$	average log-weighted term frequency

Document Frequency Factor

Symbol	Value	Description
n	1	it means there is no change in weight
t	$log\frac{N}{df_t}$	inverse document frequency (*idf*)
p	$max\{0, log\frac{N-df_t}{df_t}\}$	probabilistic *idf*

Normalization Factor

Symbol	Value	Description
n	1	without any normalization
c	$\frac{1}{\sqrt{w_1^2+w_2^2+...+w_m^2}}$	cosine normalization, w_m indicates the term weight estimated using term frequency and/or document frequency factor
u	$\frac{1}{u}$	pivoted normalized document length, which is used to compensate the impact of widely varying document length [106]
b	$\frac{1}{CharLength^{\alpha}}, \alpha < 1$	normalized by document length in bytes

Table 2.3: SMART notation for the term-weighting scheme [103]

where the denominator is the product of the Euclidean lengths of document and query vector.

The vector space model (VSM) method enables the terms to be weighted with their importance. Finally, the documents are ranked according to their similarity to the given query.

Chapter 3

Experimental Setup

This chapter introduces our experimental setup. Section 3.1 describes the databases selected for the evaluation task. Section 3.2 introduces our 20k word recognizer and its recognition performance on different data sets. The description of our retrieval task can be found in the Section 3.3. Section 3.4 introduces the different evaluation measures applied in this thesis. Our baseline SDR system is described in section 3.5.

3.1 Databases

Reuters Corpus is a large English text Corpus. It consists of 806791 English broadcasts (the annual output of Reuters) covering the period from $1996 - 08 - 20$ to $1997 - 08 - 19$. The news stories are saved as xml files in the newsML format. This corpus was originally designed to develop topic identification systems. It is now mainly used in the development of automatic text mining technologies (clustering, categorizing, topic detection, etc.). In this thesis, word statistics are gathered from the first half of the Reuters Corpus (CD1). This part consists of 473876 documents and a total of 97 million words, with an average of 205 words per document. The vocabulary consists of 311706 different words.

TIMIT acoustic-phonetic Continuous Speech Corpus [36] is a corpus of read speech. It is designed for acoustic-phonetic studies and for the development and evaluation of automatic speech recognition systems. The TIMIT corpus includes broadband recordings of 630 speakers of 8 major dialects (New England, Northern, North Midland, South Midland, Southern, New York City, Western and so-called Army Brat) of American English. Each speaker reads 10 phonetically rich sentences. The speech is saved in mono 16 bit, 16kHz waveform. The time-aligned orthographic, phonetic and word transcriptions are also available. This Corpus is used to initialize the parameters of the acoustic model.

WSJ [90], the Wall street Journal (WSJ) CSR Corpus, is a general-purpose English corpus with large vocabulary, natural language, and a high complexity. It contains a significant quantity of speech data (more than 400 hours) and text data (ca. 47 million words) and is widely used in speech processing technology. The WSJ Corpus consists of three main parts: the training part for the estimation of the acoustic models' parameters a development set for optimizing the recognition decision criterion and speech data for performance evaluation. The 20k word recognizer built for this thesis is trained on WSJ data.

The **spoken document test collection** consists of data sets selected from WSJ-Corpus (si_dt_s2, si_dt_s5, si_dt_s8 and si_dt_jd). It contains 8158 spoken sentences. The information about every selected WSJ data set is listed in the Table 3.1: The vocabulary of our test collection contains 17.52% OOV words. The occurrence of OOV words corresponds to 3.9% of the total. The shortest document consists of 4 words. The maximum length of a document is 59 words. There are on average 18 words per document.

Set	**Data**	**Evaluation task**
si-dt-s2	10 speaker * 20 Utts from ATIS domain; 10 speakers * 0utts. from Mercury	domain-independence
si-dt-s5	200 utts spoken by 10 speaker	Microphone-Independence
si-dt-s8	10 speakers * 10 utts * 2 soures * 3 levels	noisy environment
si-dt-jd	10 speakers * 20 utts = 200utts	spontaneous dictation
total	8158 files	10901 different words with OOV rate about 3.9%

Table 3.1: Test data information

3.2 20k Word Recognizer

This section introduces the 20k word recognizer applied in this thesis. This word recognizer is trained on 16kHz mono audio speech, the training part from the WSJ corpus. Acoustic models were initialized with the TIMIT corpus [32] and trained on entire WSJ training sets. The bigram language model was trained on the NoV-92 LM training data (20k vocabulary). This recognizer achieves an accuracy of 95.21% on clean read speech; the Nov92 test-set ($5k$ closed-set test data) in the WSJ corpus. Further evaluations have been made on test datasets selected for our SDR task. The experiment results are presented in Table 3.2. The performance of our 20k word recognizer on different data sets is shown in Figure 3.1. In this figure, mu_0,mu_10 and mu_20 denote

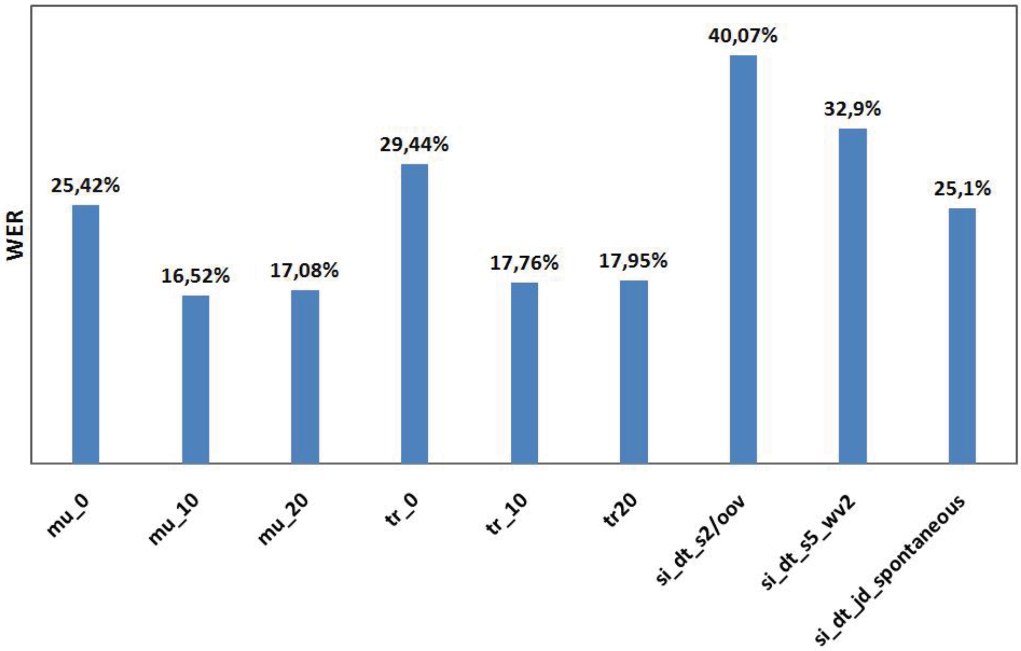


Figure 3.1: Evaluation 20k word recognizer on different datasets.

speech with music in the background with SNR values of 0, 10, and 20dB, respectively tr_0, tr_10 and tr_20 are speech data with a talking radio in the background with SNR values of 0, 10 and 20dB; si_dt_s2 is clean read speech with 7 % OOV words; $si\_dt\_s5\_wv2$ consists of records with the microphone away from the mouth of the speakers, and si_dt_jd is spontaneous dictation. The following observations can be made:

- With a talking radio in the background, more errors occur ($WER = 29.44$ % at an SNR-level of 0 dB) than musical background noise ($WER = 25.42$ % at an SNR-level of 0 dB);
- In comparison with an SNR-level of 0 dB, a clear improvement could be viewed at an SNR-level of $10dB$;

Data	#wrds	OOV	#corrs	#Del	#Subs	#Ins	WER	CORR	Acc
mu_20	3466	0.17%	2886	246	334	12	17.08%	83.30%	82.90%
mu_10	3002	0.27%	2512	182	308	6	16.52%	83.70%	83.50%
mu_0	3352	0.66%	2544	262	546	44	25.42%	75.90%	74.60%
tr_20	3442	0.29%	2840	268	334	16	17.95%	82.50%	82.10%
tr_10	3614	0.17%	2996	230	388	24	17.76%	82.90%	82.20%
tr_0	2982	0.25%	2222	236	524	118	29.44%	74.50%	70.60%
si_dt_s5	3345	0.33%	2283	362	700	40	32.90%	68.30%	67.10%
si_dt_s2	3454	7.06%	2149	381	924	79	40.07%	62.22%	59.93%
si_dt_jd	113451	2.04%	88056	7124	18271	3033	25.10%	77.60%	74.90%

Table 3.2: Performance on different data sets.

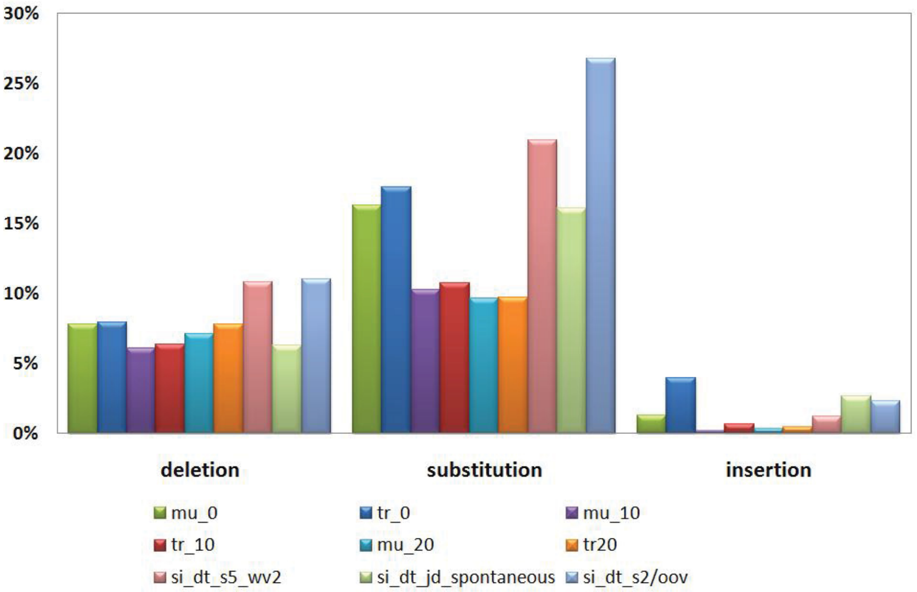


Figure 3.2: Effect of different environmental factors on the distribution of errors in the transcription.

- The worst performance ($WER = 40.07$ %) can be observed by clean read speech with 7 % OOV-words (si_dt_s2);

- The impact of varying speech style ($WER = 25.1$ %, si_dt_jd) and channels ($WER = 32.9$ %, si_dt_s5_wv2) should also be observed.

Three kinds of errors can be observed in the transcription: deletion, insertion and substitution errors. The influence of different environment factors on the error distribution in the transcription is shown in Figure 3.2. It is clear that strong background noise (music or talking radio) may bring more substitution and insertion errors. Most deletion errors are caused by the out-of-vocabulary words (si_dt_s2, with 11.03 %) and channel variation (si_dt_s5_wv2, with 10.82 %). The out-of-vocabulary words will be replaced with acoustically similar words from the vocabulary. In this case, the percentage of the substitution errors is high (26.75 %). A loud talking radio in the background (at an SNR-level of 0 dB) will cause more insertion errors, because spoken words from the background radio are added to the transcription. An average word error rate of 25 % can be achieved on the test collection.

3.3 Single-word Query Retrieval Tasks

A set of textual single-word queries was selected for the evaluation task. The system aims at returning a list of spoken documents, that contain the query formed by a user. This retrieval task is different from conventionally keyword-spotting referring to a predefined keyword set. The keyword-spotting is a widely studied topic and has achieved currently significant advances by hierarchical keyword modelling approach [17], by using a long short-term memory recurrent neural network-based feature extractor [16] and many other technologies ([5], [121], etc.). However, keyword spotting is out of the scope of this thesis.

Our query set consists of 200 in-vocabulary words, as shown in appendix A and the 30 out-of-vocabulary words (Table3.3) with the highest occurrence in the test collection. The distribution of in-vocabulary query length over the number of phones is shown in Figure 3.3. The average query length is 6 phones.

Word	Occ	Word	Occ	Word	Occ	Word	Occ	Word	Occ
CLINTON'S	35	ARTIFACTS	21	MUTATION	21	PHILOSOPHERS	21	BISMARCK	20
CONTINENTS	20	ECLIPSES	20	EXPECTANCY	20	GOATS	20	HERDING	20
HUMANKIND	20	INORGANIC	20	INTERVALS	20	MECHANIZED	20	MICROBIOLOGY	20
OCEANS	20	RADIATOR	20	SHAFT	20	CORP.	12	KORESH	11
RODHAM	11	WORKFORCE	11	HILLARY	9	SPOKESPERSON	9	SPILLER	8
POLYSTYRENE	7	BIKING	6	CANINE	6	CHAVIS	6	IMPLANTS	6

Table 3.3: 30 OOV queries with occurrences.

Logan has reported that in real applications, about 13 % of the query words are out-of-vocabulary words [67]. We simulate the real application case and keep an out-of-vocabulary word rate of 12.7 % in the query set.

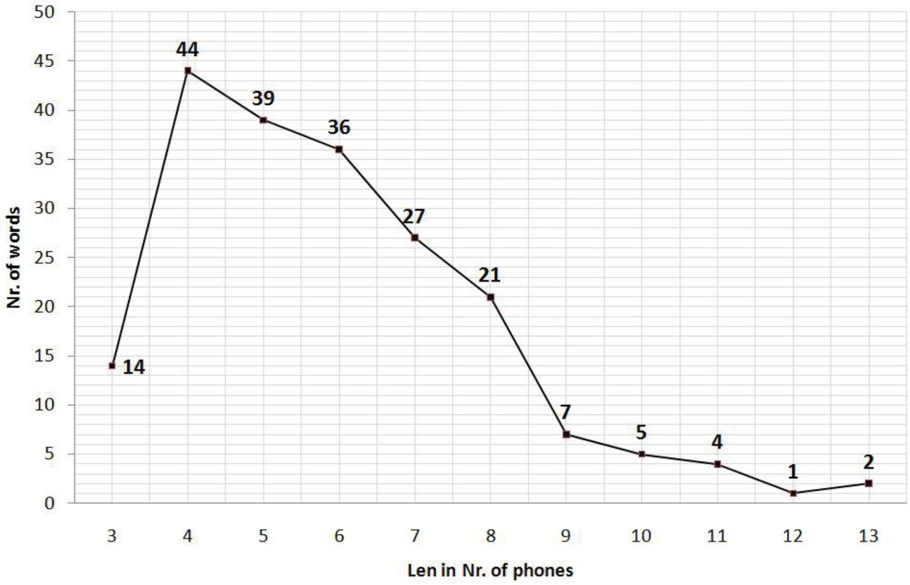


Figure 3.3: Length distribution of 200 in-vocabulary queries.

3.4 Evaluation Strategies

The performance of an automatic speech recognition system can be judged objectively. The recognizer accuracy is obtained by comparing the string of words output by the recognizer and the reference string of words. The performance evaluation of an information retrieval system depends on some human factors. The user who forms the query will assess the relevance of documents retrieved. Four evaluation measures are applied in this thesis: E1-E5 evaluation, precision/recall lot, mean average precision value (mAP) and accuracy.

E1-E5 Evaluation

The E1-E5 metric was proposed by Choi [18]. It has been widely used in retrieval performance evaluation on restricted test data.

- $E1$ denotes the number of queries for which the relevant document is at first place
- $E2$ represent the number of queries for which relevant documents are within the top 10 documents
- $E3$ indicates the mean answer rank

- $E4$ denotes the mean answer rank after removing the outliers. The mean answer rank measure drops noticeably when one of the queries with their first answer at a poor rank. Therefore, it is allowed to ignore one or two of worst queries for which the first answer is retrieved at a very poor rank when computing the mean answer rank.
- $E5$ represents the mean reciprocal rank

A document is relevant to a given query means this document is exact the one that the user is looking for. For one query there could be more than one relevant document in a Corpus to be retrieved.

Precision/Recall

Precision and recall are the most popular measures for the evaluation of information retrieval effectiveness. Precision represents the number of relevant documents retrieved over the total number of documents retrieved. It is defined as:

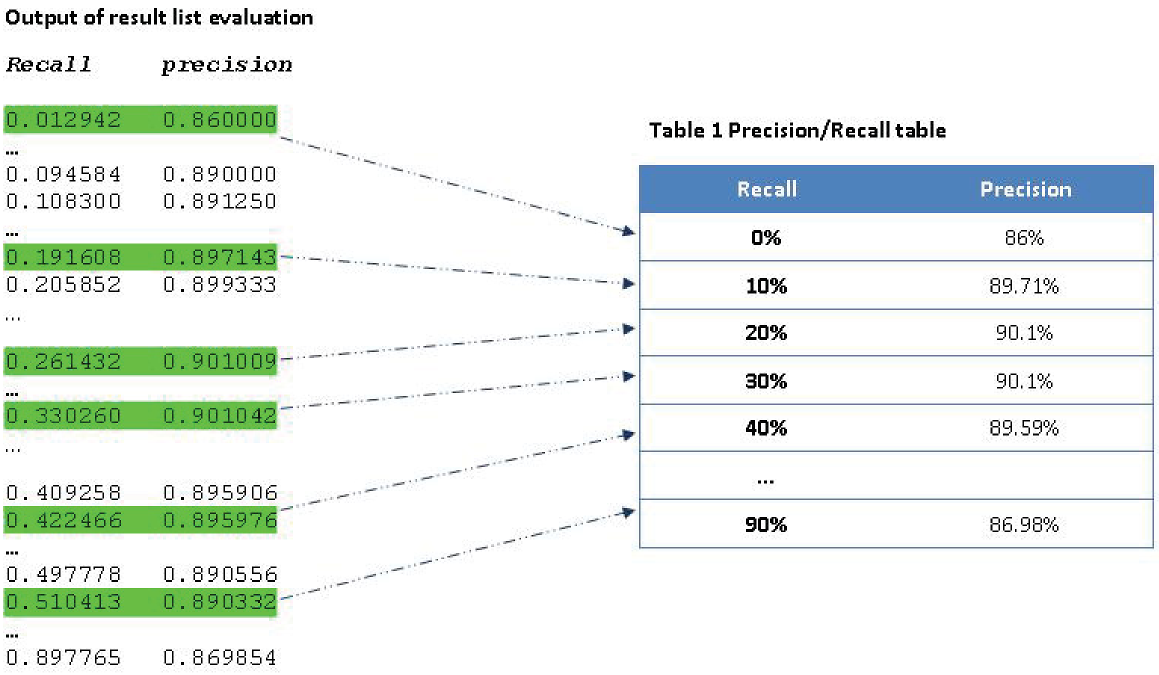
$$precision = \frac{R_{retrieved}}{N_{retrieved}} \tag{3.1}$$

Recall is the number of relevant documents retrieved over the total number of relevant documents in the collection. It is defined as:

$$recall = \frac{R_{retrieved}}{R_{collection}}; \tag{3.2}$$

The precision/recall curve is gathered by plotting precision against recall after each new document is added to the list of retrieved documents. Precision values at 11 standard recall levels $(0.0, 0.1, ...1.0)$ are interpolated using the modified plot normalization method in TREC evaluation [110]. The interpolated precision at standard recall level i is the maximum precision obtained for the given query for any actual recall level greater than or equal to i but less than $i + 1$. In this thesis, the precision at a recall of 0.0 is set to the precision value achieved at the nearest recall level. The Figure 3.4 presents an example for the precision/recall calculation.

This measure depends on the number of documents. Table 3.4 gives the definitions of quantities for the calculation of precision and recall.



Output of result list evaluation

```
Recall      precision

0.012942    0.860000
...
0.094584    0.890000
0.108300    0.891250
...
0.191608    0.897143
0.205852    0.899333
...

0.261432    0.901009
...
0.330260    0.901042
...

0.409258    0.895906
0.422466    0.895976
...
0.497778    0.890556
0.510413    0.890332
...
0.897765    0.869854
```

Table 1 Precision/Recall table

Recall	Precision
0%	86%
10%	89.71%
20%	90.1%
30%	90.1%
40%	89.59%
...	
90%	86.98%

Figure 3.4: Example of precision/recall value selection.

	relevant doc.	non-relevant doc.
retrieved doc.	true positives (tp)	false positives (fp)
not retrieved doc.	false negatives (fn)	true negatives (tn)

Table 3.4: Definition of tp,fp,fn,tn

The definition of precision and recall can be expressed as:

$$precision = \frac{tp}{tp + fp} \tag{3.3}$$

$$recall = \frac{tp}{tp + fn} \tag{3.4}$$

Mean Average Precision (mAP)

Sometimes it is very hard to make a decision between different retrieval systems just using precision-recall curves. In this case, a single value called mean average precision (mAP) can be used to compare the performance of different retrieval systems [43]. The mAP is computed by averaging the precision values at recall points for each query and then averaging them over all queries.

Accuracy

The accuracy of an information retrieval system is defined as:

$$Accuracy = \frac{tp + tn}{tp + fp + fn + tn} \tag{3.5}$$

The definition of tp, tn, fp and fn is as given in Table 3.4. This measure indicates the fraction of true classifications. This measure is not fit for the evaluation of a large collection information retrieval task, as 99% of documents in collection are usually in the non-relevant category.

3.5 Baseline SDR System

Our baseline SDR system is shown in Figure 3.5. It consists of three main components: word recognizer, indexing and vector space model based retrieving.

First, the word recognizer transforms digital spoken information into a sequence of words. The indexing module removes all stop words from the spoken document transcription provided by a word recognizer. The word recognition and indexing will be performed in advance. An index database is generated parallel to the test speech collection. When a query is formed by a user, the vector space model based method will search the index database for all splits that are relevant to the given query word. A ranked list of relevant documents will be returned to the user.

recall	0	10	20	30	40	50	60	70	80	90	mAP	max. RE
BL-INV	86.00	89.12	89.93	90.12	89.59	89.05	88.48	88.10	87.30	86.98	84.17	89.77
BL-Total	74.78	77.90	78.08	78.18	77.70	77.08	76.66	75.89	-	-	61.67	78.06

Table 3.5: Precision/recall of the baseline SDR system (in %).

This baseline system (BL) is evaluated on our test collection, including 8158 files. The experimental results are shown in Figure 3.6 and in Table 3.5. The retrieval performance for in-vocabulary queries $BL - INV$ yields a mean average precision of

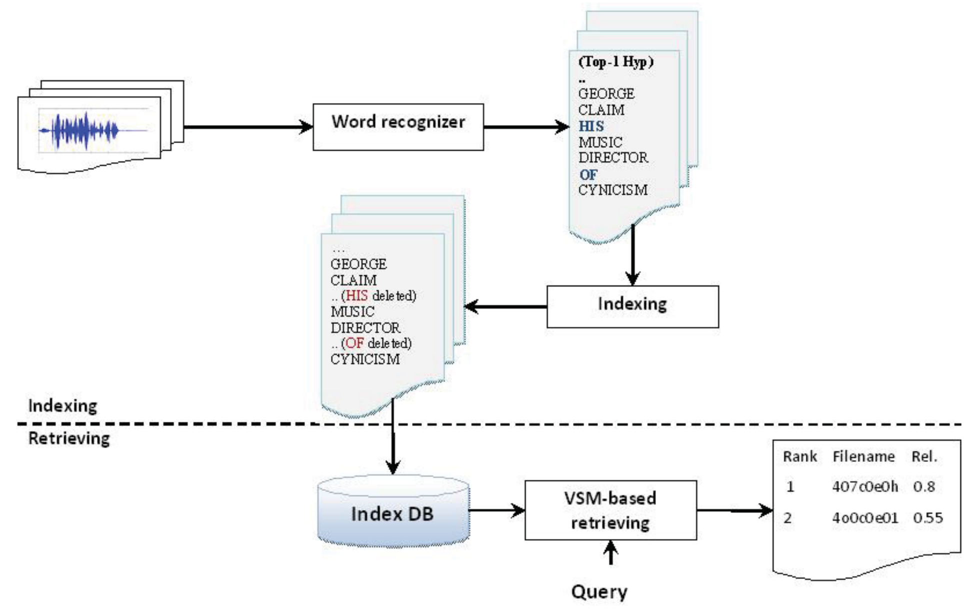


Figure 3.5: Structure of the baseline SDR system

84.17 % as shown in Table3.5. One can observe that the mean average precision drops (mAP = 61.67 %) when retrieving both in-vocabulary and OOV queries ($BL-Total$). The maximal recall rate is also degraded by about 11.7 %. The reason is that the out-of-vocabulary word will be recognized as the most acoustically similar in-vocabulary word. In speech transcript provided by a word recognizer, all out-of-vocabulary words are missing. The information bearing in the out-of-vocabulary words is lost.

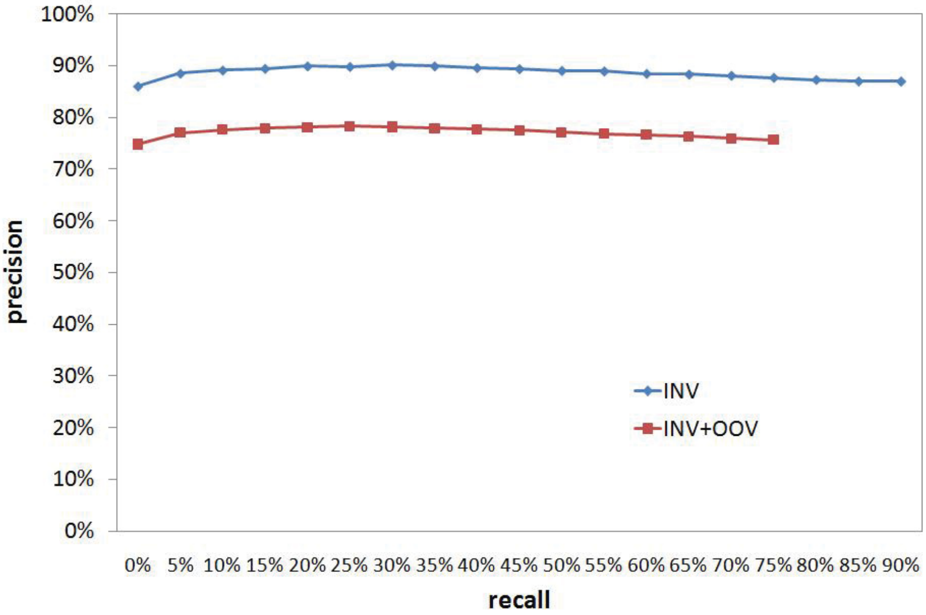


Figure 3.6: Precision/recall Plot of baseline SDR system.

Chapter 4

Enrichment of a Spoken Document Representation with Multiple Recognition Hypotheses

As mentioned before, the task of a spoken document retrieval system is to find all splits relevant to the given queries. A traditional spoken document retrieval system uses a word recognizer for automatic transformation of speech into text. Then the classic information retrieval methods can be applied for spoken information retrieval. There are some particular issues in the field of spoken information retrieval that need to be addressed:

- How to extract and represent the spoken content accurately in a form that can be stored and searched efficiently
- How to deal with the uncertainty in the transcription, or more precisely the recognition errors

This chapter focuses on the second main issue of spoken document retrieval. Recognition errors in the transcription are mainly caused by:

- Mismatch between training and real application scenario, such as varying background noise, speech variability, speech type, speaker change, application domain and more.

- The out-of-vocabulary words which do not appear in the vocabulary of the speech recognizer.

The misrecognition problem leads directly to the term mismatch problem. Different terms could be recognized for the same spoken content. Adding multiple ASR hypotheses to the transcription can effectively reduce the impact caused by misrecognition. This chapter will focus on the exploration of the spoken document retrieval approaches based on the transcription including multiple recognition hypotheses. N-best list and word lattice are two common representations of multiple recognition hypotheses. We are going to describe N-best transcription based SDR methods in Section 4.1. Section 4.2 discusses the word lattice based spoken document retrieval issue. The word confusion network as the most compact form of the lattice will be investigated in Section 4.3. In Section 4.4 we mainly focus on the exploration of SDR approaches that deal with the out-of-vocabulary problem by the way of representation expansion.

4.1 N-best Recognition Hypotheses based Spoken Document Retrieval

In this case, the representation of a spoken document prepared for the information retrieval includes N-best recognition hypotheses. The Table 4.1 shows the representation of a spoken document ($4p0a0101$) from the WSJ Corpus which includes 10 best recognition hypotheses. It can be seen that words 'The', 'young', 'in', 'November' and 'December' appear in every recognition hypothesis listed in the document representation. Therefore, the term frequency of those terms in the word-10best representation is 10. Those words are dominant terms in retrieval because of their high term frequency in the document representation. In addition to the term frequency, there are some other factors which are very useful for retrieving information from the document word-N-best representation, e.g. the rank of the hypothesis containing the given query term and the occurrence of the query term (content words) in the entire language. This extra information should be considered by forming document term weights. We are going to discuss two word-N-best representation specified document term weighting methods in the following sections.

Siegler [105] researched a set of techniques for the estimation of term presence and counts from the hypotheses contained in N-Best lists. The term presence was a Boolean

Rank	Hypothesis
1	The female produces only to two for a young in November and December
2	The female produces only to pay for a young in November and December
3	The female produces relative to four young in November and December
4	The female produces only to two for young in November and December
5	The female produces only to pay for young in November and December
6	The female produces the food for a young in November and December
7	The female produces only to two four young in November and December
8	The femail produces only to two four young in November and December
9	The female produces only to food for a young in November and December
10	The female reduces the food for a young in November and December
Ref	The female produces a litter of two or four in November and December

Table 4.1: N-best Lists of WSJ sentence 4p0a0101

value, that indicated the occurrence of a term in a document. The term counts were the number of occurrences of query terms in a document which are used to estimate the relevance/similarity score of a document given query. His experiments showed the improvements in retrieval precision and recall using the n-Best document source as opposed to the 1-best source.

4.1.1 Integrating the Rank of a Hypothesis into Document Term Weighting Scheme

The classical term-frequency $tfidf$ weighting scheme (introduced in Section 2.3.2) will be modified so that it can deal with the word-N-best representation of a spoken document. Information on the rank of the hypothesis in the word-N-best representation that contains the query term is integrated into the term weight. The query term is weighted as follows:

$$w_{tfidf}(t,q) = tf \cdot idf_{t,q} = \underbrace{\frac{tf_{query}}{|q|}}_{\text{tf}} \cdot \underbrace{\log \frac{|D|}{|\{d : t \in d^{(H)}\}|}}_{\text{idf}} \tag{4.1}$$

where tf_{query} denotes the occurrences of term t_q in query; $|q|$ indicates the size of query in account of terms; $|\{d : t \in d^{(H)}\}|$ are the number of documents in the collection including term t_q; $|D|$ indicates the number of documents in collection D. According to the SMART notation listed in Table 2.3, the weighing scheme of the query term is labeled as nt.

As reported by Siegler [105], the terms appearing in a hypothesis which is in the front rank of N-best hypotheses should be assigned with higher weight. We use $n_{t,d}^{(H)}$ to represent the average occurrence of the considered term t in the document d and it is computed as follows:

$$n_{t,d}^{(H)} = \frac{1}{H} \sum_{h=1}^{H} (n_{t,d}^{(h)}) \tag{4.2}$$

where $n_{t,d}^{(h)}$ represents the number of occurrences of the considered term t in the hth hypothesis of the document d.

The modified $(tf - idf)^{(H)}$ weight is correspondingly re-estimated as:

$$W(t,d) = (tf \cdot idf)_{t,d}^{(H)} = \frac{n_{t,d}^{(H)}}{\sum_i n_{i,d}^{(H)}} \cdot log \frac{|D|}{|\{d : t \in d^{(H)}\}|} \tag{4.3}$$

where $\sum_i n_{i,d}^{(H)}$ denotes the total average number of occurrences of all terms in document d; $|D|$ is the number of documents in the collection and $|d : t \in d^{(H)}|$ denotes the number of documents including the considered term t in one of the N-best hypotheses. We define a new symbol h in addition to the SMART notation, to label the document weighting scheme as h.

4.1.2 Document Term Weighting with General Probability

In recent studies, the hidden Markov Model is applied for the information retrieval task ([74], [73], [9] [107]). Given queries q, finding relevant documents in the collection will be achieved in a probabilistic manner.

The main issue of probabilistic information retrieval is how to estimate the probability that indicates a document is relevant to the given query. Different strategies are suggested for this issue. The most popular variation is proposed by Song ([107]).

This method combines the probability of the query term occurring in the document and its general probability for the document probability estimation. Here, the general probability of a term/word means the occurrence of a term in the natural language. In this work, we weight the query terms with binary information.

$$w(t,q) = \begin{cases} 1, & \text{if } t \in q \\ 0, & \text{otherwise.} \end{cases} \tag{4.4}$$

Let $P(t|d)$ be the probability that query term t is in document d and $P(t|G)$ be the general probability that term t appears in the natural language queries. Stop-words are not considered here. The terms in the document will be weighted as:

$$W(t,d) = \alpha P(t|d) + \beta P(t|G) \tag{4.5}$$

where the probability $P(t|d)$ is estimated as:

$$P(t|d) = \frac{n_{t,d}^{(H)}}{|d|^{(H)}}; \text{ where } |d|^{(H)} = \frac{1}{|H|} \sum_{h \in H} |h|. \tag{4.6}$$

As defined in the Equation 4.2, $n_{t,d}^{(H)}$ is the average number of occurrences of considered term t in the word-N-best representation of a spoken document d; $|d|^{(H)}$ indicates the average length of the document d in number of terms. The general language model term $P(t|G)$ is defined as:

$$P(t|G) = \frac{\sum_i \text{number of times } t \text{ appears in } d_i}{\sum_i \text{length of } d_i} \tag{4.7}$$

In this work, $P(t|G)$ is estimated using the Reuters Text Corpus introduced in 3.1. We use symbol p to represent this weighting scheme.

4.1.3 Experiment and Discussion

This section summarizes the experimental results of word-N-best based spoken document retrieval approaches. 200 in-vocabulary queries are selected for the evaluation task. We first explore the impact of the number of considered hypotheses on the retrieval performance. Then we compare word-N-best based spoken document retrieval methods with the different weighting schemes introduced in Section 4.1.1 and 4.1.2, with the baseline system which is based on the word-1best representation of a spoken document.

The Impact of the Number of Considered Hypotheses on the Retrieval Performance

This section focuses on the experimental exploration of the effect of the number of considered hypotheses on the retrieval performance. In order to increase the chance of capturing the correct hypothesis, the word-N-best represent of a spoken document should contain all reliable recognition hypotheses. However, there is a potential danger of using too many recognition hypotheses. Erroneous terms from low scoring hypotheses could be included in the document representation, which will result in the decrease of retrieval precision.

Document terms are weighted with modified term-frequency defined by the Equation 4.2 and 4.3. The results of the in-vocabulary queries retrieving experiments are presented in Figure 4.1 and Table 4.2.

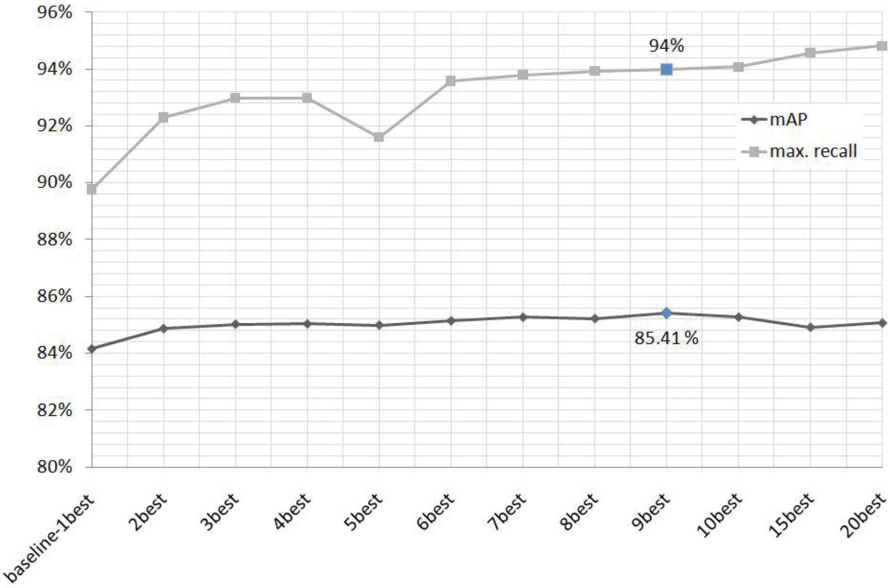


Figure 4.1: Retrieval performance in mAP and max. recall value for different N values.

Figure 4.1 shows the variation of mean average precision and recall of a spoken document retrieval system as the number of considered hypotheses is increased. It can be seen that the recall improves with growing number of considered recognition hypotheses. Compared with the word-1best based baseline system, word-20best spoken document retrieval approach yields better recall (5.05 %). The mean average precision (mAP) rises with a growing number of considered recognition hypotheses and reaches the maximum value (85.41 %) at $N = 9$. The mean average precision drops after $N = 9$. This result indicates that adding more than 9best recognition hypotheses to the document representation will not improve the retrieval performance.

recall	baseline	2best	3best	4best	5best	6best	7best	8best	9best	10best	15best	20best
0	86.00	89.70	89.70	89.70	88.50	90.40	91.09	91.09	91.70	92.40	91.70	91.70
10	89.12	90.70	91.60	91.60	91.80	91.90	92.10	92.30	92.40	92.40	92.70	93.00
20	89.93	90.90	91.80	91.80	91.70	92.00	92.30	92.10	92.20	92.30	92.40	92.60
30	90.12	91.10	91.60	91.60	92.20	92.20	92.27	92.30	92.40	92.50	92.40	92.60
40	89.59	90.10	90.90	90.90	91.40	91.50	91.60	91.80	91.97	92.00	91.90	92.10
50	89.05	89.70	90.30	90.30	90.70	90.70	90.90	91.10	91.10	91.00	91.30	91.40
60	88.48	89.10	89.40	89.50	90.00	89.80	90.00	90.00	90.00	90.00	90.10	90.20
70	88.10	88.20	87.70	87.70	88.00	87.60	87.50	87.40	89.00	87.30	87.20	87.10
80	87.30	86.70	85.40	85.40	84.90	84.80	84.50	84.20	84.00	83.80	80.00	82.70
90	86.98	85.90	84.20	84.20	83.20	83.30	82.70	82.30	81.90	81.57	80.40	79.60
mAP	84.17	84.87	85.03	85.04	84.99	85.14	85.27	85.22	85.41	85.27	84.92	85.07
max.RE	89.77	92.30	92.98	92.98	91.60	93.58	93.79	93.93	94.00	94.08	94.57	94.82

Table 4.2: Retrieval Performance (precision, recall, maximal recall and mAP value in % for 200 in-vocabulary queries for different numbers of considered recognition hypotheses.

Together with recall and mean average precision, the precision/recall values are listed in Table 4.2. There is a general trend of improved precision at recall levels from 1 to 60 %, with an increasing number of considered recognition hypotheses. The precision drops at higher recall levels ($recall > 60$ %) when the number of considered recognition hypotheses increases.

The experimental results show that adding N-best recognition hypotheses to the spoken document representation benefits the retrieval performance. However, at the same time more recognition errors are added to the spoken document representation, causing more confusion in the retrieval task. The number of recognition errors increases with a growing number of considered recognition hypotheses. This is the reason why the precision decreases at high recall levels with a growing number of considered recognition hypotheses. These experimental results serve to verify that the most important or useful terms are often contained in the hypothesis in the front rank of the N-best hypotheses. In our case, adding more than 9 recognition hypotheses to the spoken document representation does not help.

Comparison of Different Weighting Schemes

This section focuses on the performance comparison between two N-best specified document term weighting schemes described in Section sec:nbestrank and 4.1.2. We evaluate the retrieval performance of different document weighting schemes on the word-9best representation of a spoken document. The experimental results are summarized in Figure 4.2 and Table 4.3. In our experiment, we set both α and β in Equation 4.5 to 1.

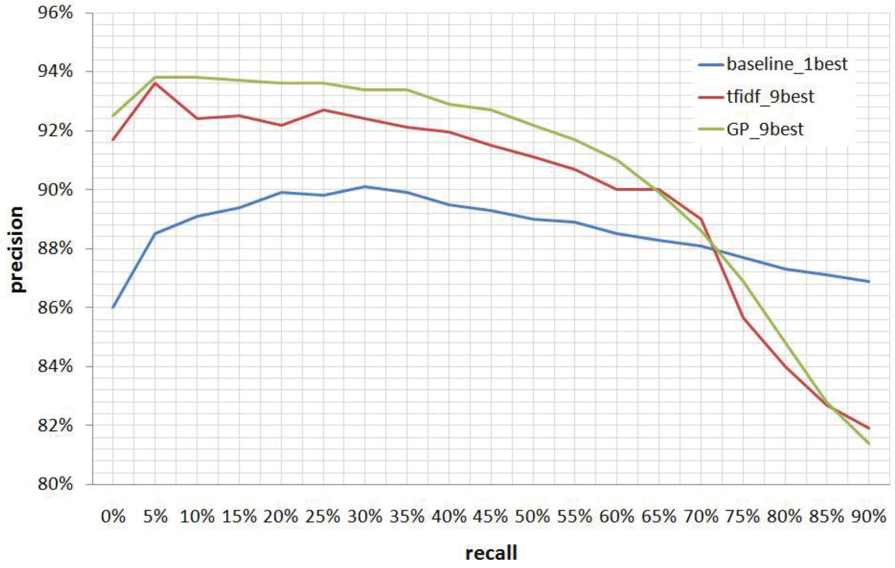


Figure 4.2: Comparison of 9-best specified document term weighting scheme.

In Figure 4.2, the green line labeled GP_9best indicates the weighting scheme that integrates the general term appearance probability into the document term weight (described in Section 4.1.2). The red lines labeled with $tfidf\_9best$ indicates the modified term-frequency weighting method, which takes the rank of the recognition hypotheses containing the query term into account introduced in Section 4.1.1. Improvement in retrieval performance can be achieved with any of the two N-best document term weighting schemes. In comparison with the $tfidf\_9best$ method, the GP_9best weighting method achieves better retrieval precision. Even though there are some fluctuations, a general trend of improved retrieval precision can be observed when we weight the document term with probability. An improvement in precision of about 1.4 % compared to the *tfidf_9best* method can be observed at a recall level of 20 %. However the *tfidf_9best* weighting scheme achieves the highest recall (94%).

recall	0	10	20	30	40	50	60	70	80	90	max.RE	mAP
baseline	86.00	89.12	89.93	90.12	89.59	89.05	88.48	88.10	87.30	86.98	89.77	84.17
tfidf_9best	91.70	92.40	92.20	92.40	91.97	91.10	90.00	89.00	84.00	81.90	94.00	85.40
GP_9best	92.50	93.80	93.60	93.40	92.90	92.20	91.00	88.60	84.80	81.40	92.30	86.10

Table 4.3: Comparison of different N-best specified weighting scheme.

If we look at the results listed in Table 4.3, we find out that the probabilistic weighting scheme *GP_9best* benefits the precision value for recall level from 0 to 70 %. An improvement in mean average precision of about 1.2 % is achieved using the *tfidf_9best* method. Integrating general term appearance probability into the document term weight (the *GP_9best* method) improves the mean average precision by a further 0.7 %. However, in comparison with the *tfidf_9best* method, a slight drop of about 1.7 % in recall can also be observed by the *GP_9best* method.

Conclusions

Experimental results have verified that the N-best transcription can keep more useful information than the baseline word-1best approach. More useful or important terms are included in lower ranked hypotheses. Adding more than 9 best recognition hypotheses into the spoken document representation does not help to improve the retrieval performance, because the number of incorrectly recognized terms will increase with the growing number of recognition alternatives. Integrating general term appearance probability into the document term weight increases the retrieval precision further. The *tfidf_9best* method yields the highest recall.

4.2 Lattice

A lattice [85] is an acyclic graph (as shown in Figure the 4.3). A lattice consists of nodes and edges. A node represents a word that is spoken in a particular period of time. An edge corresponds to the transition between words. The transition is weighted by acoustic and language model probabilities.

A lattice provides word alternatives in different time intervals of the speech signal.

The word error rate of a lattice is defined as the word error rate of the best path in the lattice closest to the reference transcription. It has been verified that the word error rate of a lattice is much lower than that of the 1-best hypotheses. Chelba [15] has reported when the word error rate of a 1-best transcription of a spoken document is 50 %, the word error rate of the lattice representation of this spoken document is only 30%. However, the incorporation of word lattices dramatically increases the size of the representation collection. The size of a lattice could become ten times or even larger than that of the 1-best representation. Information retrieving in the lattice representation of a spoken document is a time-consuming task. In addition, the lattice representation contains only paths allowed by the recognition grammar, which is less flexible for robust domain independent keyword matching.

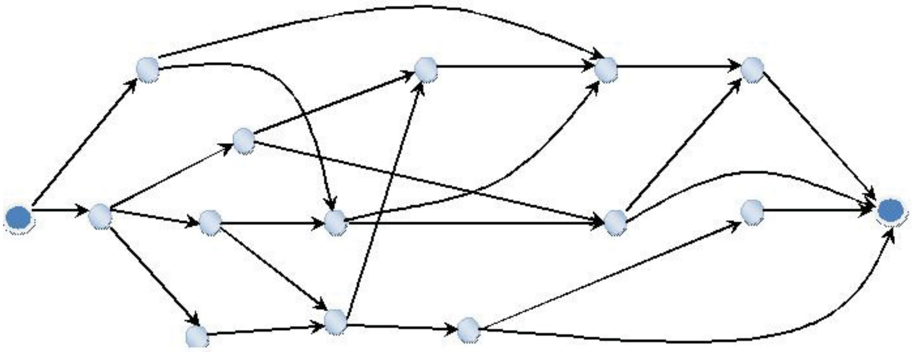


Figure 4.3: A lattice example

A lattice was first applied for a spoken document retrieval task in 1994 by James and Young [26]. They estimated phone level lattice for the spoken documents. The retrieval of spoken content was achieved via fuzzy matching between the query phone representation and phone labels in the lattice. At first, potential start positions of a query, that are lattice nodes labeled with the first phone of the query representation, are identified. Then fuzzy matching is performed for the detection of query word. The main drawback of this approach is that if the first phone label is incorrectly recognized, the query word will not be detected and information will be lost. In James' later work ([51], [52]), vector space retrieval model was modified for the lattice-based spoken document retrieval. A duration-normalized log likelihood ratio (DNLLR) [96] defined in the Equation 4.8 is estimated for each match,

$$S_{DNLLR,W} = \frac{\log P(W|O)}{T_{end} - T_{start}} \tag{4.8}$$

where $P(W|O)$ is the word likelihood provided by the automatic speech recognition system; T_{end} and T_{start} is the time in number of frames;

The term frequency in the term weighting scheme of the vector space retrieval model is defined as the number of lattice nodes labeled with the given term and with a DNLLR value exceeding a predefined threshold. In this section, we will focus on following issues:

- How to estimate DNLLR threshold correctly (discussed in Section 4.2.1)
- Investigation of different document term weighting schemes (described in Section 4.2.2)

The experimental results and conclusions will be summarized in Section 4.2.3.

4.2.1 DNLLR-Threshold Estimation

The definition of the DNLLR threshold plays a very important role in lattice retrieval. The retrieval performance relies heavily on the DNLLR threshold. The DNLLR value distribution in word-1best document representation and word lattice representation is presented in Figure 4.4. The red line indicates the lattice DNLLR value distribution and the blue line represents the DNLLR distribution of ASR 1-best output.

In Figure 4.4, the reasonable region of the threshold should not be far from the range of the DNLLR value of the ASR 1-best output. We then select the first dip (*ca.* -160.0) on the left of one-best range as the minimal threshold for the DNLLR value. Using the same strategy, the maximum DNLLR value is set to -70.0. Ca. 99% of the lattice DNLLR-values fall in this range. To achieve better retrieval performance, fine tuning of the DNLLR threshold is required. The final DNLLR threshold will be determined experimentally, which we will discuss in more detail in Section 4.2.3.

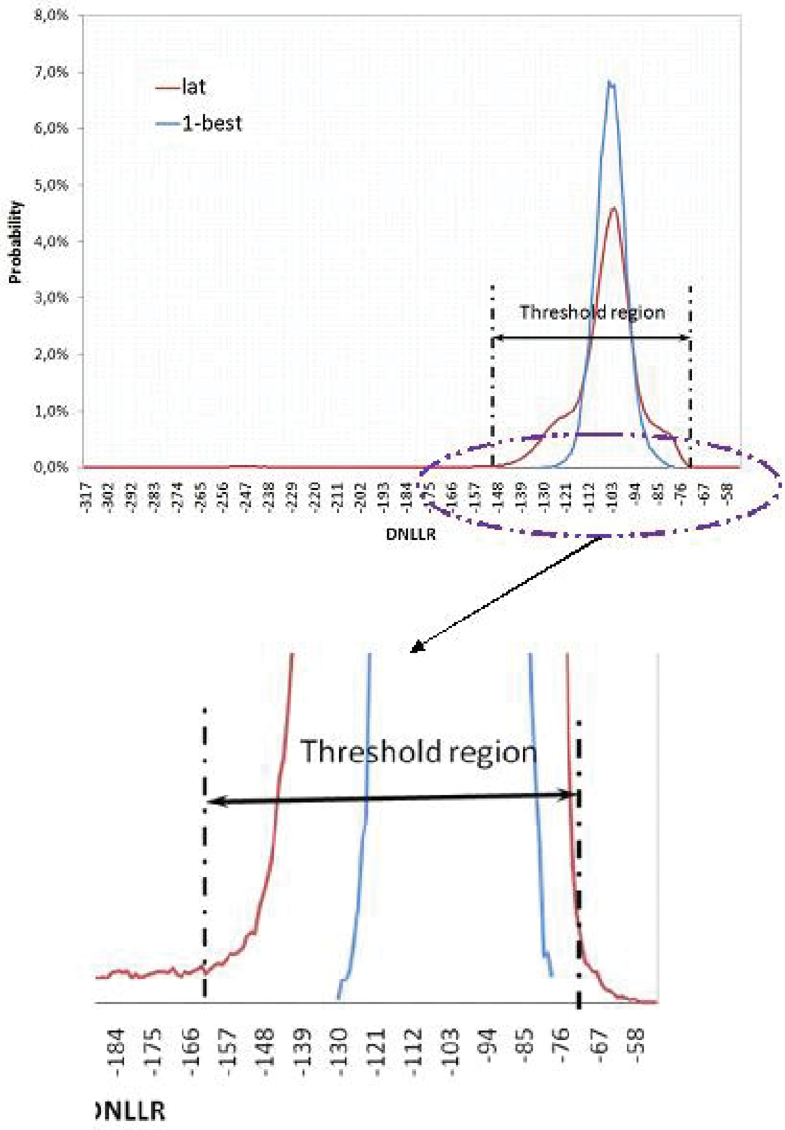


Figure 4.4: DNLLR Thresholding.

4.2.2 Document Term Weighting

We weight query terms with binary values and research the impact of different document term weighting methods on the retrieval performance. **Classical Term-frequency weighting method** can also be applied to the word lattice-based spoken document retrieval. In this case, the term-frequency is assigned to the number of matches, of which DNLLR values exceed the given threshold. In order to compensate the effect of wide variant document length, the term frequency in the weighting scheme will be normalized with the document length. We call this method **normalized term-frequency weighting**. Document length is defined as the number of words. Normalized document term weight $w(t,d)$ is defined as,

$$w(t,d) = \frac{tf_{t,d}}{\sum_{j \in d} tf_{j,d}} \log \frac{N}{df_t} \tag{4.9}$$

where $tf_{t,d}$ indicates the number of matches to the given term t in the lattice representation, of which DNLLR value lies in pre-defined range; $\sum_{j \in d} tf_{j,d}$ denotes the document length; $\log \frac{N}{df_t}$ is the normal inverse document frequency defined in the Equation 2.18.

As written in the following equation, another possibility is to use **DNLLR** value (defined in equation 4.8) directly as the weight for the terms in the document.

$$w(t,d) = \max_{t=t_d, t_d \in d} S_{DNLLR,t_d} \tag{4.10}$$

t_d is a document term that is same to query term t. The effect of different document term weighting on retrieval performance will be experimentally researched in the following Section 4.2.3.

4.2.3 Experiment and Discussion

This section summarizes the experimental results of lattice-based spoken document retrieval methods. Like in the evaluation of the word-N-best based spoken document retrieval methods, 200 in-vocabulary queries are selected for the evaluation task. We first work on fine tuning the DNLLR threshold and its impact on the retrieval performance. The appropriate range of the DNLLR value is then experimentally established. We will ignore all lattice links with the DNLLR values out of range. The effects of different document term weighting schemes on the retrieval effectiveness will be compared.

Fine Tuning of the DNLLR Threshold

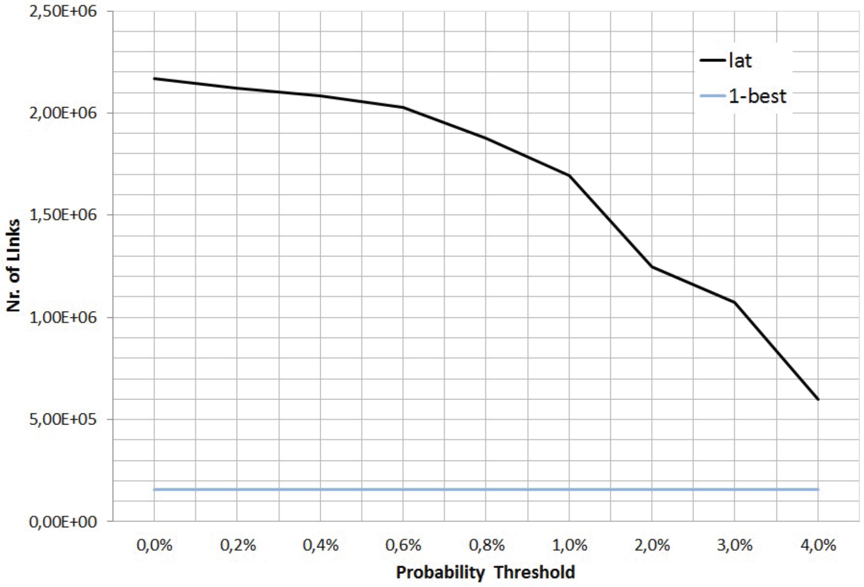


Figure 4.5: Number of valid links in the ASR lattice transcription.

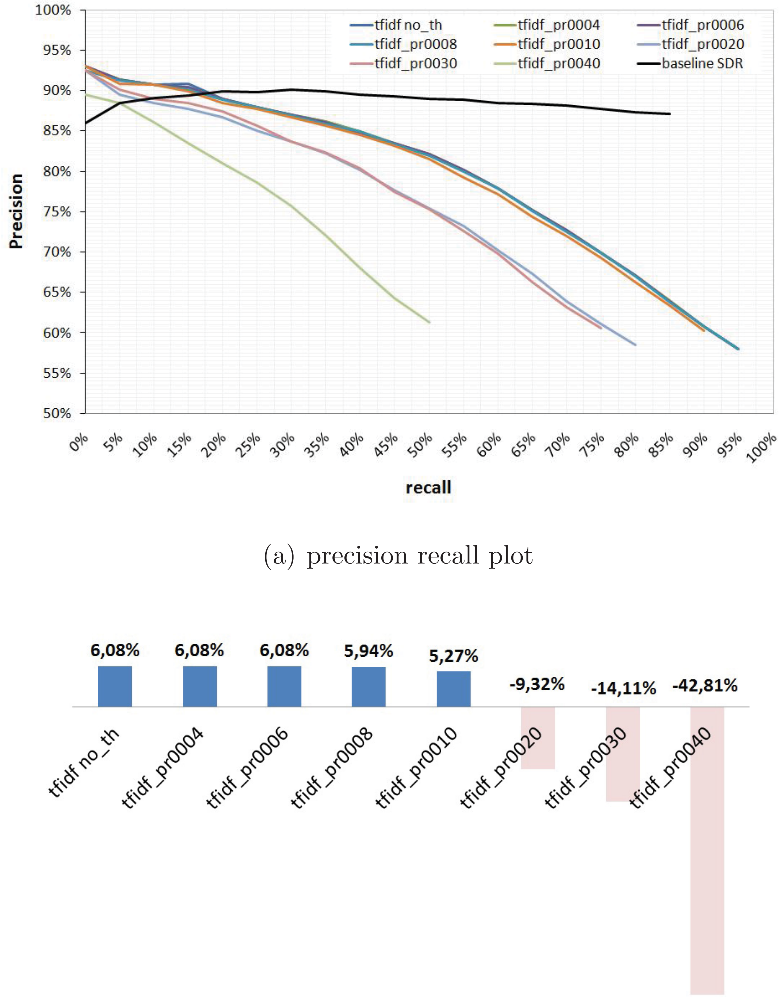
Tuning of the DNLLR threshold is realized by varying the threshold for the probability distribution of the DNLLR value. As shown in Figure 4.5, the number of the valid links decreases with increased probability threshold. The corresponding DNLLR-thresholds are listed in Table 4.4.

Probability (1E-03)	min. DNLLR	max. DNLLR	Nr. of links	label in fig.
2	-137	-75	2124132	tfidf_pr0002
4	-132	-77	2086107	tfidf_pr0004
6	-129	-80	2029050	tfidf_pr0006
8	-125	-87	1879456	tfidf_pr0008
10	-118	-90	1694347	tfidf_pr0010
20	-111	-96	1245673	tfidf_pr0020
30	-108	-96	1071801	tfidf_pr0030
40	-105	-99	598514	tfidf_pr0040

Table 4.4: Selected DNLLR thresholds.

We have selected 200 in-vocabulary queries to evaluate the retrieval performance. The classic $tfidf$ weighting scheme is applied for document term weighting. Figure

4.6 shows the impact of DNLLR threshold on system retrieval performance. The precision/recall values are presented in Table B.1. The mean average precision values are listed in Table 4.5.



(a) precision recall plot

(b) Achieved improvement on recall compared with baseline SDR

Figure 4.6: Effect of DNLLR threshold on system retrieval performance.

recall	baseline	tfidf_ noth	tfidf_ pr0004	tfidf_ pr0006	tfidf_ pr0008	tfidf_ pr0010	tfidf_ pr0020	tfidf_ pr0030	tfidf_ pr0040
mAP (%)	84.17	79.60	79.60	79.60	79.50	76.20	66.20	63.20	42.40

Table 4.5: mAP values for different DNLLR-thresholds

It can be observed that the word-lattice based spoken document retrieval methods provide better retrieval precision at a lower recall level. Compared with the baseline system, an improvement in precision at the forefront of result-list (lower recall-level) can be viewed. An improvement of about 6 % in recall was found at thresholds $tfidf\_noth$, $tfidf\_pr0004$, $tfidf\_pr0006$, $tfidf\_pr0008$ and $tfidf\_pr0010$. However, the retrieval recall drops beyond threshold $tfidf\_pr0010$. Compared with the baseline system, the recall decreases for about 42% at threshold $tfidf\_pr0040$. Links bearing important information are ignored. Therefore, the threshold $tfidf\_pr0010$ (DNLLR value range $[-118, -90]$) is selected for further evaluation of different document weighting schemes. With this threshold, we can achieve an improvement in both recall and precision at lower recall levels. The decrease in mean average precision is kept in an acceptable range (below 3.3 %). From the results of this study, we can conclude that all links in the lattice representation with the DNLLR out of range $[-118, -90]$ can be ignored in the retrieval process.

Comparison of Different Weighting Schemes

In this section, we are going to evaluate the impact of different weighting methods on retrieval performance. The following weighting methods will be evaluated: the classic $tfidf$ method labeled as $tfidf\_pr\_0010$, the normalized $tfidf$ method labeled as $ntfidf\_pr\_0010$ and the document term weighting with DNLLR values ($dnllr\_pr\_0010$). $\_pr\_0010$ indicates the DNLLR range $[-118, -90]$ established in the last section. $ntfidf$ indicates the weighting schema defined in Equation 4.9. $dnllr$ represents the weighting schema defined in Equation 4.10. All lattice links with the DNLLR value out of range $[-118, -90]$ will be ignored during the retrieval. The experiment results are presented in Figure 4.7 and Table 4.6.

The classic term-frequency weighting method $tfidf\_pr\_0010$ and the normalized term-frequency weighting method $ntfidf\_pr\_0010$ yield similar retrieval performance. This result indicates that the documents in the collection have almost the same length. Compared with the baseline system, the $ntfidf\_pr\_0010$ and $tfidf\_pr\_0010$ methods improve the retrieval precision at lower recall levels. The retrieval precision drops at lower recall levels, when the DNLLR values are used directly as document term weight. A degradation of the retrieval precision of about 13% is observed at a recall level of 25%. The precision at higher recall levels yielded by the $dnllr\_pr\_0010$ weighting method-based spoken document retrieval system is comparable to the baseline system.

The results listed in Table 4.6 indicate that any of the three weighting scheme will

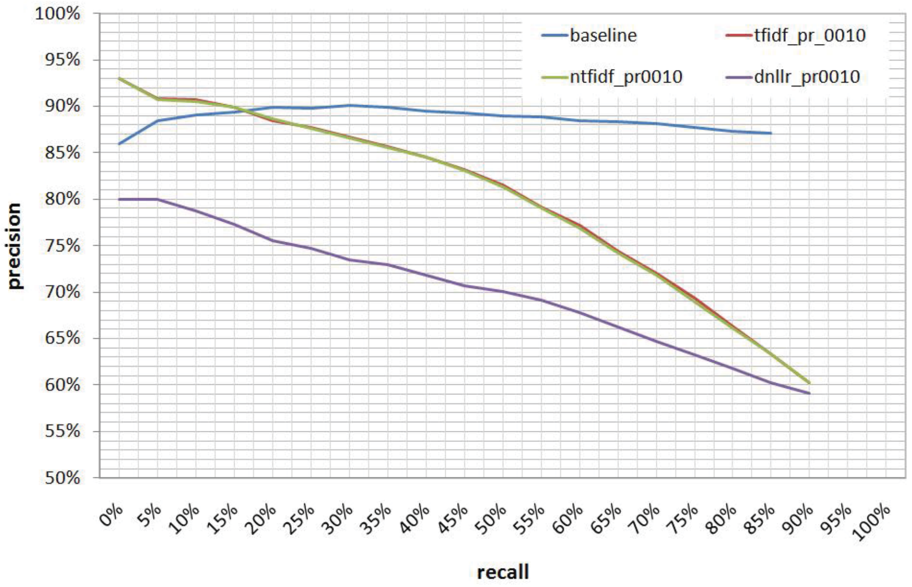


Figure 4.7: Precision/recall curves comparing the performance of different weighting schemes.

	baseline	tf.idf	ntf.idf	dnllr
max. recall (%)	89.77	94.50	94.50	94.50
mAP (%)	84.17	76.20	76.10	66.90

Table 4.6: Result of different weighting scheme for lattice-based SDR.

reach a retrieval recall of 94.5%. Weighting the document term with $tfidf\_pr\_0010$ and $ntfidf\_pr\_0010$ achieves a mean average precision of about 76%.

Conclusion

A word-lattice spoken document representation may contain far more hypotheses than a word N-best representation. DNLLR score was proposed by James [52] for the lattice-based word-spotting task. Term correctness/false alarm trade-off can be varied by setting a threshold on DNLLR score. His experiments showed that term detection were more certain with a higher values of this threshold. Therefore, in this work, the lattice representation of a spoken document was pruned by setting threshold on DNLLR score with aim to keep more correct links. Hypotheses contain correctly and incorrectly recognized terms with similar DNLLR values. A threshold is assigned to the DNLLR value to limit the indexing space, ignoring all links outside of a predefined DNLLR range. In our experiments, the best range of DNLLR is experimentally fixed at $[-118, -90]$.

Different document term weighting schemes are also explored. The *dnllr* weighting scheme treats correct and incorrect terms in the same way. Therefore, more false alarms are added to the final retrieval list. The $tfidf\_pr\_0010$ and $ntfidf\_pr\_0010$ weighting schemes take the number of valid links into account. Here, valid means the link with DNLLR value in the pre-defined range. The experiments show that terms which really appear in the speech signal will occur in more hypotheses. High retrieval precision is achieved with the $tfidf\_pr\_0010$ and $ntfidf\_pr\_0010$ weighting methods.

Chelba and Acero [14] proposed another lattice-based information retrieval method, a position-specific posterior lattice method, that took into account the position of query words within given lattice. The distance of a word from the start node of the lattice was used to form the forward probability. The posterior probability of a word occurring at a given position was computed using this modified forward probability and given standard backwards probability. The relevance score of a document was estimated by aggregating the expected count of each subsequence of query terms in the document according to position. Kazemian [97] reported that PSPL and confusion networks reached a comparable recall. As the confusion networks is known as the most compact representation of multiple hypotheses, it will be discussed more in detail in following sections.

4.3 Confusion Networks

A word confusion network (WCN) is a kind of normalized lattice. It is the most compact representation of multiple hypotheses. The word confusion network was proposed by Mangu [71]. It was originally designed to minimize word errors in recognition output. Based on the assumption that word error and sentence error are highly correlated, the sentence errors will also be minimized when word errors are minimized. A confusion network enforces the alignment of the words that occur at the same approximate time in the lattice. Based on their time span, links in the lattice are clustered. The general structure of a word confusion network is shown in Figure 4.8. Each edge of the confusion network is labeled with word hypotheses and their posterior probabilities. Mangu has verified that the word error rate of the 1-best path in the word confusion network is lower than that of 1-best path in the normal lattice output of ASR.

A confusion network is built of four main steps: an estimation of the posterior probabilities for each link in the lattice, lattice pruning, intra-word clustering and inter-word

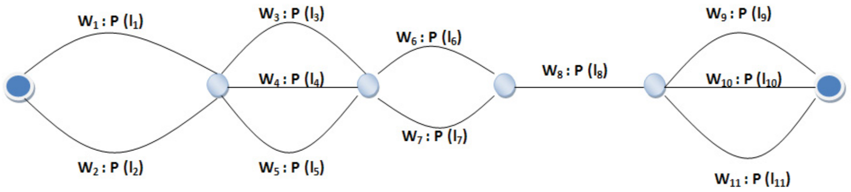


Figure 4.8: General structure of a word confusion network.

clustering. Section 4.3.1 describes how we construct a word confusion network from the recognition lattice. We investigate different methods for word confusion network specified term weighting in Section 4.3.2. The experiment results and discussions are presented in Section 4.3.3.

4.3.1 Construction of a Word Confusion Network

Each link in a lattice can be represented with: starting node $Inode(l)$, ending node $Fnode(l)$, starting time $Itime(l)$, ending time $Ftime(l)$ and word label $Word(l)$. The link posterior probability represents the reliability of $Word(l)$ recognized. After the link posterior probability is estimated, an ordered link equivalence is established. This link equivalence is consistent with the lattice order. The link equivalence is initialized with a class consisting of all links with the same label and within the same period of time. After that, clustering algorithm merges equivalence classes until a totally ordered equivalence is obtained. This is realized in two stages: intra-word clustering and inter-word clustering. In the intra-word clustering stage, clusters with the same word instance are merged. Based on the phonetic similarity between word components, heterogeneous clusters are grouped in the inter-word clustering step.

Section 4.3.1 introduces how we estimate the posterior probability for each arc in the lattice. The links with very low posterior probabilities will then be ignored for further steps. Intra-word clustering is described in Section 4.3.1, which also discusses the inter-word clustering stage.

Posterior Estimation and Lattice Pruning

We use L to represent the set of links in a word lattice. Each link l contains information about the starting node $Inode(l)$, ending node $Fnode(l)$, starting time $Itime(l)$, ending time $Ftime(l)$ and word label $Word(l)$. The posterior probability $p(l)$ of each link is computed from the acoustic and language model scores in the lattice, as the sum of the posterior probabilities of all paths passing through link l. Given a sequence of acoustic observation $o_1^T = o_1, ..., o_T$, the posterior probability of path q with word sequence $w_1^M = w_1, ..., w_M$ is expressed as:

$$p(q|o_1^T) = p(w_1^M|o_1^T) = \frac{p(o_1^T|w_1^M) \cdot p(w_1^M)}{p(o_1^T)} \propto p(o_1^T|w_1^M) \cdot p(w_1^M) \tag{4.11}$$

where $p(w_1^M)$ denotes the language model probability; $p(o_1^T|w_1^M)$ is the acoustic model probability; $p(o_1^T)$ represents the probability of the acoustic observations. As written in the Equation 4.12, the posterior probability of a link $p(l)$ is the sum of posterior probabilities of all paths that pass through link l.

$$p(l) = \sum_{q \in Q_l} p(q|o_1^T) \tag{4.12}$$

where Q_l represents all paths that pass through link l; For a large scale lattice, the link posterior probability is estimated using the forward-backward algorithm shown in Figure 4.9.

We are going to estimate the posterior probability of link $l(n_i, n_f, t, t-1, w_l)$. Forward-backward probabilities are estimated inductively. In this thesis, the forward probability $\alpha_t(i)$ is defined by the Equation 4.13 as the probability of observing the t feature vectors and being in node n_i at time t. It can be deduced by summing the forward probabilities for all possible predecessor nodes j weighted by the acoustic and language model probability. The forward probability of start node n_i is then computed as follows:

$$\alpha_t(i) = \sum_j \alpha_{t-1}(j) \cdot P(o_t|w_j) \cdot P(w_j|w_{j-m}^{j-1}) \tag{4.13}$$

where j indicates all possible predecessor nodes of n_i $P(o_t|w_j)$ represents the acoustic model probability, and $P(w_j|w_{j-m}^{j-1})$ represents the m-gram language model probability.

The backward probability $\beta_{t+1}(f)$ of end node n_f of link l is estimated as in the

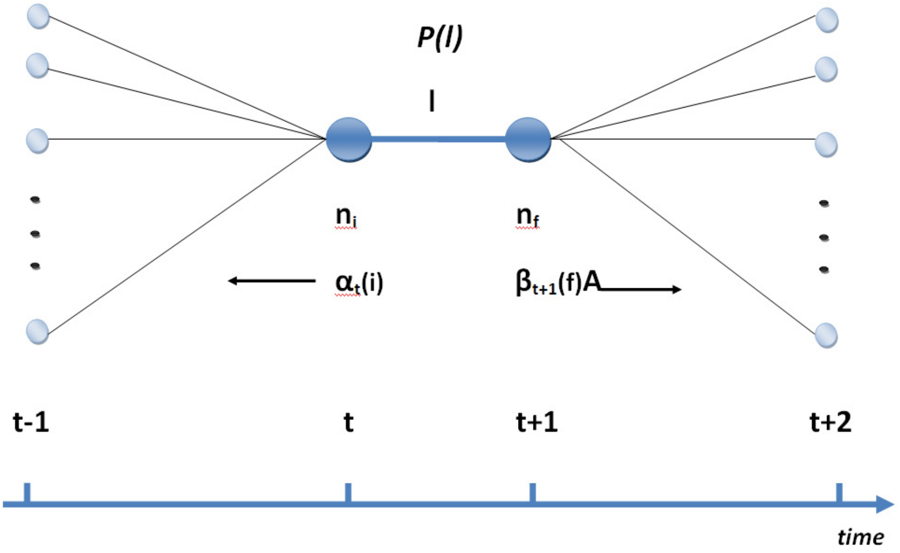


Figure 4.9: Link posterior estimation using the forward-backward algorithm

forward case,

$$\beta_{t+1}(f) = \sum_{k} \beta_{t+2}(k) \cdot P(o_{t+2}|w_k) \cdot P(w_k|w_{k-m}^{k-1}) \tag{4.14}$$

where $\beta_{t+2}(k)$ denotes the backward probabilities of all possible successor nodes k weighted by the acoustic model probability $P(o_{t+2}|w_k)$ and the m-gram language model probability $P(w_j|w_{j-m}^{j-1})$. Both forward and backward probabilities are initialized with value 1. The posterior probability of link l is then estimated as:

$$p(l) = \alpha_t(i) \cdot \beta_{t+1}(f) \cdot P(o_{t+1}|w_l) \cdot P(w_l|w_{l-m}^{l-1}) \tag{4.15}$$

Links with very low posterior probabilities will be removed from the lattice.

Intra-word Clustering

The arcs with the same word label starting times and ending times are merged in intra-word clustering stage. The following metric is used as similarity measure between two sets of links:

$$SIM(E1, E2) = MAX_{e1 \leq E1, e2 \leq E2} Overlap(e_1, e_2) \cdot p(e_1) \cdot p(e_2) \tag{4.16}$$

where $Overlap(e_1, e_2)$ indicates the temporal overlap between two link equivalences (e_1 and e_2) normalized by their maximum lengths. $p(e_1)$ and $p(e_2)$ are posterior probabilities of the link equivalence.

The links are clustered iteratively. At each step similarities between all possible pairs of cluster candidates are computed. The most similar clusters will be grouped together. This process will terminate when there is no potential pair of link equivalences left to be merged.

Inter-word Clustering

The goal of inter-word clustering is to merge phonetically similar word-labeled link equivalences that are competing for the same position in the reference transcription. Inter-word clustering is based on the similarity $SIM(E_1, E_2)$ between two links equivalences (E_1 and E_2), and computed as follows

$$SIM(E_1, E_2) = AVG_{w_1 \in Words(E_1), w_2 \in Words(E_2)} sim(w_1, w_2) \cdot p_{E_1}(w_1) \cdot p_{E_2}(w_2) \quad (4.17)$$

where

$$p_{E_1}(w_1) \; = \; p\{e \in E_1 : Word(e) = w_1\} \quad (4.18)$$

$$p_{E_2}(w_2) \; = \; p\{e \in E_2 : Word(e) = w_2\} \quad (4.19)$$

$sim(w_1, w_2)$ is the phonetic similarity of two words (w_1 and w_2) and is estimated as:

$$sim(w_1, w_2) = 1 - \frac{Edit\_Distance(PS_1, PS_2)}{max(len(PS_1), len(PS_2))} \quad (4.20)$$

where $Words(E_1)$ represents the words clustered in Equivalence class E_1. PS_1 is the phone sequence of word w_1; The phone sequence of word w_2 is represented with PS_2. $Edit\_Distance(PS_1, PS_2)$ is the number of operations required to transform PS_1 into PS_2. The Levenshtein distance is normally used to compute this metric. The edit distance between two phone sequences will be normalized with the length of the longer sequence. The length of a phone sequence is defined as the number of phones in a sequence. The inter-word clustering will be continued until total order has been achieved. Here, total order means that there are no potential link equivalences left to be merged.

4.3.2 Word Confusion Network Specified Term Weighting Methods

We weight the query term with binary information and explore different possibilities for document term weighting. The **classical $tfidf$ weighting scheme** introduced in Section 2.3.2 is used directly for document term weighting. Here, the term frequency means the number of occurrence of the terms in the word confusion network representation. This method does not take the uncertainty of the speech recognition system into account, because it was originally developed for use on text document collections, where the words are assumed to be known with certainty. The **posterior probability** of a word represents the reliability that a word w is correctly recognized. This information should be useful for spoken document retrieval. The posterior probability of a term is directly used as document term weight. The document term weight is estimated as:

$$w_{t,d} = \sum_{w(l)=t} p(l) \tag{4.21}$$

where $w(l)$ means the word of link l; $p(l)$ means the posterior probability of link l. The document relevance score is estimated as defined in equation 2.22. The limitation of this weighting scheme is that content words and function words are treated equally. In other words, the function/stop words are as important as content words, as the posterior probability of a word can only provide information about the reliability of the recognized word. However, it has been verified that removing the function words from the document representation improves the retrieval performance [98].

Combining $tfidf$ and Posterior Probability for Document Term Weighting

We propose a new method to document term weighting, where the uncertainty of the speech transcription and term frequency are integrated. As defined in the following equation, the new document weight is the linear combination of word posterior probability and term-frequency weight.

$$w(t,d) = \alpha \cdot \underbrace{\frac{tf_{t,d}}{len_d}}_{tf} \cdot \underbrace{\log \frac{N}{df_t}}_{idf} + (1-\alpha) \sum_{l(w)=t} p(l) \tag{4.22}$$

where $tf_{t,d}$ is the frequency of term t in document d; len_d indicates the total number of terms in document d; the total number of documents in the entire spoken document collection is denoted by N; the number of documents including term t is denoted by df_t.

4.3.3 Experiment and Discussion

In this section, we evaluate the retrieval performance with 200 in-vocabulary queries. First, we perform retrieval based on the word confusion network representation of the spoken documents. We make a comparison between the word confusion network-based and the lattice-based spoken document retrieval system. Then we evaluated the impact of term-weighting methods on the retrieval performance.

Word Confusion Network vs. Lattice

Compare the word confusion network representation of a spoken document with its lattice representation, we find out that the total number of links in the document representation is significantly reduced. This result can be seen in Figure 4.10. In this Figure, the label *lattice_noth* denotes the original lattice provided by an automatic speech recognizer. The labels $tfidf\_pr0004$ to $tfidf\_pr0040$ represent the lattice after DNLLR-value based pruning, according to different DNLLR thresholds. The label WCN denotes the word confusion networks. The label *baseline* is the number of links in the word-1best representation.

In comparison with the original lattice, it can be seen in Figure 4.10 that the number of links contained by the work confusion network has been reduced by about 76.5 %.

Figure 4.11 and Table 4.7 present the results of the performance evaluation experiments. We make a comparison of different spoken document retrieval systems that are based on word-1best, lattice and word confusion network representation. It can be observed that the highest recall 95.2 % is reached using the lattice-based spoken

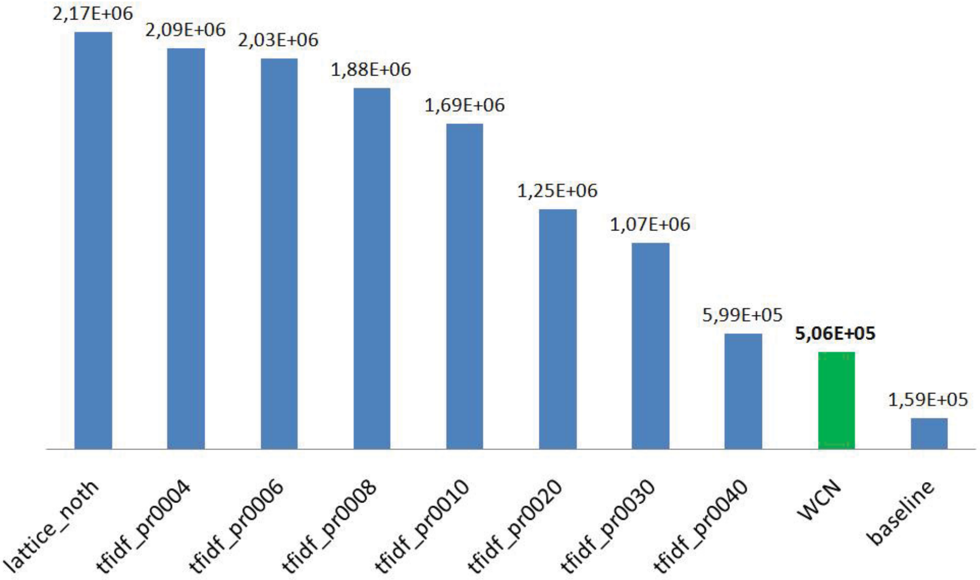


Figure 4.10: Number of links included in lattice, WCN and baseline word-1best representation.

document retrieval method. In comparison with the word-1best based methods, an improvement of about 5.5 % in recall has been achieved. Pruning the lattice representation according to the DNLLR value leads to a drop in recall and precision.

The recall is kept (95.2 %) when using word confusion network as the representation of a spoken document. The word confusion network representation of a spoken document contains the same number of links as in the pruned lattice representation (DNLLR range lies in $[-105, -99]$). An improvement of about 22.1 % in mean average precision can be achieved when using a word confusion network instead of a pruned lattice as the representation of a spoken document. The results of this study indicate that even though a number of links have been ignored or merged during the build-up of a word confusion network representation, relevant information is kept. Similar behavior of the retrieval performance is observed in results listed in Table 4.7. The word confusion network based spoken document retrieval method yields a mean average precision

	lattice									
	noth	pr0004	pr0006	pr0008	pr0010	pr0020	pr0030	pr0040	**WCN**	baseline
Nr. links(E+05)	21.70	20.90	20.30	18.80	16.90	12.40	10.70	5.99	5.06	1.59
mAP (%)	79.60	79.59	79.57	79.46	76.21	66.16	63.23	42.43	64.52	84.17
max.RE (%)	95.23	95.23	95.23	95.10	94.50	81.40	77.10	51.34	95.23	89.77

Table 4.7: Performance (mAP and max. RE) comparison between baseline, lattice and WCN.

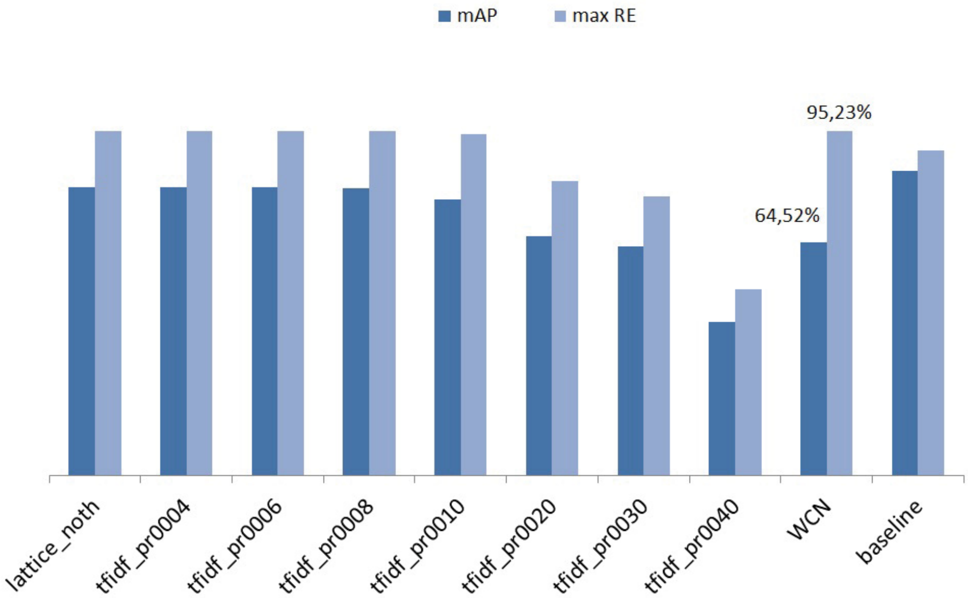


Figure 4.11: The plot of mAP and recall for spoken document retrieval on lattice, confusion network and 1-best transcription.

of about 64.5 %.

Term Weighting with Posteriors

We evaluate the performance of the word confusion network-based spoken document retrieval that directly uses term posterior probabilities as document term weights. The vector space model is applied for the retrieval task. The experimental results are presented in Table 4.8. The query terms are weighted with a binary value. *WCN_tfidf* represents the WCN-based method using *tfidf* weighting. *WCN_posterior* denotes the WCN-based method using posterior weighting schema. In comparison with *WCN_tfidf* method, it can be observed that the *WCN_posterior* method improves the retrieval precision at lower retrieval recall levels. The retrieval precision drops with increasing recall rate. The improvement in $E1$ identifies that compared with the *WCN_tfidf*

recall	0	10	20	30	40	50	60	70	80	90	max. RE	mAP
baseline	86.00	89.12	89.93	90.12	89.59	89.05	88.48	88.10	87.30	86.98	89.77	84.17
WCN_tfidf	66.00	70.30	68.12	67.58	66.78	65.35	64.18	62.54	60.78	58.77	95.23	64.92
WCN_posterior	74.50	68.90	67.20	66.20	65.40	64.50	62.80	61.40	60.20	-	87.19	58.70

Table 4.8: Performance (recall, precision, mAP and max.RE) comparison of baseline and WCN based SDR.

($E1$ = 66 %), more queries ($E1$ = 74.5 %) had their right answer in the front of the retrieved result list when the posterior probabilities are directly used as document term weights (*WCN_posterior* method). However, the recall drops with posterior probability weighting represented by *WCN_posterior*, and only 87.19% is achieved. Some words bearing important information are ignored during the retrieval process, because of their low posterior probabilities. The decrease in mean average precision (about 6.22%) is caused by misrecognized words with high posterior probabilities.If there are no alternative word hypotheses in the same segment of time, the word posterior is set to 1.

Combining *tfidf* and *posterior* Probabilities for Term Weighting

This study evaluates the new term weighting method proposed in Section 4.3.2. We first examine the impact of the α selection on retrieval performance. The experimental results are presented in Figure 4.12 and Table 4.9.

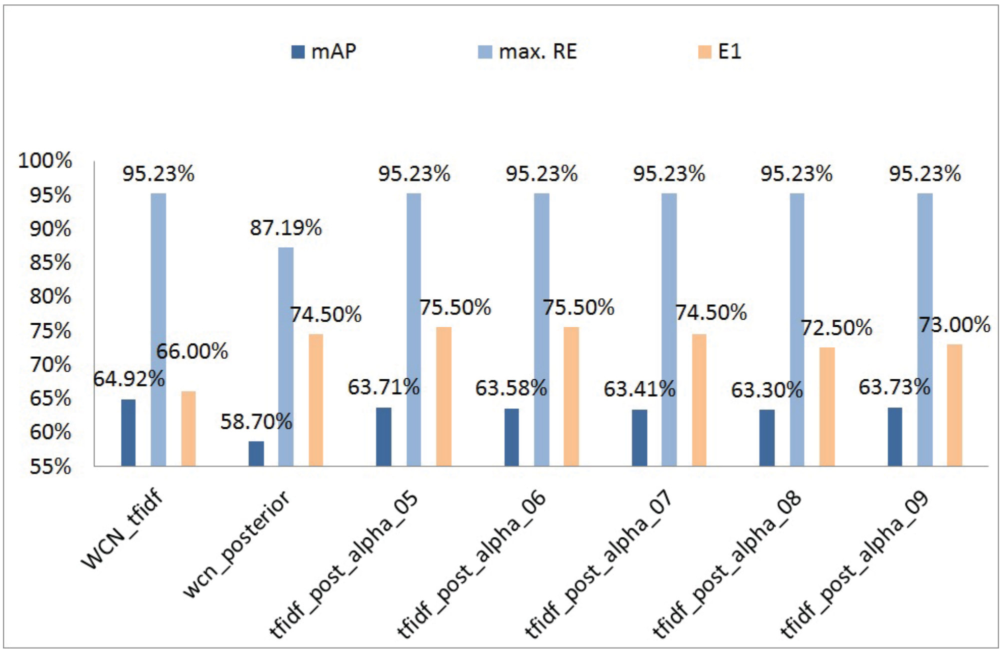


Figure 4.12: Comparison of different weighting schemes for word confusion network based spoken document retrieval approaches.

It can be observed that the number of queries with correct answers in the front rank of the retrieved result list has increased. When using the combined weighting scheme (α = 0.5 and α = 0.6 ($E1$)) about 75.5 % of the queries have their correct answer at rank 1. There are more than one correct answer for each query. The $E1$ value indicates whether one of the correct answers of a query listed at first rank. The recall reaches 95.23 %. A slight decrease in mean average precision value is observed (1.2 % compared to the $tfidf$ weighting).

			Combined weighting (tfidf-post)				
	WCN-tfidf	WCN-posterior	$\alpha = 0.5$	$\alpha = 0.6$	$\alpha = 0.7$	$\alpha = 0.8$	$\alpha = 0.9$
mAP (%)	64.92	58.70	63.71	63.58	63.41	63.30	63.73
max. RE (%)	95.23	87.19	95.23	95.23	95.23	95.23	95.23
E1(%)	66.00	74.50	75.50	75.50	74.50	72.50	73.00

Table 4.9: Performance of WCN-based SDR when using different weighting schemes.

Comparison of different robust word-based spoken document retrieval methods

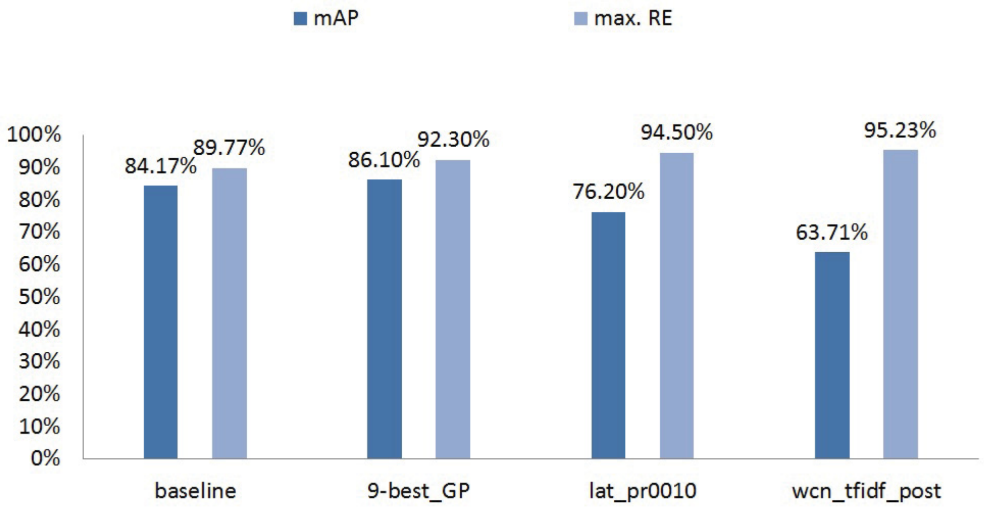


Figure 4.13: Performance comparison of robust SDR approaches using different ASR-representation containing multiple hypotheses.

It has been confirmed by experiment in Section 4.1.3 that $9best\_GP$ is the best performed SDR method based on N-best spoken document transcription. The experiments in Section 4.2.3 verified that the lat_pr0010 method (The range of the DNLLR value of links in the lattice is set to $[-118, -90]$) outperforms other pruned lattice based SDR approach. We compare those robust word-based spoken document retrieval systems with WCN-based method in this study.

The experimental results are presented in Figure 4.13 and Table 4.10. The label *GP_9best* indicates the spoken document retrieval approach based on the representation of a spoken document that contains 9-best recognition hypotheses, and at the same time the word general probability is integrated in the document term weights. The *lat_pr*0010 method represents the pruned lattice-based spoken document retrieval approach. The label $WCN(\alpha = 0.5)$ represents the word confusion network based spoken document retrieval method with a novel weighting scheme that linearly combines the term frequency and posterior probabilities as document term weights with $\alpha = 0.5$.

recall (%)	0	10	20	30	40	50	60	70	80	90	max. RE	mAP
baseline	86.00	89.12	89.93	90.12	89.59	89.05	88.48	88.10	87.30	86.98	89.77	84.17
GP_9best	92.50	93.80	93.60	93.40	92.90	92.20	91.00	88.60	84.80	81.40	92.30	86.10
lat_pr001	93.00	90.70	88.50	86.70	84.50	81.50	77.20	72.00	66.30	60.30	94.50	76.20
WCN(α=0.5)	75.50	69.60	67.10	65.20	63.80	62.50	61.40	60.30	59.20	58.20	95.23	63.71

Table 4.10: Precision/Recall, mAP and max. Recall (in %) value of SDR approaches using different ASR transcription containing multiple hypotheses.

From the data in Figure 4.13 and Table 4.10 that compared the baseline spoken document retrieval system that is based on the 1-best recognition result, we can see that the retrieval performance is improved when using document representation containing multiple ASR hypotheses. The best mean average precision (86.1%) is observed at 9-best transcription based spoken document retrieval methods. A relative improvement in mean average precision of 2.3% is reached, compared with the baseline system. A maximal retrieval recall of 95.2% is achieved using the word confusion network based spoken document retrieval approach, which is corresponding to a relative improvement of about 6%, compared with the baseline system. A drop in mean average precision can be observed.

Conclusions

The results of these studies verify that the word confusion network is a very good alternative to lattice transcription for spoken information retrieval. Compared with normal lattice-representation, the number of links in the word confusion network is reduced by 76.5 %. The word confusion network based spoken document retrieval achieves a recall of about 95.23 %. However, compared with the lattice-based spoken document retrieval methods, a drop in mean average retrieval precision (about 15.1 %) is observed using the word confusion network as the document representation. Using posterior probabilities directly as the document term weights improves the retrieval precision at low recall levels. However misrecognized terms with high posterior probability in the document representation result in the drop of the mean average precision (6.22 % compared with *tfidf* weighting). The number of queries that have their right answer at rank 1 in the result list increases when we use the new weighting method, which combines term frequency and posterior probability for the document term weighting.

The maximal retrieval recall is achieved with the word confusion network based spoken document retrieval approach (95.2%), which is corresponding to a relative improvement of about 6 % compared to the baseline system.

4.4 Dealing with OOV Words via Representation Expansion

Mismatch between the terms in the query and the terms that compose relevant documents is one fundamental problem in information retrieval. Term mismatch is often caused by recognition errors in the spoken document representation. We have said before that the recognition errors in document representation are caused by two factors: mismatch between training and application environment and out-of-vocabulary words in the query and the speech segment.

In this section, we are going to investigate a robust spoken document retrieval method dealing with the out-of-vocabulary words. In speech transcription, out-of-vocabulary words spoken will be replaced with the most acoustically similar terms in the recognition vocabulary. Consequently, relevant documents composed of out of vocabulary words cannot be detected. The recognition vocabulary also limits the query vocabulary. The problem caused by the out-of-vocabulary words could be handled with subword-based spoken document retrieval methods, which will be discussed later in Chapter 5. An alternative solution to the out-of-vocabulary word problem is to find a way to better represent need for information, namely the expansion of the query and/or the document representation. Representation expansion, as the term suggests, means that the query or document representation will be modified with additional terms. For each out-of-vocabulary word, a set of similar words in the recognition vocabulary will be added into the representation. In this thesis, we ignore the context similarity between terms and focus only on their acoustic similarity.

The acoustic similarity between two words is often set to the edit distance between their phonetic sequences. The edit distance between two phone sequence is computed with the Levenshtein algorithm. The alternative in-vocabulary term with minimal edit distance to the out-of-vocabulary words will be added into the query representation or even replace the original one.

Query Expansion Model

Crestani proposed to expand the query representation with more approximate terms [19]. In his work, phonetic confusion information is integrated into the similarity measure between two words. The phonetic confusion information is often saved as a matrix and includes statistics on the substitution, insertion and detection errors of a phone recognizer. Section 5.3.1 will describe phonetic confusion information in more detail. Approximate alternatives are selected in the following steps:

- Assign a threshold to the confusion value;
- All phone pairs in the confusion matrix whose confusion value is below the threshold will be selected
- The approximate terms are built via replacing the original phone in each query term with the hypothesis phone in selected phone pairs.

This representation expansion runs independently from the document collection. A fixed number of new similar terms will be added to the query representation. The query contains more terms. If a new term l is selected to approximate an original term t, it will be weighted by:

$$w(l,q) = sim(t,l) \cdot w(t,q) = \frac{\sum\limits_{m=1}^{N} C(t[m], l[m])}{\sum\limits_{m=1}^{N} C(t[m], t[m])} \cdot w(t,q) \tag{4.23}$$

where $sim(t,l)$ represents the similarity between phone sequences of the new alternative term l and the original term t; w(t,q) is the weight of original query term t; and phonetic confusion information is used to compute the $sim(t,l)$. The phone sequences of the original term t and the phone sequence of alternative term l include N phones. $C(t[m],t[m])$ comes from the phone confusion information matrix and indicates the probability that phone $t[m]$ is correctly recognized. $C(t[m],l[m])$ denotes the substitution probabilities in the confusion matrix. It represents the probability that phone $t[m]$ is recognized as phone $l[m]$.

However, it is very hard to find a suitable threshold value. An incorrectly defined confusion threshold results in too many similar terms being inserted; thus, the retrieval precision drops. In addition, this method only takes substitution errors into account. The effect of insertion and deletion error on the retrieval performance is ignored. Similar approach was also proposed by Saraclar et. al. [101]. In their proposal, the word or subword sequence of each query will be transformed into the sequences that the query tends to be mi-recognized to.

Moreau proposed an alternative query expansion method that takes all insertion, deletion and substitution errors into account [77]. The terms t in the query representation ($t \in (Q \cap \overline{D})$) will be replaced with the most similar terms l in the document representations. Queries contain a same number of terms. The relevance score $Sim_{Exp}(q,d)$ of the document d is then computed as:

$$Sim_{Exp}(q,d) = \sum_{t \in Q} P(u_t|t) \cdot w(t,q) \cdot w(u_t,d) \tag{4.24}$$

$$u_t = \begin{cases} t & if \quad t \in (Q \cap D) \\ arg[max_{t' \in D} P(t'|t)] & if \quad t \in (Q \cap \overline{D}) \end{cases} \tag{4.25}$$

where $P(u_t|t)$ denotes the confusion probability between term u_t and t. $w(t,q)$ is the query term weight; $w(u_t,d)$ is the weight for term u_t in the document. The out-of-vocabulary-term t ($t \in (Q \cap \overline{D})$ will be replaced in the document by the most similar term t'. This query-expansion methods proposed by Moreau will be experimentally researched in the following section.

Lee ([61]) discussed about expanding the representation of a query with a set of acoustic patterns automatically learned from the target spoken archive in an unsupervised way. Two levels of acoustic patterns are in use, the word-like patterns and subword-like patterns. The experimental results of the Mandarin news retrieval-task verified the usefulness of this computationally intensive method.

Yi and Allan ([120]) proposed expanding the query with the words in documents, that are semantic similar to the given query word. They discussed several ways of calculating a query-specific topic either from feedback documents or from the whole corpus. Their experiments showed that topics discovered in the whole corpus are too coarse-grained to be useful for query expansion. Topics detected in the query related relevant documents can benefits the retrieving performance. However, the methods supporting this kind of topics are very sensitive to the parameters. Incorrectly tuned method-parameters may lead to the drops of retrieving performance.

Document Expansion Model

The **document expansion** method replaces the term that appeared only in the document with the most similar term in the query or extend the document with similar terms in the equivalence list of a given query term. The term similarity here means the acoustic similarity or/and contextual/semantic similarity Jourlin et. al. [87] proposed to extend a document representation with synonyms of query words. An equivalence list was built for each query term. In addition, other specific entities that belong to the same general semantic entity of a query term were also added into the equivalence list, the so-called semantic *poset*. For example, the equivalence list of the query word *Europe* may contain the names of all European countries, regions and cities. The Information Retrieval techniques were implemented within the Probabilistic Retrieval Model. Documents are ranked in order of decreasing estimated probability of relevance. However, very little improvement in Information Retrieval performance has been verified with their experiments. In their experiments, a gain of 17% in average precision on speech transcripts provided by a word recognizer can only be achieved by combining semantic *poset* with blind relevance feedback.

But from the document point of view, however, it is a computationally intensive task to add similar terms to each document in a huge collection. For this reason, we only focus on query expansion model in this thesis.

Experiment and Discussion

We are going to evaluate the query expansion model proposed by Moreau [77] and its ability to deal with recognition errors, more specifically its ability to deal with recognition errors caused by out-of-vocabulary words. 200 in-vocabulary queries and

30 out-of-vocabulary queries are selected for the retrieval performance evaluation task. The 30 out-of-vocabulary queries are replaced with their acoustically most similar terms in recognizer vocabulary. Table 4.11 shows some examples.

words	**sim. terms**	**words**	**sim. terms**
BISMARCK	EARMARKED	CONTINENTS	CONTINENT
HERDING	ADDING	INORGANIC	BARGAINING
RADIATOR	CREATOR	RODHAM	ADAM
WORKFORCE	ENFORCE	HILLARY	ARTILLERY
SPOKESPERSON	SPOKESWOMAN	SPILLER	DISTILLER
POLYSTYRENE	ASPIRANTS	BIKING	BITING
CANINE	NINTH	IMPLANTS	COMPLAINTS

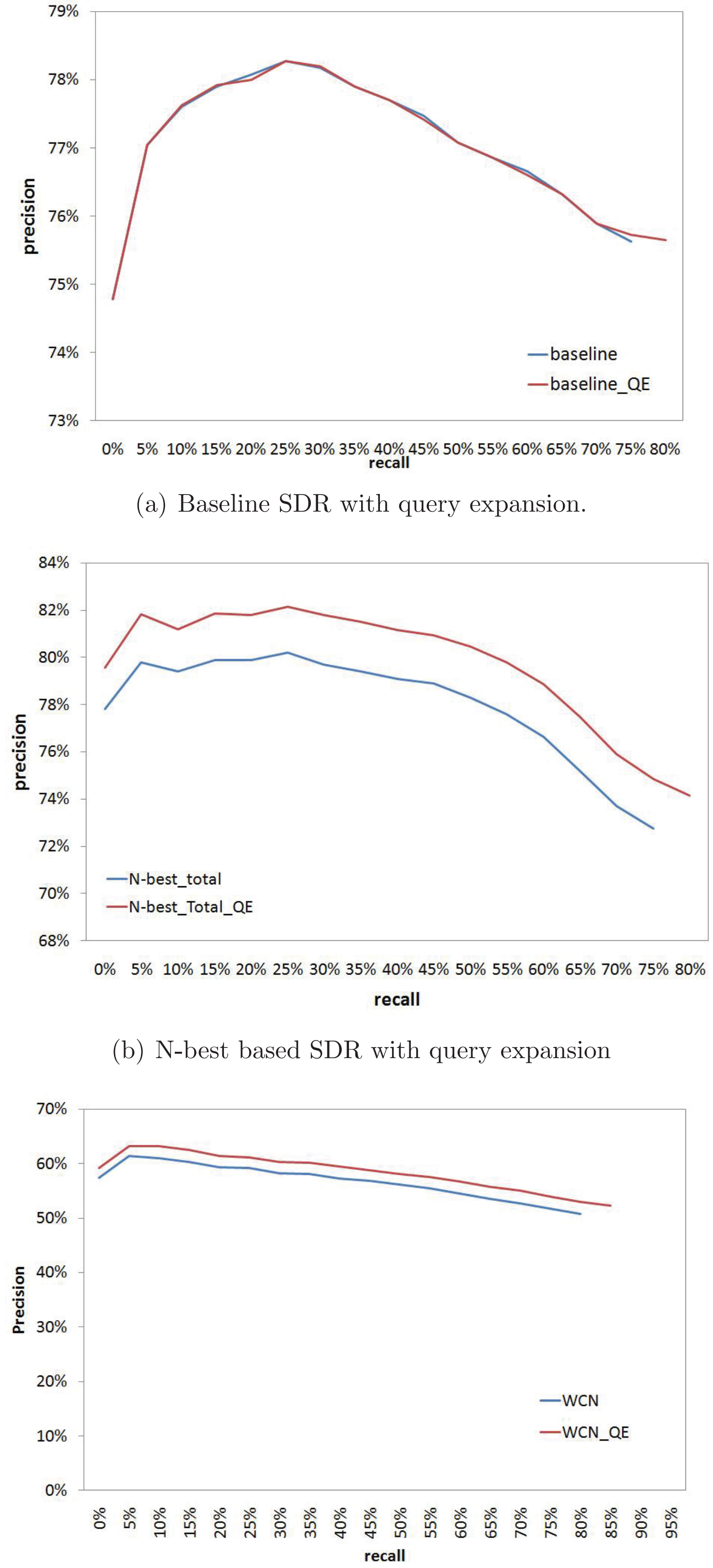
Table 4.11: OOV words and similar terms.

The experimental results are presented in Figure 4.14 and Table 4.12. The retrieval performance of our baseline system based on 1-best ASR output before and after query expansion is shown in Figure 4.14(a). It can be observed that the mean average precision is improved by the query expansion model, while the recall remains the same.

recall	onebest	onebest+QE	5-best	5-best+QE	WCN	WCN+QE
0	74.78	74.78	77.80	79.56	57.39	59.13
10	77.60	77.63	79.40	81.19	61.05	63.21
20	78.08	78.00	79.90	81.80	59.37	61.43
30	78.18	78.20	79.70	81.81	58.26	60.34
40	77.70	77.70	79.10	81.15	57.27	59.51
50	77.08	77.08	78.30	80.46	56.11	58.04
60	76.66	76.60	76.63	78.86	54.54	56.77
70	75.89	75.89	73.69	75.90	52.74	55.00
80	-	75.65		74.16	50.79	52.96
mAP	61.67	65.45	62.41	67.77	48.19	52.59
max.recall	78.06	78.06	79.76	80.67	82.80	84.26

Table 4.12: The effect of query expansion on the performance of a word-based SDR.

A clear improvement in both mean average retrieval precision and recall can be observed with the N-best and word confusion network based spoken document retrieval with query expansion (see figures 4.14(b) and 4.14(c)). Integrating query expansion into the N-best based SDR approach achieves the highest improvement in mean average precision (5.35 %), while the highest improvement in maximal achievable retrieval recall is observed with the word confusion network based spoken document retrieval approach (1.46 %):



(a) Baseline SDR with query expansion.

(b) N-best based SDR with query expansion

(c) WCN based SDR with query exansion.

Figure 4.14: Performance evaluation of word-based SDR with query expansion.

The expansion of the query representation with their acoustic similar candidates in the documents enables the word-based spoken document retrieval system to deal better with OOV queries. The results for retrieving OOV queries are shown in Figure 4.15.

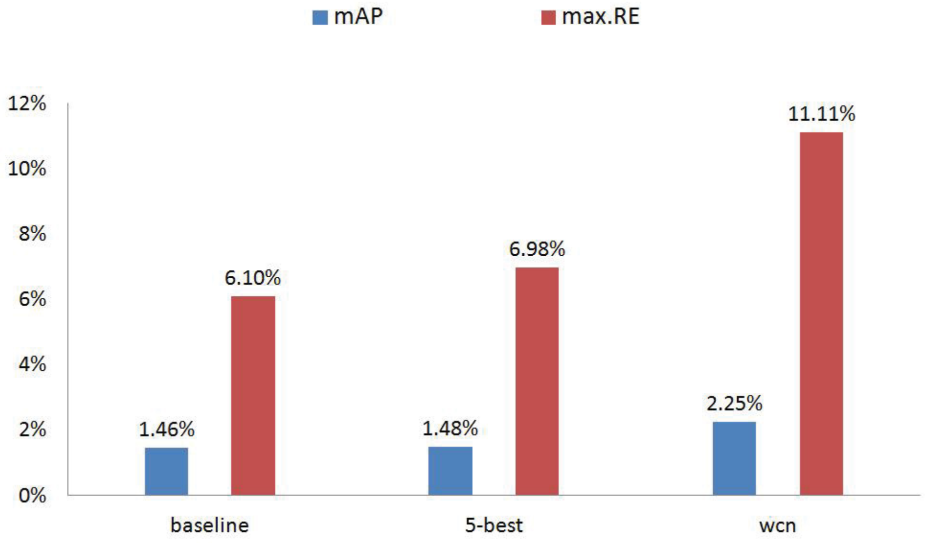


Figure 4.15: Performance of retrieving OOV queries with query expansion

4.5 Summary

In this chapter, we have explored a number of robust word-based spoken document retrieval methods in an effort to improve the retrieval performance when there are recognition errors in the word transcription of a spoken document. Recognition errors in the spoken document transcription are often caused by two factors: mismatch between training and application environments; and words that are not included in the recognizer vocabulary. Recognition errors caused by mismatch between training and application environments can be handled by adding multiple ASR hypotheses into the spoken document transcription. Representation expansion enables the word-based spoken document retrieval methods to deal with out-of-vocabulary words. Various robust spoken document retrieval methods based on the document representations containing multiple recognition hypotheses were investigated. These were spoken document retrieval methods based on the N-best, lattice and word confusion network representation of a spoken document.

In Section 4.1, the spoken document retrieval methods based on the document representation including N-best recognition hypotheses were investigated. The results of this study indicate that the number of considered recognition hypotheses has a significant impact on the retrieval performance. It was verified that the recall rate increased with the growing number of recognition hypotheses considered. In our experiments, an improvement of about 5.05 % in recall was achieved, if we took $N = 20$ best hypotheses into account during the retrieval process. However, the best mean average precision (85.41) % was observed by $N = 9$, and dropped when $N > 9$. A comparison of different weighting scheme was made. The experimental results show that in comparison to the term-frequency weighting method, an improvement of 0.7 % in mean average precision was reached by the probability weighting method. This study verified that with the growing number of recognition hypotheses considered, more misrecognized terms are added to the spoken document representation. Consequently, the precision at a higher recall levels drops. Integrating the general probability of word appearance into the term weights improves the precision further.

The lattice transcription of a spoken document contains much more hypotheses than the N-best transcription. However, more misrecognized terms will remain in the document representation. In order to limit the search space, the lattice representation of a spoken document is pruned based on the duration-normalized log likelihood ratio DNLLR values of the links. Invalid links with a DNLLR value below the threshold will be ignored in the retrieval process. The results of this study verified the improvement in precision and recall achieved by the DNLLR-based lattice pruning, when the valid range of the DNLLR values are set to $[-118, -90]$. A valid link in the lattice can represent a misrecognized word. Using DNLLR-value directly as the term weighting may cause more confusion during the retrieval. The experimental results showed that many false alarms appear in the retrieved list when using DNLLR-value directly as the document term weights. In this case, weight document term with term frequency $tfidf$ performed better.

A word confusion network is the most compact representation of a lattice. The results of this study indicate that the total number of links in the document representation is reduced by about 76.5%. The maximum retrieval recall reaches 95.23%, which is comparable to the lattice-based methods. Retrieval precision reaches 64.5%. Using the word posterior probability directly as the weight of a document term improves the precision rate at lower recall levels. However, the mean average precision achieved with this weighting method drops for about 6.22% in comparison with the term-frequency *tfidf* weighting method. The experimental results verified that the precision at lower recall levels is improved with this new weighting method. The number of queries for which the relevant document is in the front rank of the retrieved list also increases. With this new weighting method we achieve a 63.71% mean average precision.

We made a comparison of the different robust word-based spoken document retrieval methods listed above. The experimental results indicate that the best retrieval precision is achieved by the 9-best based spoken document retrieval method. The maximal recall is yielded by the word confusion network based spoken document retrieval methods.

We carried out some additional research into representation expansion, which might enable the word-based spoken document retrieval methods to deal with out-of-vocabulary words. However, the experimental results showed that only slight improvement in retrieval performance can be achieved by the query expansion method proposed by Moreau [77]. As we stated before, the out-of-vocabulary words are a major problem that word-based spoken document retrieval systems have to deal with. The limited recognizer vocabulary restricts the range of the query. Replacing the original out-of-vocabulary query with the acoustically most similar one in a recognizer vocabulary (query expansion) diminishes this problem and enables word-based SDR doing retrieval of out-of-vocabulary query words. The experimental results verified that integrating the query expansion into the word-based spoken document retrieval methods can further improve the retrieval performance.

Chapter 5

Subword-based SDR Approaches

The performance of a spoken document retrieval system relies heavily on the quality of the spoken document transcription generated by an automatic speech recognizer. Errors in a spoken document transcription are often caused by the misrecognition problem and the out-of-recognition-vocabulary words in spoken information. In the last chapter, we thoroughly discussed how we deal with the recognition-errors in the spoken document transcription. In this chapter, we will explore the methods dealing with the recognition-errors caused by the out-of-recognition-vocabulary (OOV) words, the so-called OOV problem.

As we discussed before, the word-based spoken document retrieval method can work with the collection covering a large vocabulary. The word recognizer with very large vocabulary is required to index the diverse spoken information. There are two main drawbacks of this kind of spoken document retrieval methods. A huge amount of training data are needed to build a reliable large vocabulary continuous speech recognition system (LVCSR) for the indexing task. In addition, the size of the recognizable vocabulary is limited and fixed, which restricts the vocabulary of queries.

The size of vocabulary grows with the increasing number of available speech data. Even when we use very large vocabulary speech recognizer for the indexing task, there are still some new out-of-vocabulary words in the spoken document collection. This statement has been verified by Ng's experiments [84]. Most of the out-of-vocabulary words are proper names bearing important information.

We can solve the OOV problem by regularly updating the recognizer-vocabulary with new words. Every time we change the recognizer-vocabulary, the entire spoken information collection must be re-indexed again, which is a very computationally intensive task. Another way to handle the OOV-problem is the use of subword index-terms. These two advantages have convinced me to investigate the subword-based issues in the following sections:

- The number of indexing terms required to cover the language is dramatically reduced.
- It makes the indexing and retrieval processing independent from any word vocabulary. Virtually, open-vocabulary query retrieval is allowed.

This chapter is organized into five sections. First, Section 5.1 gives an overview of different subword units and evaluates their feasibility in spoken document retrieval. Then we focus on the shortest unit, the phones, and research different phone-based spoken document retrieval methods. The phone-transcription of a spoken document is generated using the methods introduced in section 5.2. Methods for retrieving information in phone-transcription are discussed in sections 5.3 and 5.4. Our research findings are summarized in Section 5.5.

5.1 Subword as Indexing Units

As mentioned in Section 2.2, different kinds of subword-units, such as morphemes, syllables or even phones, are used for speech-indexing task. Different subword-representation of the word 'performed' is shown in table 5.1.

Units	**Indexing Terms**
word	performed
phone	p er f ao r m d
phone sequence (n=3)	p_er p_er_f er_f_ao f_ao_r ao_r_m r_m_d m_d
broad phone class sequence (c=8,n=3)	$c_8\_c_3\_c_7$ $c_3\_c_7\_c_2$ $c_7\_c_2\_c_3$ ..
syllable (VCV)	pe erfo orme ed
morpheme	perform ed

Table 5.1: Examples of indexing terms for different units

Indexing a spoken document collection with subword units reduces the size of vocabulary. Consequently, the efficiency of spoken document retrieval is improved. From the data in Figure 5.1, we can visualize the reduction in vocabulary when we use a subword instead of a word as the index-unit. Figure 5.1 indicates how many unique subwords are required to represent a vocabulary of words. We have built word vocabularies with size from $1k$ to $10k$ on the Wikipedia word frequency list [3]. The Wikipedia word frequency list has been estimated from a collection of TV and movie scripts/transcripts downloaded from the Internet which includes $30M$ words. Vocabularies including $20k$ and $64k$ words have been estimated using the LDC's Gigaword text corpus [41].

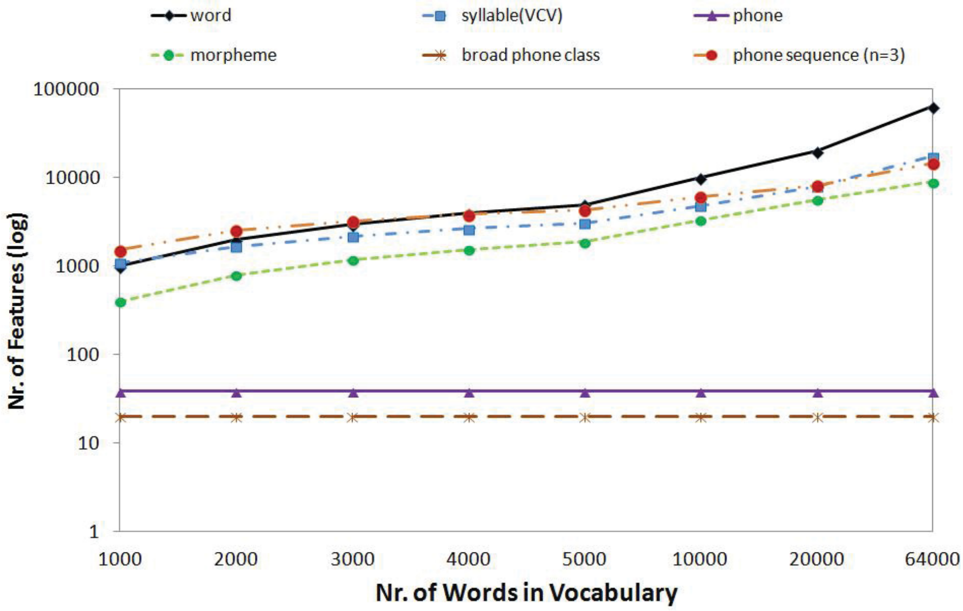


Figure 5.1: Number of unique terms for each type of index unit.

It can be observed that the number of unique phones or broad phone-classes required is small and fixed. When we use the subword-like syllable (vowel-consonant-vowel, VCV-features), phone-3gram or morpheme to represent the words in the vocabulary, the number of unique units required increases slightly with the growing number of words in the vocabulary. However, a clear reduction in vocabulary-size can be achieved by large-vocabulary indexing task ($64k$-vocabulary).

We gathered different subwords via word-segmentation. In order to avoid a short word to be split into too small segments, a minimal word-length is defined. In our case, the word-length is defined as the number of characters. The most frequent word-length will be selected as minimal word-length. We research the distribution of word-length on different collections with a varying number of words and summarize the experiment results in Figure 5.2. It can be viewed that the most frequent word-length varies from four to seven characters. In this work, the minimal word-length is set to four, which means that only words with more than four characters will be split into small fragments like vowel-consonant-vowel features, or morphemes.

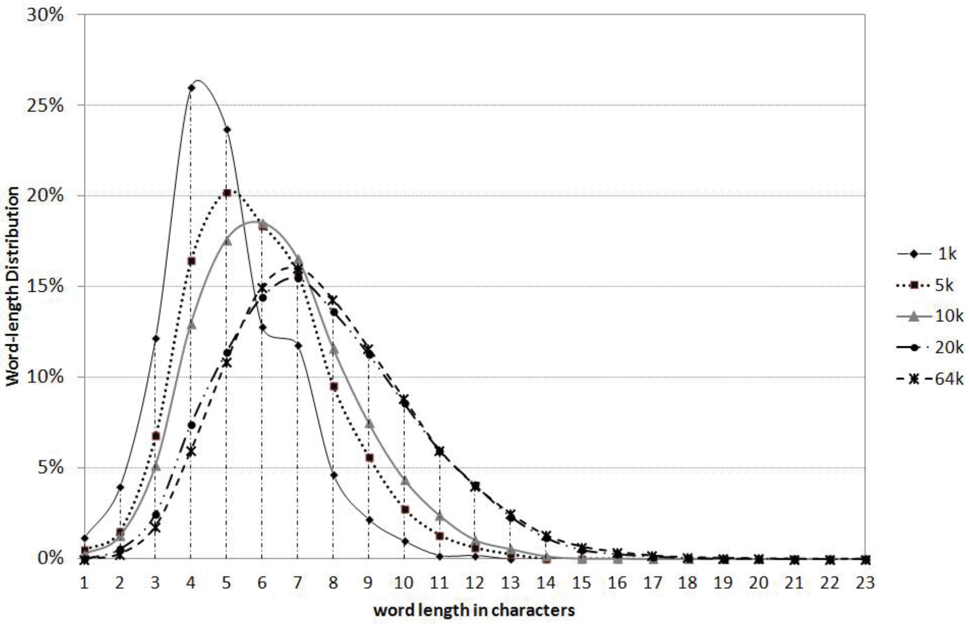


Figure 5.2: Word length distribution.

Extraction of Syllable Similar Index-units

Teufel [109] has proposed an automatic method for the extraction of vowel-consonant-vowel index-units. We have applied his method to our 64k vocabulary for syllable analyzing. Three special kinds of syllable-unit are selected: vowel-consonant-vowel(VCV), consonant-vowel (CV) and vowel-consonant (VC) unit. Generally, VC and CV units occur on the boundaries of a word. A CV feature can only appear at the beginning of a word, while VC features are always located at the end of a word. A cut is made at the position of a vowel in the middle of the word. For example, the word *communication* can be segmented into $co, ommu, uni, ica, atio, ion$.

Extraction of Morphemes

Morphemes are extracted using open-source software called Morfessor 1.0 [23]. Morfessor 1.0 is based on the algorithm proposed by Creutz and Lagus [22] [21] and has been developed for automatic morphological analysis. Based on a segmentation model learned previously, Morfessor can produce morpheme-like units (morphs) for each input word. A segmentation model consists of a morph lexicon and morph probabilities, and can be learned in an unsupervised manner from text. The Morfessor Toolkit can be used for learning a segmentation model and segmenting a word into morphemes. In our work, the segmentation model is learned from a vocabulary including $64k$ words.

The extraction of phone and phone N-grams will be discussed in more detail in section 5.2.

Experiment and Discussion

We have introduced a number of different subwords which can be applied for spoken information indexing and have described how to extract them from text. In this section, we carry out a feasibility study into using subword for the information retrieval task.

We run retrieval experiments on a collection of error-free text transcriptions of spoken documents ($WSJ, si-dt-s2$). Subword units are directly extracted from the ground-truth word annotation. We can imagine that the error-free text transcription has been produced by a perfect speech recognition system. The vector-space based retrieval model introduced in Section 2.3.2 is applied. Document terms are weighted by the method described in the following equation:

$$w(t_j, d) = \frac{n_{t_j,d}}{\sum_i n_{t_i,d}} \cdot log\frac{|D|}{|\{d : t_j \in d\}|} \tag{5.1}$$

where $w(t_j, d)$ represents the weight which will be assigned to the term t in document d; and $n_{t_j,d}$ is the occurrences of term t_j in document d, which is called term frequency. $\sum_i n_{t_i,d}$ is the total occurrences of all terms in the document d; $|D|$ indicates the number of documents in the collection and document frequency ($|\{d : t_j \in d\}|$) denotes the number of documents including the term t_j. This weighting method is called $tf-idf$, with SMART-Notation *ntu.ntu*. Query terms are simply weighted with '1'. This weighting method assigns more weight to the terms with high term frequency and low document frequency.

After eliminating common terms, there are still 1214 different content words left in the test collection ($WSJ, si-dt-s2$). These content words are selected as queries. The mean average precision (mAP), E1 (percent of queries with hit at first place in the ranked list) and E5 (mean reciprocal rank) are used to evaluate the retrieval performance of applying different kinds of subwords as index-units. Experimental results are shown in Figure 5.3 and Table 5.2.

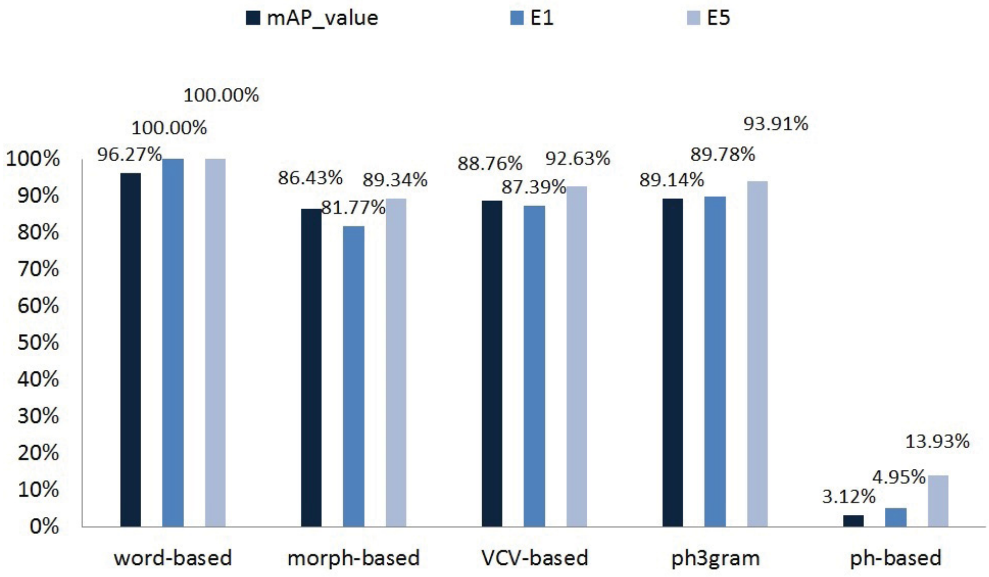


Figure 5.3: Retrieval performance of different index units derived from text reference transcription.

	word	morphem	syllable	phone-3gram	mono-phone
mAP	96.27	86.43	88.76	89.14	3.12
E1	100.00	81.77	87.39	89.78	4.95
E5	100.00	89.34	92.63	93.91	13.93

Table 5.2: Retrieval performance (mAP, E1 and E5 in %) of different index units (subword) on IR task.

The *syllables* used in our experiment are vowel-consonant-vowel features. The $phone - 3gram$ indicates that the index-unit is a sequence of three phones. It can be viewed that the word-based IR reaches the most mean average retrieval precision of 96%. The applied retrieval model is not based on the exact matching between query and document term, but on the vector-space based retrieval model that integrates inverse document frequency into the weighting scheme. Therefore the mean average precision of the word-based retrieval system cannot reach 100%.

We find out that longer subword units perform better than shorter ones. The syllable-based retrieval approach could reach a mean average retrieval precision of about 88.76%. Indexing spoken documents with syllable-like units could achieve better retrieval performance than that with the morpheme ($mAP = 86.43\%$). As shown in Figure 5.1, the number of unique syllables is more than that of morphemes, which means syllable-like units have better discriminative power than morpheme-like units. In addition, the syllable-like vowel-consonant-vowel terms are overlapped. The overlap between index-units enables more precise partial matching between the query and document. With the traditional vector-space retrieval model and $tfidf$ weighting, the mono-phone based SDR system can only achieve a mean average retrieval precision of about 3%. Only $4, 9\%$ of total queries have their answer at the first rank in the retrieved list. A significant improvement in retrieval performance was observed when using an overlapped phone-3gram instead of mono-phone for document indexing. Approximately 89% queries will have their answer at first rank in the retrieved list. This result reflects that longer units are useful for discriminating documents in the collection. Our experiment results have verified that indexing documents with subword units could achieve reliable retrieval performance.

In selection of an index-unit, we have to consider the information coverage achieved with the index-unit. However, the ability of an index-unit in precisely describing information is also a very important factor. A small index-unit like phones could provide very good information coverage but with poor precision. In this thesis, we are more concerned about the ability of an index-unit in information coverage. Therefore we will focus on the investigation of phone-based SDR approaches in this section.

Phone-based spoken document retrieval approaches must overcome the problem caused by the misrecognized terms in the transcription of a spoken document. After discussing how to estimate reasonable phone transcription of the speech segmentation, we explore some robust phone-based methods that attempt to compensate the errors in transcription by integrating ASR uncertainty into the scoring scheme. According to the applied retrieval model, the phone-based approaches can be roughly classified in two categories. These are:

- Phone 3-Gram based application using modified vector-space retrieval model.
- Probabilistic phone string matching-based spoken document retrieval.

5.2 Construction of Phone Transcription

The mono-phone transcription of a speech segment can be estimated in two ways: using a phone recognizer or post-processing of the word one-best transcription.

Applying Phone Recognizer

We built a phone recognizer on 16kHz mono audio speech data. 64 phones of TIMIT-alphabet are grouped into 39 phone-cluster to make the phone recognizer less sensitive to background noise. A context independent Left-to-right hidden Markov model (HMMs) with 128 Gaussians per state is used to model the acoustic features of a phone-cluster. The HTK toolkit [119] is used to train the set of phone models. The initial value of an acoustic model-parameter is estimated on the training set of the TIMIT corpus. The acoustic model is trained on the WSJ corpus. A phone-loop language model is estimated for the phone recognizer. The TIMIT and WSJ corpus were introduced in section 3.1.

Post-processing of Word One-best Transcription

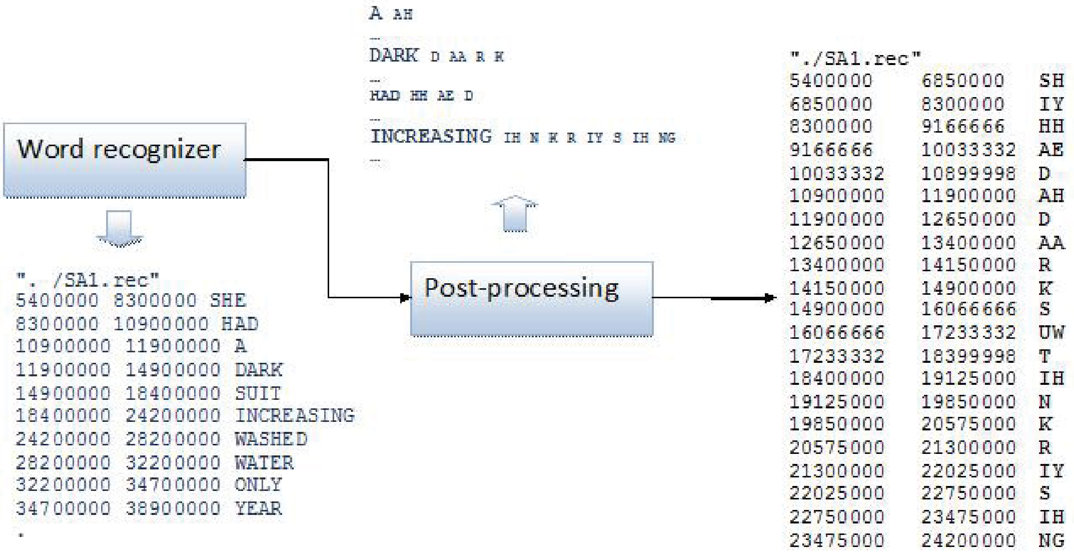
How to produce the monophone transcription of a spoken document directly from its word one-best recognition output can be viewed in Figure 5.4.

A Term in the word one-best transcription of a spoken document is first replaced with a sequence of phones according to a pronunciation dictionary. It is assumed that every phone has the same duration. Based on this assumption, the time marker of each phone can be computed as follows:

$$time_{start,ph} = time_{start,word} + (pos - 1) * Avg\_Duration_{ph} \tag{5.2}$$

$$time_{end,ph} = time_{start,ph} + Avg\_Duration_{ph} \tag{5.3}$$

where $time_{start,ph}$ and $time_{end,ph}$ indicate the start and end time mark of a phone; $time_{start,word}$ and $time_{end,word}$ are time marks of a word; and pos is the position of the considered phone in the phonetic representation of a word provided by a pronunciation



```
A AH
…
DARK D AA R K
…
HAD HH AE D
…
INCREASING IH N K R IY S IH NG
…
```

```
". /SA1.rec"
5400000 8300000 SHE
8300000 10900000 HAD
10900000 11900000 A
11900000 14900000 DARK
14900000 18400000 SUIT
18400000 24200000 INCREASING
24200000 28200000 WASHED
28200000 32200000 WATER
32200000 34700000 ONLY
34700000 38900000 YEAR
.
```

```
"./SA1.rec"
5400000    6850000    SH
6850000    8300000    IY
8300000    9166666    HH
9166666    10033332   AE
10033332   10899998   D
10900000   11900000   AH
11900000   12650000   D
12650000   13400000   AA
13400000   14150000   R
14150000   14900000   K
14900000   16066666   S
16066666   17233332   UW
17233332   18399998   T
18400000   19125000   IH
19125000   19850000   N
19850000   20575000   K
20575000   21300000   R
21300000   22025000   IY
22025000   22750000   S
22750000   23475000   IH
23475000   24200000   NG
```

Figure 5.4: Construction mono-phone transcription from word one-best output of the ASR.

dictionary; The average phone duration ($Avg\_Duration_{ph}$) is estimated as:

$$Avg\_Duration_{ph} = \frac{(time_{end,word} - time_{start,word})}{Nr\_of\_Phs} \tag{5.4}$$

where Nr_of_Phs is the number of phones in the phonetic representation of a word provided by a pronunciation dictionary.

Experiment and Discussion

As mentioned before, the performance of spoken document retrieval relies heavily on the quality of the transcription. In this section, we study the quality of the monophone transcriptions produced in two different ways. The test-set of the TIMIT corpus is selected for the evaluation task. The word one-best transcription of the document is produced by the word recognizer, introduced in section 3.2. The quality of the monophone transcription is judged by the rate of correctly recognized phones ($Correct$) and the accuracy ($Accuracy$). The experimental results are summarized in Table 5.3.

In Table 5.3, $CORR$ is the number of phones recognized correctly, DEL indicates

	CORR	DEL	SUB	INS	%Correct	%Accuracy
Phone recognizer	32915	11703	19652	3673	51.21	45.50
Estimation from Word one-best	36804	13255	14211	3624	57.26	51.63

Table 5.3: Quality of mono-phone transcription.

the number of phones spoken but not recognized, SUB represents the number of phones that are substituted in the transcription, and INS is the number of phones that appear in the transcription that are not spoken.

The experiment results confirm that the monophone transcription estimated from the word one-best transcription is of better quality. In comparison with the mononphone transcriptions produced by the phone recognizer, a rise in the rate of correctly recognized phones (ca. 6%) in the estimated monophone transcription can be viewed. The number of substitution and insertion errors are also reduced. The only advantage of applying the phone recognizer is that less spoken phones are missing in the transcription. Given these facts, we do further research on the estimated monophone transcriptions of a spoken document archive.

5.3 Robust SDR Approach based on Phone N-gram

This section focusses on the exploration of robust spoken document retrieval methods, which are based on transcription consisting of the overlapping phone N-grams. The vector space model introduced in Section 2.3.2 is modified for retrieving information in the phone N-gram transcription of a spoken document. The vector space model is originally applied for the retrieval of information in the error free documents. As the approximately matching between query terms and document terms can not be realized using the vector space model, misrecognized terms in the phone N-gram transcription of a spoken document will not be detected using the vector space model. Consequently, how to modify the vector space model for approximately phone N-gram matching has become the main focus of this section. Ng [84] has proposed some phone N-gram based spoken document retrieval methods. With the retrieval experiments on a collection of broadcasting news, he confirmed that:

- The retrieval performance is improved with an increasing length of phone-Ngram units. However, when the length of Ngram units is greater than 3, the terms become too specific and the retrieval performance drops. Phone-3gram performs best.

- Better performance is achieved when doing information retrieval in the transcription consisting of overlapped phone-Ngram. The reason is that the overlapped indexing terms provide more chances for partially matching between query and document terms.
- Indexing spoken document with broad phone classes could also provide enough information for effective retrieval.

Therefore, the following research in this section is based on the transcription of a spoken document that consists of overlapping phone-3grams that is produced by post-processing the estimated monophone transcriptions of spoken documents(as shown in Figure 5.5).

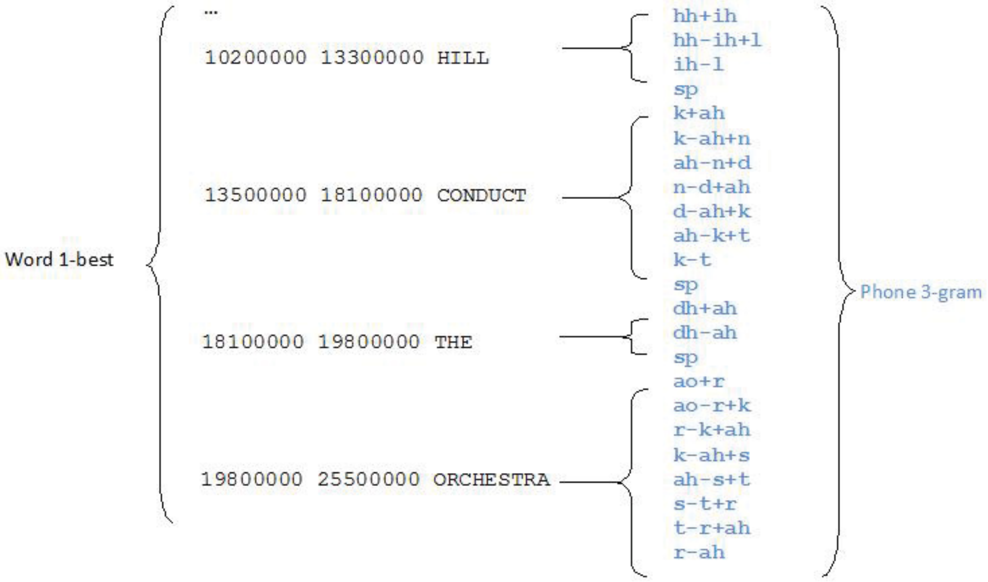


Figure 5.5: WSJ 4o0c0e0a word 1-best in phone-3gram.

5.3.1 Weighting with Confusion Information for Approximate Matching

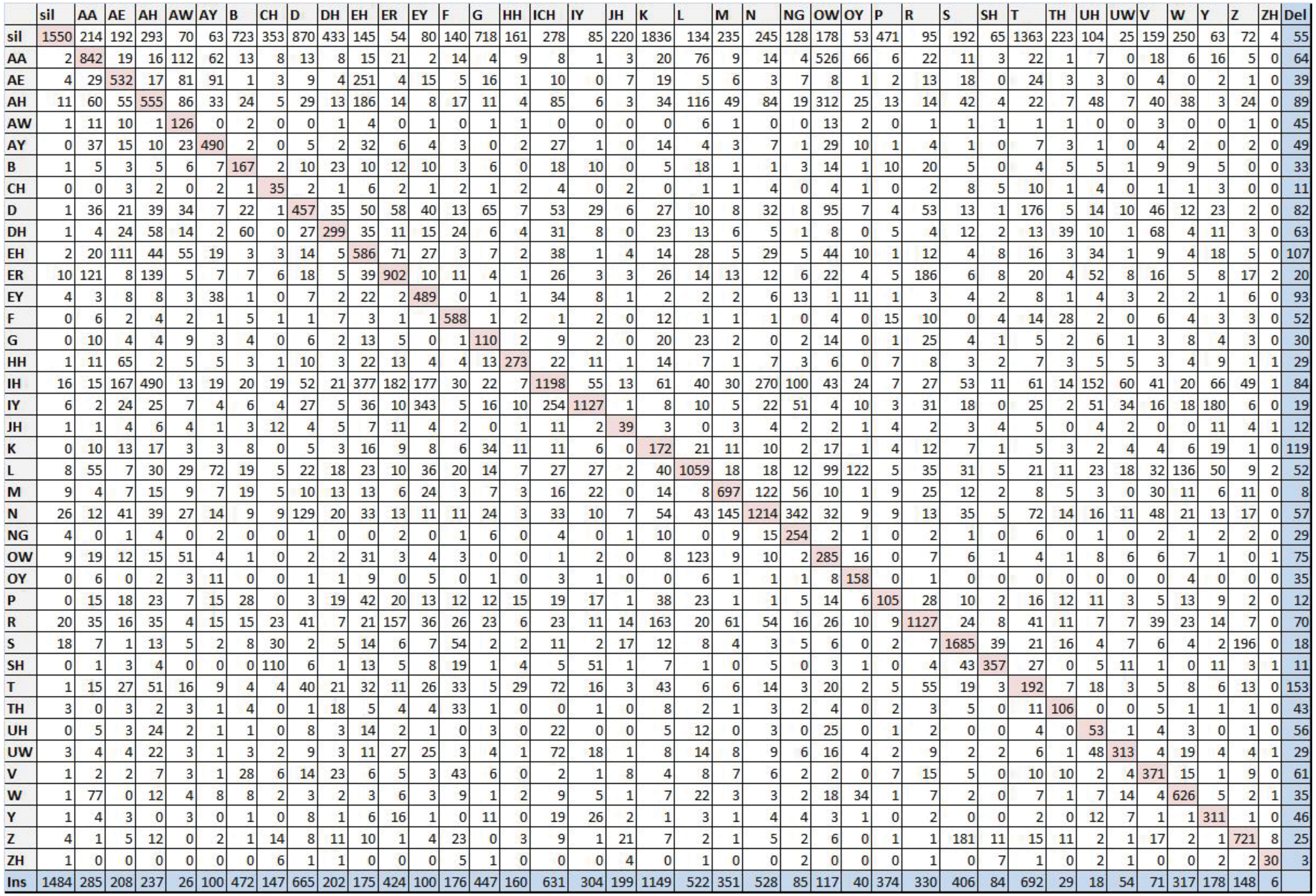
The approximate term-matching method proposed by Ng [83] considers all possible matches between the query and document terms. The phone confusion information provided by a phone recognizer is integrated into the document term weight. Chaudhari [13] expanded the method proposed by Ng with a kind of high order confusions that updated the phone confusion matrix with additional phone *bi*-gram, *tri*-gram and in general N-gram confusions. Only slightly improvement in precision and recall could be confirmed with Chaudhari's experiments. However, this kind of phone confusion matrix that includes high order confusions has very high dimensionality even only none zero components are stored in this matrix. This method is computationally very expensive. Srinivasan [108] proposed to formulate a probabilistic term weighting in a Bayesian framework. The Bayesian probabilistic model estimates a term weight based on phonetic confusions. In this work Ng's proposal will be further researched.

The phone confusion information of 39 English phone classes, including the silence provided by a phone recognizer, is summarized in Figure 5.6. **Phone confusion information** is the rate of substitution, insertion and deletion errors in the transcription of a spoken document produced by a phone recognizer. It is saved in a matrix; the so-called phone confusion matrix. The English phone confusion information is gathered from the TIMIT corpus. The dimension of our confusion matrix is 40×40. The upper left sub-matrix with dimension 39×39 includes all the statistics on substitution errors. In substitution sub-matrix, component $C(r, l)$ represents how many times a phone r is recognized as another phone l. The last column contains the number of deletion errors. The deletion error means that a phone is spoken but not recognized. The last row contains the number of insertion errors. The insertion error indicates how many times a phone is not spoken but recognized.

The similarity $sim(t', t)$ between query phone-3gram term t and document phone-3gram term t' is then expressed as:

$$sim(t', t) = DP(l_{t'}, l_t) \tag{5.5}$$

where $l_{t'}$ and l_t represent the length of the query term t and the document term t'. $l_{t'}$ and l_t are equal to 3 in case of phone-3gram. The similarity between two phone-3grams



	sil	AA	AE	AH	AW	AY	B	CH	D	DH	EH	ER	EY	F	G	HH	ICH	IY	JH	K	L	M	N	NG	OW	OY	P	R	S	SH	T	TH	UH	UW	V	W	Y	Z	ZH	Del
sil	1550	214	192	293	70	63	723	353	870	433	145	54	80	140	718	161	278	85	220	1836	134	235	245	128	178	53	471	95	192	65	1363	223	104	25	159	250	63	72	4	55
AA	2	842	19	16	112	62	13	8	13	8	15	21	2	14	4	9	8	1	3	20	76	9	14	4	526	66	6	22	11	3	22	1	7	0	18	6	16	5	0	64
AE	4	29	532	17	81	91	1	3	9	4	251	4	15	5	16	1	10	0	7	19	5	6	3	7	8	1	2	13	18	0	24	3	3	0	4	0	2	1	0	39
AH	11	60	55	555	86	33	24	5	29	13	186	14	8	17	11	4	85	6	3	34	116	49	84	19	312	25	13	14	42	4	22	7	48	7	40	38	3	24	0	89
AW	1	11	10	1	126	0	2	0	0	1	4	0	1	0	1	1	0	0	0	0	6	1	0	0	13	2	0	1	1	1	1	1	0	0	3	0	0	1	0	45
AY	0	37	15	10	23	490	2	0	5	2	32	6	4	3	0	2	27	1	0	14	4	3	7	1	29	10	1	4	1	0	7	3	1	0	4	2	0	2	0	49
B	1	5	3	5	6	7	167	2	10	23	10	12	10	3	6	0	18	10	0	5	18	1	1	3	14	1	10	20	5	0	4	5	5	1	9	9	5	0	0	33
CH	0	0	3	2	0	2	1	35	2	1	6	2	1	2	1	2	4	0	2	0	1	1	4	0	4	1	0	2	8	5	10	1	4	0	1	1	3	0	0	11
D	1	36	21	39	34	7	22	1	457	35	50	58	40	13	65	7	53	29	6	27	10	8	32	8	95	7	4	53	13	1	176	5	14	10	46	12	23	2	0	82
DH	1	4	24	58	14	2	60	0	27	299	35	11	15	24	6	4	31	8	0	23	13	6	5	1	8	0	5	4	12	2	13	39	10	1	68	4	11	3	0	63
EH	2	20	111	44	55	19	3	3	14	5	586	71	27	3	7	2	38	1	4	14	28	5	29	5	44	10	1	12	4	8	16	3	34	1	9	4	18	5	0	107
ER	10	121	8	139	5	7	7	6	18	5	39	902	10	11	4	1	26	3	3	26	14	13	12	6	22	4	5	186	6	8	20	4	52	8	16	5	8	17	2	20
EY	4	3	8	8	3	38	1	0	7	2	22	2	489	0	1	1	34	8	1	2	2	2	6	13	1	11	1	3	4	2	8	1	4	3	2	2	1	6	0	93
F	0	6	2	4	2	1	5	1	1	7	3	1	1	588	1	2	1	2	0	12	1	1	1	0	4	0	15	10	0	4	14	28	2	0	6	4	3	3	0	52
G	0	10	4	4	9	3	4	0	6	2	13	5	0	1	110	2	9	2	0	20	23	2	0	2	14	0	1	25	4	1	5	2	6	1	3	8	4	3	0	30
HH	1	11	65	2	5	5	3	1	10	3	22	13	4	4	13	273	22	11	1	14	7	1	7	3	6	0	7	8	3	2	7	3	5	5	3	4	9	1	1	29
IH	16	15	167	490	13	19	20	19	52	21	377	182	177	30	22	7	1198	55	13	61	40	30	270	100	43	24	7	27	53	11	61	14	152	60	41	20	66	49	1	84
IY	6	2	24	25	7	4	6	4	27	5	36	10	343	5	16	10	254	1127	1	8	10	5	22	51	4	10	3	31	18	0	25	2	51	34	16	18	180	6	0	19
JH	1	1	4	6	4	1	3	12	4	5	7	11	4	2	0	1	11	2	39	3	0	3	4	2	2	1	4	2	3	4	5	0	4	2	0	0	11	4	1	12
K	0	10	13	17	3	3	8	0	5	3	16	9	8	6	34	11	11	6	0	172	21	11	10	2	17	1	4	12	7	1	5	3	2	4	4	6	19	1	0	119
L	8	55	7	30	29	72	19	5	22	18	23	10	36	20	14	7	27	27	2	40	1059	18	18	12	99	122	5	35	31	5	21	11	23	18	32	136	50	9	2	52
M	9	4	7	15	9	7	19	5	10	13	13	6	24	3	7	3	16	22	0	14	8	697	122	56	10	1	9	25	12	2	8	5	3	0	30	11	6	11	0	8
N	26	12	41	39	27	14	9	9	129	20	33	13	11	11	24	3	33	10	7	54	43	145	1214	342	32	9	9	13	35	5	72	14	16	11	48	21	13	17	0	57
NG	4	0	1	4	0	2	0	0	1	0	0	2	0	1	6	0	4	0	1	10	0	9	15	254	2	1	0	2	1	0	6	0	1	0	2	1	2	2	0	29
OW	9	19	12	15	51	4	1	0	2	2	31	3	4	3	0	0	1	2	0	8	123	9	10	2	285	16	0	7	6	1	4	1	8	6	6	7	1	0	1	75
OY	0	6	0	2	3	11	0	0	1	1	9	0	5	0	1	0	3	1	0	0	6	1	1	1	8	158	0	1	0	0	0	0	0	0	0	4	0	0	0	35
P	0	15	18	23	7	15	28	0	3	19	42	20	13	12	12	15	19	17	1	38	23	1	1	5	14	6	105	28	10	2	16	12	11	3	5	13	9	2	0	12
R	20	35	16	35	4	15	15	23	41	7	21	157	36	26	23	6	23	11	14	163	20	61	54	16	26	10	9	1127	24	8	41	11	7	7	39	23	14	7	0	70
S	18	7	1	13	5	2	8	30	2	5	14	6	7	54	2	2	11	2	17	12	8	4	3	5	6	0	2	7	1685	39	21	16	4	7	6	4	2	196	0	18
SH	0	1	3	4	0	0	0	110	6	1	13	5	8	19	1	4	5	51	1	7	1	0	5	0	3	1	0	4	43	357	27	0	5	11	1	0	11	3	1	11
T	1	15	27	51	16	9	4	4	40	21	32	11	26	33	5	29	72	16	3	43	6	6	14	3	20	2	5	55	19	3	192	7	18	3	5	8	6	13	0	153
TH	3	0	3	2	3	1	4	0	1	18	5	4	4	33	1	0	0	1	0	8	2	1	3	2	4	0	2	3	5	0	11	106	0	0	5	1	1	1	0	43
UH	0	5	3	24	2	1	1	0	8	3	14	2	1	0	3	0	22	0	0	5	12	0	3	0	25	0	1	2	0	0	4	0	53	1	4	3	0	1	0	56
UW	3	4	4	22	3	1	3	2	9	3	11	27	25	3	4	1	72	18	1	8	14	8	9	6	16	4	2	9	2	2	6	1	48	313	4	19	4	4	1	29
V	1	2	2	7	3	1	28	6	14	23	6	5	3	43	6	0	2	1	8	4	8	7	6	2	2	0	7	15	5	0	10	10	2	4	371	15	1	9	0	61
W	1	77	0	12	4	8	8	2	3	2	3	6	3	9	1	2	9	5	1	7	22	3	3	2	18	34	1	7	2	0	7	1	7	14	4	626	5	2	1	35
Y	1	4	3	0	3	0	1	0	8	1	6	16	1	0	11	0	19	26	2	1	3	1	4	4	3	1	0	2	0	0	2	0	12	7	1	1	311	1	0	46
Z	4	1	5	12	0	2	1	14	8	11	10	1	4	23	0	3	9	1	21	7	2	1	5	2	6	0	1	1	181	11	15	11	2	1	17	2	1	721	8	25
ZH	1	0	0	0	0	0	0	6	1	1	0	0	0	5	1	0	0	0	4	0	1	0	0	2	0	0	0	1	0	7	1	0	2	1	0	0	2	2	30	3
Ins	1484	285	208	237	26	100	472	147	665	202	175	424	100	176	447	160	631	304	199	1149	522	351	528	85	117	40	374	330	406	84	692	29	18	54	71	317	178	148	6	

Figure 5.6: 39 English phone confusion matrix.

$DP(m,n)$ is computed recursively as:

$$DP(m,n) = \begin{cases} 1, & m = 0; n = 0 \\ DP(0, n-1) \cdot Ins(t[n-1]), & m = 0, n > 0 \\ DP(m-1, 0) \cdot Del(t'[m-1]), & m > 0, n = 0 \\ max \begin{cases} DP(m-1, n) \cdot Del(t'[m-1]) \\ DP(m-1, n-1) \cdot Sub(t'[m-1], t[n-1]) \\ DP(m, n-1) \cdot Ins(t[n-1]) \end{cases} & m > 0, n > 0 \end{cases} \tag{5.6}$$

where $Ins(t[n-1])$ indicates the insertion error rate of phone $t[n-1]$; $Del(t'[m-1])$ is the deletion error rate of phone $t'[m-1]$; $Sub(t'[m-1], t[n-1])$ indicates the rate that phone $t'[m-1]$ is substituted by the phone $t[n-1]$; The computation of $DP(m,n)$ is terminated when m reaches $l_{t'}$ and n reaches l_t. As we announced before, the rate of substitution, insertion and deletion errors is summarized in the phone confusion information matrix (5.6, [56]).

The similarity defined by Equation 5.5 is estimated for each term t' in a document, given query term t. The relevance score of the spoken document d to the given query q is then simply expressed by the following equation:

$$Sim(d,q) = \sum_{t \in q} \sum_{t' \in d} sim(t', t). \tag{5.7}$$

We can imagine that the spoken content retrieval task is very computationally intensive in this approach, as it handles all potential matches between document phone-3gram terms and phone-3gram terms in the query.

5.3.2 Integrating Position Information into Indexing and Relevance Ranking

The spoken document retrieval methods that we have discussed so far are based on the assumption that the query terms are independent from one another. The proximity information of a term is ignored. Two documents, including the same query terms, are scored equally. However, in this case, the document in which query terms are close to each other should be assigned with a higher relevance-score, which can be achieved by incorporating the term proximity information into the relevance score. The position information of a term in a document is crucial to the evaluation of the proximity information. This section focusses on the investigation of how to incorporate the term position information into the phone-3gram based spoken document scoring method.

Chelba has successfully incorporated the position information into his spoken document retrieval approach. He proposed to retrieve information on the word-level position-specific posterior lattice (PSPL) [14] of a spoken document. The position-specific posterior lattice is a compact representation of a speech recognition lattice, which contains the position as well as other contextual information for each link. Chelba made some retrieval experiments on a collection of lecture records (iCampus [37]), and verified the advantages of incorporating the position information into relevance score. His idea about the position specific posterior lattice was extended by Pan and applied to the Chinese retrieval task [88]. A Chinese word often consists of one or two characters. In Pan's application, a position specific posterior lattice was built for a spoken document at the character-level. Pan's experiments confirmed an improvement in the retrieval precision (mean average precision) of about 55%.

Position Information in Phone-3gram One-best Transcription

In our work, we extend the position specified posterior lattice to the phone-3gram level. In order to incorporate the proximity information of each triphone term into the ranking-score of a spoken document, we attempt to match a subsequence of N phone-3grams in the document transcription.

After the indexing process, each phone-3gram is represented as $(t, pos, posterior)$. $Posterior$ is the probability of the query phone-3gram term t appearing at position pos. Given query $Q = t_1, t_2, ...t_q$, the probability $P(t_{i,Q}, D)$ of a given query term $t_{i,Q}$ in the document D is computed as follows:

$$P(t_{i,Q}, D) = log(1 + \sum_{k=1}^{d} Prob_{k,D}(t_{k,D} = t_{i,Q})) \tag{5.8}$$

where k denotes the position in document D. There are in total d phone-3grams in this document; $Prob_{k,D}(t_{k,D} = t_{i,Q})$ denotes the probability of the query phone-3gram $t_{i,Q}$ appearing at position k in document D. In our case, $Prob_{k,D}(t_{k,D} = t_{i,Q})$ is estimated as:

$$Prob_{k,D}(t_{k,D} = t_{i,Q}) = \frac{d}{OCC(t_{i,Q}, D)} \tag{5.9}$$

where $OCC(t_{i,Q}, D)$ denotes the frequency of term $t_{i,Q}$, in document D containing d phone-3grams.

In order to discount the effect of large occurrences of some of the phone-3grams in a given document, logarithmic tapering off of the expected occurrences is applied. Single phone-3gram matching score is expressed by:

$$Score_1(D, Q) = \sum_{i=1}^{Q} P(t_{i,Q}, D) \tag{5.10}$$

The matching score of a subsequence of N phone-3grams is estimated as:

$$Score_N(D, Q) = \sum_{i=1}^{q-N+1} log[1 + \sum_{k=1}^{d} \prod_{n=1}^{N-1} Prob_{k+n,D}(t_{k+n} = t_{i+n,Q})] \tag{5.11}$$

The relevance or similarity between a document D and a given query $Q = t_1, t_2, ...t_q$ is estimated as:

$$Sim(D, Q) = \sum_{n=1}^{q} \omega_n \cdot Score_n(D, Q) \tag{5.12}$$

where the weight ω_n should increase with the growing n and can be assigned in two different ways:

$$\omega_n = \alpha \cdot n; \text{or } \omega_n = e^n \tag{5.13}$$

Second exponential weight is applied in the following experiments.

The Phone-3gram Position Specific Posterior Lattice (PSPL)

In the last section, we adapted PSPL ranking scheme to a document representation consisting of phone-3grams that is derived from the word one-best transcription. In this section, we focus on the estimation of the position-specific posterior lattice (PSPL) at the phone-3gram level and the corresponding scoring method incorporating the proximity information. Table 5.4 shows a part of phone-3gram PLPS estimated for the file 410a0101 in the WSJ-Corpus. A Table entry consists of a phone-3gram with its posteriors in parenthesis. The estimation of posteriors of a phone-3gram is showed in Equation 5.14.

...	n	+1	+2	+3	+4	+5	+6	...
...	ah-z (1.0)	ah (0.7)	sp(0.7)	ey (0.7)	sp(0.7)	l+eh(0.7)	l-eh+t(0.7)	...
...		l+ae(0.1)	l-ae+t(0.1)	ae-t+ah(0.1)	t-ah+t (0.1)	ah-t+uw(0.1)	t-uw+d (0.1)	...
...		l+eh(0.2)	l-eh+t(0.2)	eh-t+er (0.2)	t-er (0.2)	sp(0.2)	ah+v(0.2)	...

Table 5.4: Part of phone-3gram-PSPL for file 4l0a0101.

The position-specific posterior lattice at the phone-3gram level is derived from the word N-best transcription of a spoken document in two steps. First, the word N-best transcription is transcribed into phone-3gram 'N'-best representation. Then the phone-3grams in 'N'-best transcription are ordered according to their position in the transcription and are represented as $(t, pos, posterior)$. The posterior of a phone-3gram term $t_{i,Q}$ at position k in document D is computed as follows:

$$Posterior_D(t_{i,Q}, k) = \frac{OCC_D(t_{i,Q}, k)}{N} \tag{5.14}$$

where $OCC_D(t_{i,Q}, k)$ indicates how many times the term $t_{i,Q}$ occurs at position k in a phone-3gram position-specific posterior lattice. N is the number of considered word hypotheses. This posteriors is used for estimating the probability $P(t_{i,Q}, D)$ of a given query term $t_{i,Q}$ in the document D and is computed as:

$$P(t_{i,Q}, D) = log(1 + \sum_{k=1}^{d} Posterior_D(t_{i,Q}, k)) \tag{5.15}$$

where k denotes the position in document D. There are in total d phone-3grams in this document; $Posterior_D(t_{i,Q}, k)$, as denotes the posteriors of the query phone-3gram $t_{i,Q}$ appearing at position k in document D.

Single phone-3gram matching score is expressed by:

$$Score_1(D, Q) = \sum_{i=1}^{Q} P(t_{i,Q}, D) \tag{5.16}$$

The matching score of a subsequence of N phone-3grams is then estimated as:

$$Score_N(D,Q) = \sum_{i=1}^{q-N+1} log[1 + \sum_{k=1}^{d} \prod_{n=1}^{N-1} Posterior_D(t_{i,Q},k)] \tag{5.17}$$

The relevance or similarity between a document D and a given query $Q = t_1, t_2, ...t_q$ is estimated as:

$$Sim(D,Q) = \sum_{n=1}^{q} \omega_n \cdot Score_n(D,Q) \tag{5.18}$$

where the exponential weight ω_n should increase with the growing n.

5.3.3 Experiment and Discussion

In this section, we are going to evaluate the performance of different phone-3gram based spoken document retrieval methods that we have discussed before. We first compare their performance by the in-vocabulary query retrieval task and then we test their ability to deal with out-of-vocabulary words. Their performance on the total query set will also be evaluated. We study the advantages that are achieved by incorporating the position information into the scoring-scheme. In addition, we will do some experiments to research the effect of query-length on retrieval performance.

Retrieving In-vocabulary Query-words

There are 200 in-vocabulary queries that are selected for the evaluation task. The experimental results are summarized in Figure 5.7 and Table 5.5.

In Figure 5.7 and Table 5.5, the symbol $baseline$ represents the performance of our baseline spoken document retrieval system, based on the word one-best transcription. This baseline system was described in section 3.5. $tfidf$ represents the phone-3gram based spoken document retrieval approach used to estimate the ranking-score based on the classical term-frequency ($tfidf$). The approach used to integrate the phone confusion information into the scoring scheme is noted as $confusion\_info\_sum$ (see the discussion in Section 5.3.1). $proximity\_weighting$ denotes the approach discussed in section 5.3.2. These symbols can also be found in the following figures and tables.

In comparison with word one-best transcription based spoken document retrieval

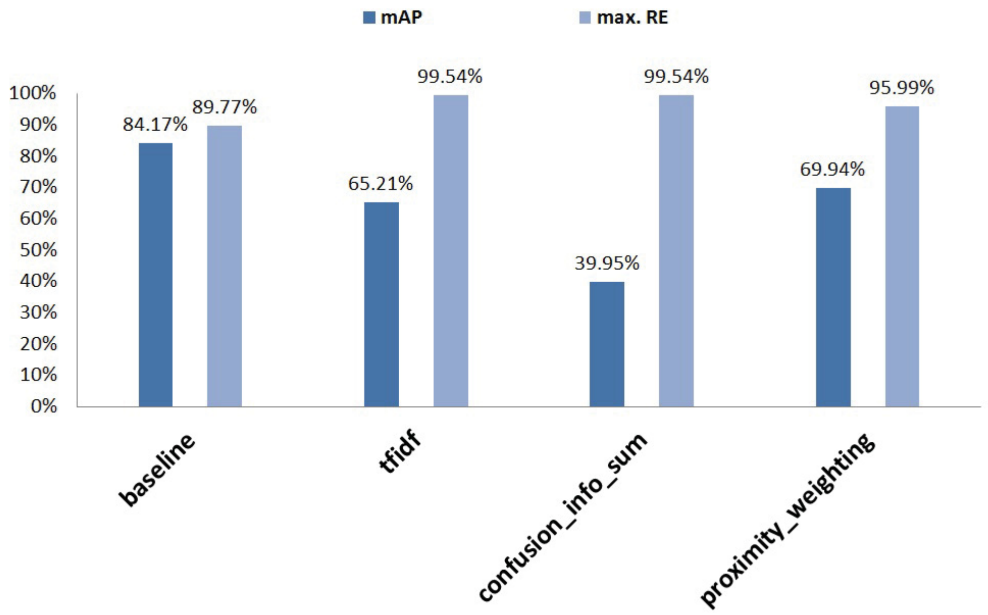


Figure 5.7: Retrieval performance of different phone-3gram based methods (in-vocabulary query words).

approaches, the phone-3gram based approaches show their superiority over maximum achievable recall rate. The phone-3gram application with classical term-frequency ($tf-idf$) based scoring scheme can reach a maximum recall of about 99.5%. However, decrease in the mean average precision value can also be observed. Incorporating the proximity information into the document relevance score improves the mean average precision rate of a phone-3gram based spoken document retrieval system. In Figure 5.7, we can see that the *proximity_weighting*-method achieves a 69.94% mean average precision rate while keeping a 95.99% maximum recall rate. We discover that weighting a document term directly with the phone confusion information cannot improve the retrieval performance; on the contrary, a clear drop in mean average precision rate can be observed.

Our experiments verified that very good information coverage can be achieved by using the phone-3gram as the index unit. In comparison with the word-based spoken document retrieval application, the phone-3gram based spoken document retrieval approach achieves a better recall rate. The maximum achievable retrieval recall rate reaches 99.54%.

		phone-3gram		
recall	baseline	tfidf	confusion_info_sum	proximity_weighting
0	86.00	79.00	58.50	84.50
10	89.12	77.80	53.15	81.30
20	89.93	77.90	51.78	82.22
30	90.12	77.07	50.14	81.75
40	89.59	75.47	49.09	81.05
50	89.05	73.30	45.35	80.40
60	88.48	70.29	39.51	78.09
70	88.10	63.96	32.10	71.49
80	87.30	51.73	23.79	56.98
90	86.98	31.22	13.35	27.91
mAP	84.17	65.21	39.95	69.94
max.RE	89.77	99.54	99.54	95.99

Table 5.5: Retrieval performance in precision/Recall, mAP, max. Recall value (%) of in-vocabulary query evaluation.

The Ability to Deal with OOV Queries

In the last section, we confirm the improvement in the recall rate achieved by using the phone-3gram as the index unit for the spoken document retrieval task. We are now going to research the ability of the phone-3gram index unit to deal with out-of-vocabulary words.

We chose 30 out-of-vocabulary words as the queries. Different spoken document retrieval approaches, based on the phone-3gram index unit, are evaluated. The experimental results are gathered in Figure 5.8 and Table 5.6.

In Figure 5.8, it can be viewed that the phone-3gram based spoken document retrieval methods may deal with out-of-vocabulary query words with poor retrieval performance. A high maximum recall rate is reached (92.83%). However, this means a significant loss in mean average retrieval precision value. The highest precision (ca. 16%) is reached by proximity-weighting at recall level of 10%. A poor mean average retrieval precision (0.7%) is observed by the classical *tfidf* scoring method. The phone-3gram based spoken document retrieval method that integrates the confusion information into the relevance score (the red line in Figure 5.8) performs slightly better than the method which forms the relevance score with the classical term-frequency weight. The phone-3gram based spoken document retrieval method that incorporates the proximity information into the relevance score achieves the best performance for

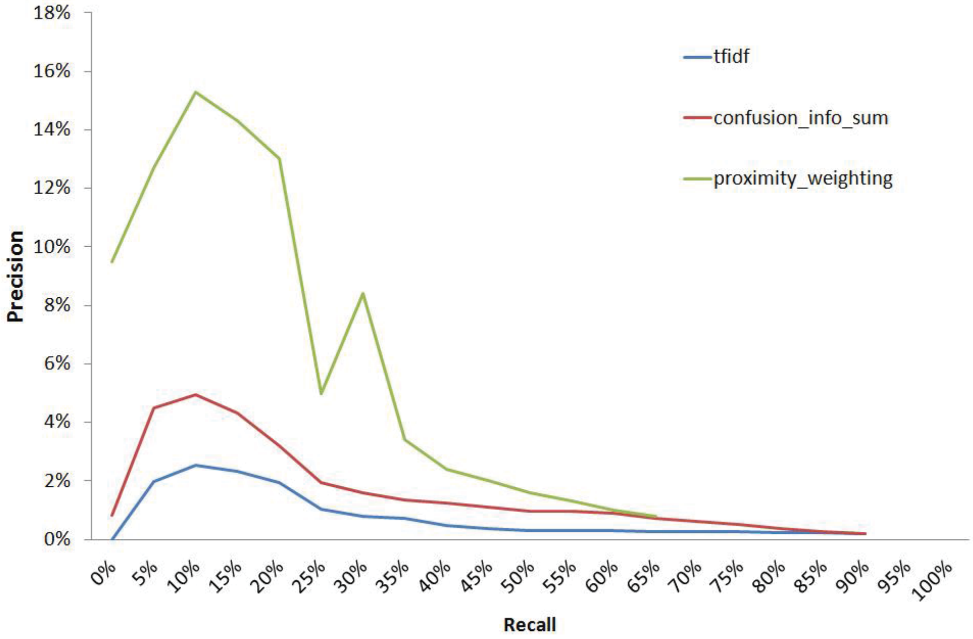


Figure 5.8: Performance of out-of-vocabulary query retrieval with different phone-3gram based spoken document retrieval methods.

the out-of-vocabulary query retrieval task. Even with a decreased maximal recall rate of 68.85%, the mean average precision rate reaches 4.54%. The results of this experiment indicate that the phone-3gram based spoken document retrieval method can hardly work with the out-of-vocabulary words. This is not a good choice for the out-of-vocabulary words retrieval task because of its low retrieval precision rate.

recall	0	10	20	30	40	50	60	70	80	90	mAP	max. RE
tfidf	0	2.55	1.94	0.80	0.48	0.32	0.29	0.28	0.24	0.20	0.73	92.83
conf_info_sum	0.83	4.94	3.20	1.58	1.23	0.98	0.89	0.63	0.38	0.21	1.53	92.74
proximity_weighting	9.50	15.30	13.00	8.40	2.40	1.60	1.00				4.54	68.85

Table 5.6: Performance of Phone-3-gram based SDR approaches for OOV queries retrieval task.

Performance Evaluation in Real Application

As we stated before (Section 3.3), a query set containing 200 in-vocabulary words and 30 out-of-vocabulary words is used to simulate the real application case. We compare the retrieval performance of different phone-Ngram based methods on this query set. The results obtained are shown in Table 5.7 and Figure 5.9(a).

		Phone 3-gram		
recall	baseline	tfidf	confusion_info_sum	proximity_weight
0	74.78	68.99	50.86	77.38
10	77.60	67.99	46.31	75.06
20	78.08	68.05	45.55	75.70
30	78.18	66.91	44.03	75.04
40	77.70	65.20	42.57	74.17
50	77.08	62.98	37.52	73.02
60	76.66	57.98	31.31	69.60
70	75.89	47.89	23.68	60.16
80	-	31.09	15.37	42.21
90	-	-	2.20	9.74
mAP	61.67	52.18	32.60	61.08
max. RE	78.06	89.04	98.65	93.40

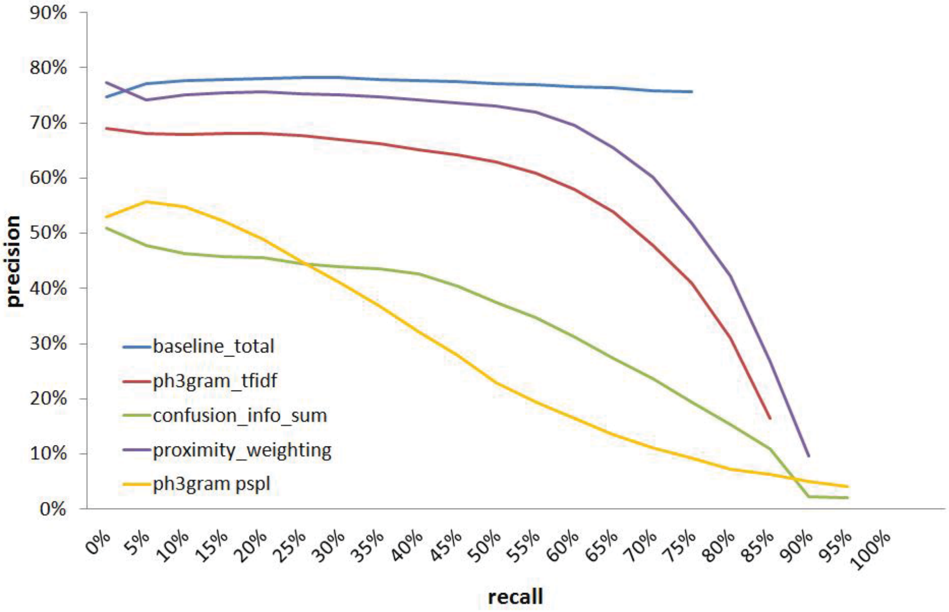
Table 5.7: Precision/Recall, mAP, max. Recall value (%) of evaluation with total query set.

In comparison with the word-based baseline system, a clear improvement in recall rate is achieved using the phone-3gram based spoken document retrieval methods. The maximum retrieval recall ($max.RE = 98.65\%$) is yielded when the phone confusion information is integrated into the document ranking score. If we compare the results, it can be seen that integrating the proximity information into the document ranking-score improves the precision rate at low recall levels. This method improves the maximal recall rate by about 15.34, in comparison with the baseline system, while keeping a 61.08% mean average retrieval precision.

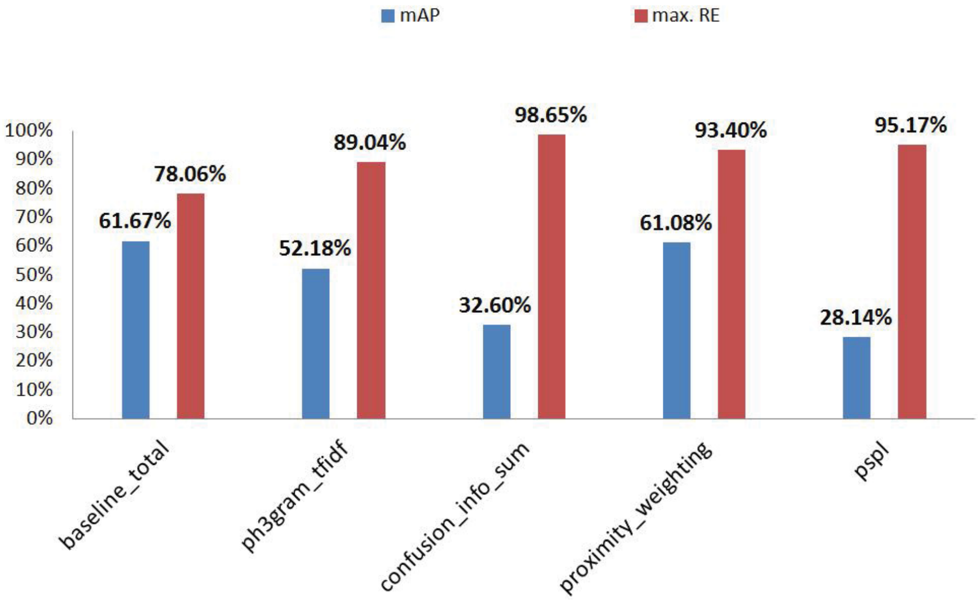
Performance of the Phone-3gram PSPL based Method

In Section 5.3.2, in the part (The Phone 3-gram position specific posterior lattice PSPL) we successfully extend the position specific posterior lattice to the phone-3gram level. The performance of a spoken document retrieval system based on the phone-3gram PSPL is evaluated in this section. Figure 5.9(a), 5.9(b) and Table 5.8 presents the experimental results.

A clear benefit of applying the phone-3gram PSPL could not be identified in this analysis. The phone-3gram PSPL based spoken document retrieval method performs slightly better than the $confusion\_info\_sum$ method listed in Table 5.7 at a lower retrieval recall level. However, the mean average retrieval precision decreases by a further 4.46%.



(a) Precision/recall plot



(b) mAP & max. RE value

Figure 5.9: Evaluation of different phone-3gram based methods with total query set.

recall	INV	OOV	total
0	59.50	10.00	53.04
10	62.55	4.00	54.78
20	56.58	2.82	48.87
30	48.84	1.60	41.07
40	40.01	1.18	32.21
50	30.21	0.89	22.90
60	21.78	0.78	16.45
70	14.70	0.56	11.11
80	10.15	-	7.27
90	6.89	-	4.97
mAP	33.63	1.83	28.14
max. RE	97.56	79.18	95.17

Table 5.8: Performance of phone-3gram PSPL based SDR method.

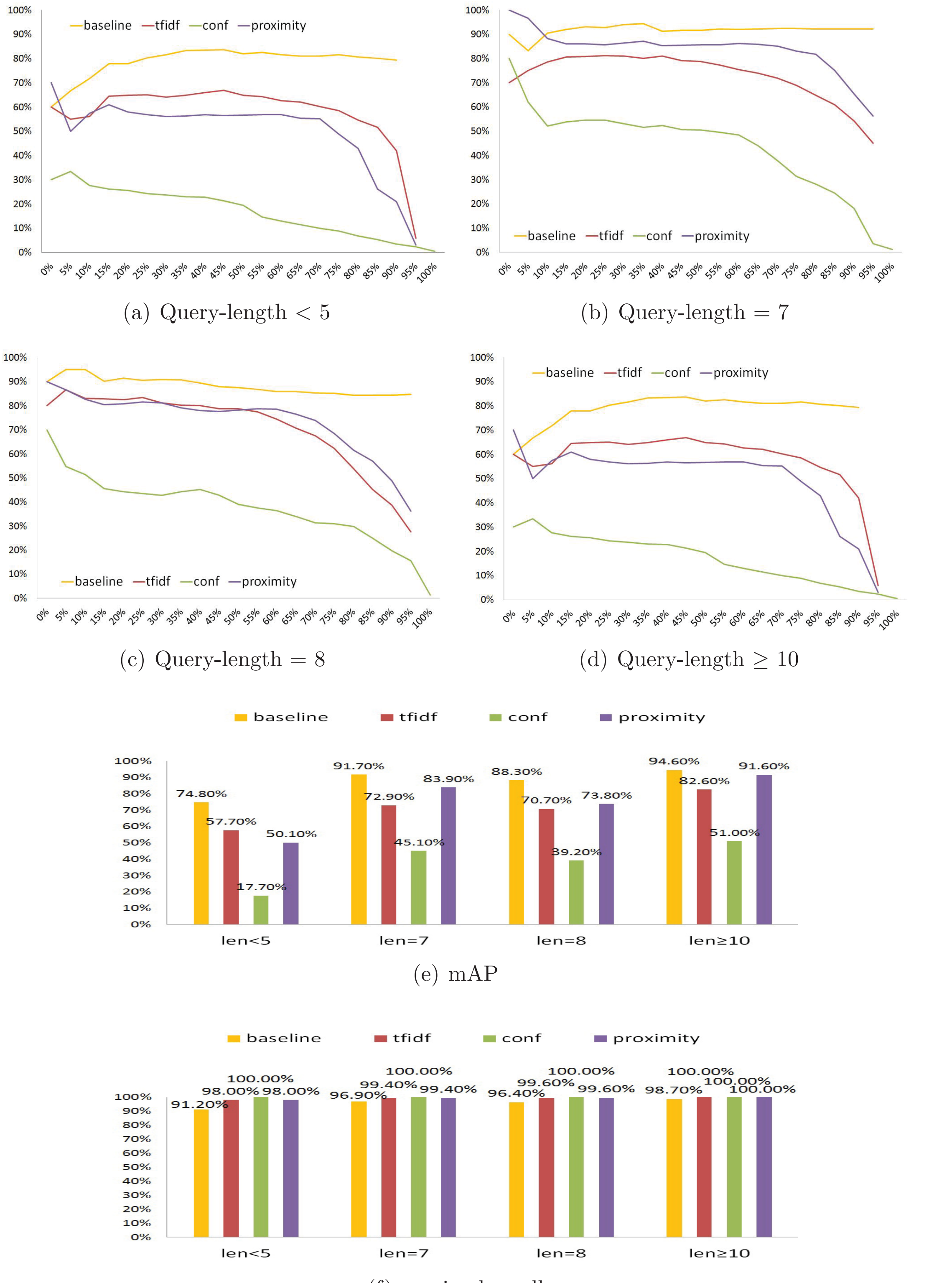
The Effect of Query-length on Retrieval Performance

This set of analyses examines the impact of query-length on the retrieval performance of phone-3gram based methods. Here, the query-length indicates the number of the phones. The experimental results in section 5.3.3 show that applying the phone-3gram PSPL method can not benefit the retrieval performance. Therefore the research on the impact of query-length on this method is not conducted in this section. As shown in Table 5.9, the query set in this evaluation task consists of 40 in-vocabulary (INV) words and 40 out-of-vocabulary (OOV) words, with different length (*len*). Figure 5.10 and Table 5.17 present the obtained results.

INV				OOV			
len < 5	len = 7	len = 8	len >= 10	len < 5	len = 7	len = 8	len $\geq$ 10
Certain	Capital	Agreement	Administration	Goats	Radiator	Clinton’s	Continents
Costs	Congress	American	Environmental	Shaft	Workforce	Mutation	Humankind
Foreign	December	Companies	Executives	Herding	Teenagers	Bismarck	Spokesperson
Funds	Economy	Exchange	Development	Oceans	Appetizers	Eclipses	Polystyrene
Meeting	Increased	Government	Investment	Koresh	Automakers	Intervals	Expectancy
Stocks	Markets	Military	Incorporated	Rodham	Firearms	Mechanized	Microbiology
Total	Program	Proposal	International	Hiliary	Infinite	Implants	Spokespeople
Value	Position	Spokesman	California	Spiller	Letterman	Hoffenberg	Cardiologists
Women	United	Students	Department	Biking	Malaria	Killington	Immunizations
Better	Yesterday	Recently	Securities	Canine	Somalia	Cheerleading	Macdonals

Table 5.9: Selected INV and OOV words with different length in number of phones.

The label *baseline* is used to represent our baseline word-based spoken document



(a) Query-length < 5

(b) Query-length = 7

(c) Query-length = 8

(d) Query-length ≥ 10

(e) mAP

(f) maximal recall

Figure 5.10: Performance (precision/recall plot) of in-vocabulary queries with different lengths.

retrieval system. $tfidf$ represents the phone-3gram based spoken document retrieval systems that use term frequency to form the document ranking-score. $conf$ indicates the phone-3gram based retrieval method that integrates the phone confusion information into the document ranking-score. The phone-3gram based retrieval method, which takes the proximity information into account while computing the document ranking score, is represented by the label *proximity*.

In comparison with the baseline system, the *proximity* method significantly improves the maximum recall rate. It can be viewed in Table 5.12, retrieving selected long queries with the *proximity* method reaches a recall rate of 100% also in case of retrieving queries *Environmental*, *Incorporated*, $California$ and *Department*. Retrieving those four queries with the *baseline* method yields a worse recall rate less than 100%. The $conf$ and $tfidf$ methods achieve also a recall rate of 100% in retrieving all selected long queries. Average retrieval performance summarized in Table 5.17 has showed that the *proximity* method yields better mAP value in comparison with the $conf$ and $tfidf$ methods. It can be observed that the *proximity* method outperforms its rivals in the long queries retrieval task (as showed in Figure5.17(f)). We assume that the recall and mAP achieved in retrieving queries with the same length are normally distributed. Therefore, the normally distributed t-test statistic is applied to check whether this statement is statistically significant. We break our statement into 3 sub-statements.

- The *proximity* method reaches better recall than the *baseline* method
- The *proximity* method yields better mAP than the $conf$ method
- The *proximity* method achieves better mAP than the $tfidf$ method

Those 3 sub-statements and their t-test statistics are listed in Table 5.10. We choose a significance level of 0.05 ($\alpha = 0.05$). The label *RE* denotes the recall value. H_1 represents our statement/conclusion. H_0 is the null hypotheses of our statement. The t-test statistic is computed as,

$$t = \frac{M_1 - M_2}{\sqrt{\left(\frac{(N_1-1)S_1^2+(N_2-1)S_2^2}{N_1+N_2-2}\right)\left(\frac{1}{N_1}+\frac{1}{N_2}\right)}} \tag{5.19}$$

The dimension *df* of t-test is computed as:

$$df = N1 + N2 - 2; \tag{5.20}$$

H_1	H_0	N		Mean		Std. Deviation		t	p-value
		N_1	N_2	M_1	M_2	S_1	S_2		
$RE_{proximity} > RE_{baseline}$	$RE_{proximity} \leq RE_{baseline}$	10	10	100.00	98.70	0	1.79	2.29	$< p_{0.025}$
$mAP_{proximity} > mAP_{tfidf}$	$mAP_{proximity} \leq mAP_{tfidf}$	10	10	91.60	82.60	5.45	7.84	2.98	$< p_{0.005}$
$mAP_{proximity} > mAP_{conf}$	$mAP_{proximity} \leq mAP_{conf}$	10	10	91.60	51	5.45	23.12	5.40	$< p_{0.0005}$

Table 5.10: t-test statistics of long queries (length $>$ 10) related statements.

N denotes the dimension of given sample vector. In our case, the *df* has a value of 18. According to the one-tailed t-table ([1]), the *p*-value of the first statement in Table 5.10 lies in range $p_{0.025}$ and $p_{0.01}$. The *p*-value of the first statement is less than 0.05 (our significant level). Therefore, the H_0 of the first statement can be rejected. Our first statement is acceptable. The *p*-value of the second statement in Table 5.10 lies in range $p_{0.005}$ and $p_{0.001}$. Thereby our second statement is also acceptable. The *p*-value of the third statement in Table 5.10 is far below $p_{0.0005}$. Therefore the third statement is acceptable.

Compared to the *baseline* system a clear improvement in the retrieval recall with *proximity* method can also be observed for the short queries (+6.8%) and for queries including 7 or 8 phones (+2.5%). The performance of retrieving very short queries with (length $<$ 5) is summarized in Table 5.15. Table 5.14 and Table 5.13 show the performance of retrieving queries ($length = 7$ and $length = 8$). Statistic significance research is also made to confirm this conclusion. We break this conclusion into two statements. The statements and their t-test statistics are listed in Table 5.11. The degree of freedom *df* is 18. According to the one-tailed t-table ([1]), we find out that the *p*-value of the first statement in Table 5.11 lies between $p_{0.001}$ and $p_{0.005}$, the p-value of the second statement is less than $p_{0.025}$ and the p-value of the third statement is less than $p_{0.005}$. All of them are less than our pre-defined significance-level ($\alpha = 0.05$). Therefore, our conclusions listed in Table 5.11 are statistically significant.

H_1	H_0	N		Mean		Std. Deviation		t	p-value
		N_1	N_2	M_1	M_2	S_1	S_2		
$RE_{short\_proximity} > RE_{short\_baseline}$	$RE_{short\_proximity} \leq RE_{short\_baseline}$	10	10	98.00	91.20	2.59	6.56	3.04	$< p_{0.005}$
$RE_{7\_proximity} > RE_{7\_baseline}$	$RE_{7\_proximity} \leq RE_{7\_baseline}$	10	10	99.40	96.90	0.83	3.33	2.30	$< p_{0.025}$
$RE_{8\_proximity} > RE_{8\_baseline}$	$RE_{8\_proximity} \leq RE_{8\_baseline}$	10	10	99.60	96.40	1.25	2.94	3.17	$< p_{0.005}$

Table 5.11: t-test statistics of statements about improved recall with *proximity* method when retrieving not very long queries (length $<$ 8).

It can also be observed in Table 5.17, that the $tfidf$ method shows its benefits in retrieving short queries. In comparison with the *baseline* system, the maximal recall rate increases by about 7.8%. Retrieving short queries with the *proximity* method

		0	10	20	30	40	50	60	70	80	90	100	mAP	max. RE
Administration	baseline	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	98.43	99.84	100.00
	tfidf	100.00	85.71	92.85	95.00	80.64	84.21	79.17	80.00	78.13	79.17	76.82	83.17	100.00
	conf	25.00	50.00	54.17	54.29	51.02	44.92	42.53	40.00	39.68	34.54	20.86	43.20	100.00
	proximity	50.00	75.00	86.67	90.47	92.59	94.12	92.50	93.62	94.34	95.00	95.45	90.98	100.00
Environmental	baseline	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	97.56	97.82	-	89.54	98.00
	tfidf	100.00	100.00	100.00	100.00	100.00	96.15	88.23	85.37	81.63	78.95	58.14	88.85	100.00
	conf	100.00	55.56	50.00	44.12	50.00	50.00	37.97	41.67	43.96	44.12	32.47	44.99	100.00
	proximity	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	1.90	90.19	100.00
Executives	baseline	100.00	71.43	83.33	87.50	90.48	92.31	93.33	94.29	95.00	95.45	95.92	89.90	100.00
	tfidf	50.00	71.43	75.00	82.35	86.36	85.71	84.85	84.62	74.51	70.00	52.22	76.71	100.00
	conf	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	97.92	99.79	100.00
	proximity	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	95.92	99.59	100.00
Development	baseline	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
	tfidf	100.00	100.00	100.00	100.00	100.00	100.00	87.88	85.00	84.78	83.02	77.78	91.85	100.00
	conf	100.00	50.00	50.00	57.69	60.61	62.50	56.86	55.74	58.21	60.27	58.33	57.02	100.00
	proximity	100.00	83.33	90.91	93.75	95.24	96.15	96.67	97.14	97.50	97.78	98.00	94.65	100.00
Investment	baseline	100.00	100.00	100.00	100.00	100.00	92.00	90.32	91.43	92.50	91.11	88.46	94.58	100.00
	tfidf	100.00	100.00	100.00	87.50	78.26	79.31	75.67	76.15	72.55	66.13	46.94	78.25	100.00
	conf	100.00	41.67	42.86	38.89	33.33	29.11	23.93	21.48	22.70	22.16	14.07	29.02	100.00
	proximity	100.00	100.00	81.82	82.35	78.26	82.14	82.35	83.78	84.09	85.41	85.19	84.54	100.00
Incorporated	baseline	100.00	100.00	100.00	100.00	100.00	100.00	96.29	96.67	97.14	97.50	-	88.76	95.35
	tfidf	100.00	100.00	100.00	100.00	85.00	88.00	86.67	85.71	80.95	73.58	10.64	81.06	100.00
	conf	16.67	35.71	42.86	36.11	32.69	34.37	31.71	25.64	23.61	18.84	10.31	29.19	100.00
	proximity	100.00	100.00	100.00	100.00	94.44	95.65	96.29	96.77	97.14	95.12	2.00	87.74	100.00
International	baseline	100.00	85.71	92.31	86.36	89.29	91.18	90.24	91.49	90.91	91.80	92.54	90.18	100.00
	tfidf	100.00	85.71	86.67	90.00	78.13	81.58	82.22	84.31	86.21	84.85	60.19	81.99	100.00
	conf	100.00	75.00	57.14	43.18	43.10	37.80	33.64	33.86	30.30	27.72	22.22	40.40	100.00
	proximity	100.00	100.00	100.00	90.47	92.59	93.94	92.50	93.48	90.91	91.80	92.54	93.82	100.00
California	baseline	100.00	90.00	90.00	93.10	94.74	95.83	96.49	95.52	94.81	95.35	-	84.58	97.80
	tfidf	100.00	90.00	90.00	90.00	92.31	91.84	93.22	88.89	82.95	83.67	81.25	88.41	100.00
	conf	100.00	90.00	90.00	90.00	92.31	91.84	88.71	90.14	91.25	89.13	85.85	89.92	100.00
	prox	100.00	100.00	94.74	96.43	97.30	95.74	96.49	96.97	97.33	97.62	79.47	95.21	100.00
Department	baseline	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	-	90.00	95.65
	tfidf	50.00	83.33	81.82	87.50	85.71	82.14	84.85	84.21	84.09	80.39	9.00	76.30	100.00
	conf	50.00	50.00	52.94	50.00	40.00	45.10	40.58	34.04	20.33	17.37	3.00	35.34	100.00
	proximity	50.00	83.33	90.00	93.33	94.73	92.00	93.33	91.42	90.24	91.11	1.52	82.10	100.00
Securities	baseline	100.00	80.00	80.00	80.00	84.21	80.77	83.33	82.86	84.62	82.22	82.00	82.00	100.00
	tfidf	100.00	80.00	72.73	66.67	66.67	60.00	60.98	64.44	62.26	58.73	43.62	63.61	100.00
	conf	100.00	80.00	66.67	60.00	43.24	40.38	41.67	43.28	41.77	41.11	40.20	49.83	100.00
	proximity	100.00	100.00	100.00	92.31	72.73	77.78	78.13	80.56	80.48	82.22	82.00	84.62	100.00

Table 5.12: Performance evaluation of in-vocabulary queries with length > 10.

achieves a recall that is comparable to the *tfidf* method. The *tfidf* method yields a better mAP value (about + 7.6%) compared to the *proximity* method. However, the mAP-value drops for about 17.1% compared to the *baseline* method. The results of research about the statistic significance related this conclusion are summarized in Table 5.16. The results in Table 5.16 confirm that retrieving short queries with the *tfidf* method achieves better recall than that with the *baseline* method. But retrieving short queries with the *baseline* method yields better mAP-value. The p-value of the second statement in Table 5.16 is greater than our pre-defined significance-level ($\alpha = 0.05$). Consequently, we can not mention that retrieving short queries with the *tfidf* method

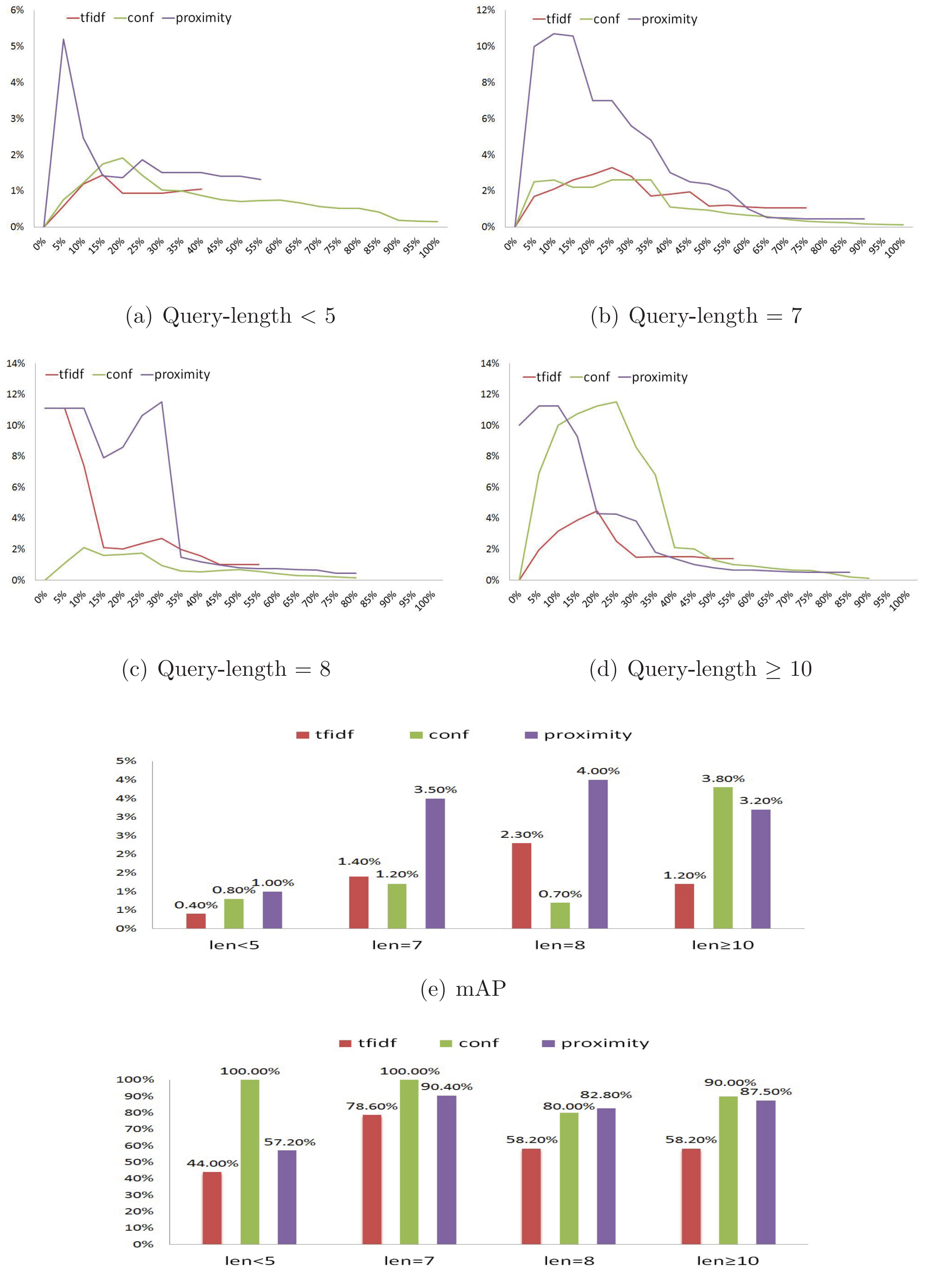
		0	10	20	30	40	50	60	70	80	90	100	mAP	max. RE
Agreement	baseline	100.00	100.00	90.91	83.33	83.33	86.20	88.57	90.00	89.13	90.12	-	80.16	98.03
	tfidf	100.00	71.43	83.33	83.33	76.92	80.64	79.49	73.47	67.21	69.71	62.20	74.77	100.00
	conf	100.00	60.00	66.67	71.43	64.52	55.56	56.36	58.06	56.16	44.23	38.35	57.13	100.00
	prox	100.00	71.43	83.33	88.24	80.00	65.79	67.39	70.59	70.69	71.88	1.73	67.11	100.00
American	baseline	100.00	100.00	100.00	97.56	96.36	95.65	95.18	95.83	95.50	95.97	-	87.21	98.48
	tfidf	50.00	63.64	75.00	72.73	61.63	62.26	62.70	62.59	61.99	62.63	59.73	64.49	100.00
	conf	50.00	24.07	27.34	36.36	28.80	27.62	23.51	19.05	17.73	12.65	9.08	22.62	100.00
	prox	100.00	100.00	100.00	100.00	94.64	94.29	95.18	95.83	92.17	90.84	2.56	86.55	100.00
Companies	baseline	100.00	68.00	66.00	67.57	62.62	60.14	61.35	63.58	64.73	64.94	-	57.89	92.22
	tfidf	100.00	45.95	45.21	39.37	39.88	38.78	38.61	34.21	32.45	31.78	7.57	35.38	100.00
	conf	100.00	48.57	49.25	49.02	50.38	44.86	40.65	36.00	34.73	35.80	9.94	39.92	100.00
	prox	100.00	65.38	58.93	60.24	61.47	54.97	55.25	55.98	57.02	59.29	10.57	53.91	100.00
Exchange	baseline	100.00	100.00	90.00	93.33	94.74	92.00	93.33	94.12	94.87	95.35	93.88	94.16	100.00
	tfidf	100.00	83.33	90.00	93.33	78.26	82.14	80.00	78.05	75.51	73.21	71.87	80.57	100.00
	conf	100.00	100.00	100.00	100.00	85.71	88.46	90.32	91.42	90.24	89.13	88.46	92.37	100.00
	prox	100.00	100.00	100.00	100.00	94.74	92.00	93.33	94.12	92.50	93.18	93.88	95.38	100.00
Government	baseline	100.00	100.00	100.00	100.00	96.97	96.39	96.97	96.55	96.24	96.64	-	87.98	96.87
	tfidf	100.00	100.00	94.12	94.12	92.75	87.91	88.07	89.60	89.51	88.89	10.67	83.56	100.00
	conf	100.00	100.00	88.89	88.89	83.12	80.81	80.00	80.00	79.01	79.56	22.32	78.26	100.00
	prox	100.00	94.12	94.12	90.57	90.14	86.96	88.07	89.60	89.51	90.00	5.43	81.85	100.00
military	baseline	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	-	90.00	90.48
	tfidf	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	87.50	98.75	100.00
	conf	100.00	40.00	27.59	28.26	34.00	37.50	41.67	44.62	30.63	24.84	15.91	32.50	100.00
	prox	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	84.00	98.40	100.00
proposal	baseline	100.00	100.00	100.00	100.00	100.00	100.00	100.00	97.06	97.44	97.67	-	89.22	97.87
	tfidf	100.00	100.00	100.00	100.00	95.00	96.00	96.55	86.84	86.36	79.25	41.96	88.20	100.00
	conf	33.33	62.50	52.94	58.33	50.00	48.98	42.42	35.48	33.93	32.56	24.23	44.14	100.00
	prox	100.00	100.00	100.00	100.00	100.00	100.00	100.00	97.06	97.44	97.67	37.30	92.95	100.00
spokesman	baseline	50.00	80.00	90.00	86.67	90.00	66.67	53.06	48.44	43.75	45.98	-	60.46	97.72
	tfidf	50.00	80.00	90.00	76.47	75.00	70.97	70.27	54.39	47.30	42.55	41.51	64.85	100.00
	conf	25.00	7.69	9.89	12.50	13.43	12.79	12.38	11.31	10.77	11.27	9.63	11.17	100.00
	prox	50.00	30.77	23.68	25.49	33.51	31.88	33.77	37.35	37.77	43.01	4.03	30.13	100.00
students	baseline	100.00	100.00	100.00	100.00	100.00	100.00	100.00	91.89	90.48	89.58	-	87.20	93.75
	tfidf	100.00	100.00	83.33	87.50	90.48	92.31	90.63	85.00	77.55	76.79	-	78.36	95.83
	conf	100.00	38.46	14.08	12.73	10.92	13.41	15.76	17.44	14.18	9.09	0.84	14.69	100.00
	prox	100.00	83.33	66.67	70.00	70.37	72.73	74.36	77.27	79.17	79.63	-	67.35	95.83
recently	baseline	100.00	100.00	76.92	76.19	72.41	76.47	73.81	73.47	76.36	74.60	-	70.02	98.07
	tfidf	100.00	100.00	71.43	69.57	65.63	70.27	73.81	73.47	68.85	63.51	2.95	65.95	100.00
	conf	100.00	33.33	27.03	29.63	35.00	28.26	28.44	23.68	24.71	25.41	2.09	25.76	100.00
	prox	100.00	55.56	62.50	66.67	63.64	68.42	70.45	73.47	76.36	75.81	1.15	61.40	100.00

Table 5.13: Performance evaluation of in-vocabulary queries with length = 8.

yields a better mAP-value than that with the *proximity* method. In this case, the null-hypotheses can not be rejected. Therefore, this conclusion can not be accepted.

The ability of phone-3gram based spoken document method to deal with out-of-vocabulary words was analysed in the previous experiments (Figure 5.8). In this section, we are going to research whether the query-length has an impact on the out-of-vocabulary query retrieval performance. The obtained results are compared in Figure 5.11 and Table 5.18.

The result of this study shows that the mean average precision of the phone-3gram



(a) Query-length < 5

(b) Query-length = 7

(c) Query-length = 8

(d) Query-length ≥ 10

(e) mAP

(f) max. recall

Figure 5.11: Performance (precision/recall plot) of OOV queries with different lengths.

		0	10	20	30	40	50	60	70	80	90	100	mAP	max. RE
Capital	baseline	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	-	90.00	95.91
	tfidf	100.00	100.00	100.00	100.00	95.24	96.00	90.63	85.00	78.00	78.57	-	82.34	97.95
	conf	33.33	35.71	25.00	28.30	26.67	15.38	13.81	14.66	14.55	14.62	0.85	18.96	100.00
	prox	100.00	100.00	83.33	83.33	83.33	82.76	85.29	87.18	88.63	86.27	-	78.01	97.96
Congress	baseline	100.00	100.00	100.00	94.44	92.00	90.63	92.11	93.18	93.88	94.55	-	85.08	98.28
	tfidf	50.00	85.71	85.71	89.47	88.46	90.63	89.74	91.11	92.00	92.86	93.55	89.92	100.00
	conf	100.00	66.67	63.16	70.83	74.19	69.05	68.63	69.50	71.88	73.24	54.72	68.19	100.00
	prox	100.00	85.71	92.31	94.44	92.00	93.55	94.60	95.35	95.83	96.29	5.09	84.52	100.00
December	baseline	100.00	100.00	100.00	100.00	94.12	95.24	96.15	96.67	97.06	97.37	95.35	97.20	100.00
	tfidf	100.00	100.00	100.00	100.00	100.00	100.00	89.29	90.63	91.67	92.50	87.23	95.13	100.00
	conf	100.00	100.00	100.00	100.00	100.00	100.00	100.00	90.32	89.19	86.04	82.00	94.76	100.00
	prox	100.00	100.00	100.00	100.00	100.00	100.00	100.00	96.67	97.06	94.87	95.35	98.40	100.00
Economy	baseline	100.00	100.00	94.12	95.83	96.88	97.50	97.92	98.21	98.41	98.59	96.29	97.38	100.00
	tfidf	100.00	100.00	94.12	85.19	83.78	82.98	85.45	85.94	81.58	82.35	78.79	86.02	100.00
	conf	100.00	100.00	100.00	92.00	88.57	88.63	87.04	87.30	86.30	85.37	82.98	89.82	100.00
	prox	100.00	100.00	100.00	95.83	96.88	97.50	97.92	98.21	98.44	98.59	97.50	98.09	100.00
Increased	baseline	100.00	83.33	91.67	94.12	87.50	81.25	80.00	80.43	82.35	-	-	68.07	88.68
	tfidf	100.00	50.00	55.00	55.17	60.00	50.98	42.67	37.76	33.60	28.57	-	41.38	98.11
	conf	100.00	20.00	22.92	20.25	18.58	18.57	16.50	14.80	12.00	7.99	1.00	15.26	100.00
	prox	100.00	83.33	78.57	84.21	87.50	81.25	84.21	84.09	85.71	1.97	-	67.08	98.11
Markets	baseline	100.00	75.00	63.64	71.43	77.78	73.91	77.78	80.00	80.00	81.58		68.11	97.14
	tfidf	50.00	50.00	50.00	55.56	29.17	25.37	19.27	18.18	18.54	16.94	13.57	29.66	100.00
	conf	100.00	15.79	15.21	18.18	17.50	12.98	13.13	13.79	14.51	14.35	13.73	14.92	100.00
	prox	100.00	100.00	53.85	55.56	56.00	60.71	63.64	64.86	66.67	65.96	67.31	65.46	100.00
Program	baseline	100.00	85.71	92.86	95.24	93.10	94.29	95.24	95.83	96.43	96.77	97.10	94.26	100.00
	tfidf	100.00	100.00	76.47	80.00	84.38	84.62	86.96	88.68	85.71	83.33	67.00	83.72	100.00
	conf	100.00	100.00	72.22	76.92	79.41	78.57	76.92	75.41	77.14	78.95	76.13	79.17	100.00
	prox	100.00	100.00	100.00	100.00	100.00	97.06	97.56	97.92	98.18	98.36	98.53	98.76	100.00
Position	baseline	100.00	100.00	100.00	100.00	100.00	100.00	90.32	89.19	88.37	87.50	-	85.54	93.61
	tfidf	100.00	50.00	64.29	73.68	79.17	82.14	80.00	63.46	50.00	48.84	48.45	64.00	100.00
	conf	4.55	4.90	3.86	3.30	2.48	2.05	1.33	1.50	1.72	1.89	1.35	2.44	100.00
	prox	100.00	50.00	56.25	53.85	43.18	31.94	32.94	34.74	34.23	35.00	0.90	37.30	100.00
United	baseline	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	-	90.00	97.69
	tfidf	100.00	100.00	100.00	97.50	98.11	97.01	97.50	97.85	98.10	98.32	-	88.44	99.23
	conf	100.00	54.17	40.63	41.05	48.15	50.78	51.32	53.53	55.62	57.35	2.44	45.50	100.00
	prox	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	-	90.00	99.23
Yesterday	baseline	33.33	84.62	88.46	91.89	93.75	94.92	95.77	95.18	95.74	96.23	-	83.66	97.35
	tfidf	33.33	78.57	88.46	87.18	90.00	90.32	90.67	89.77	87.38	88.70	-	79.11	99.12
	conf	100.00	84.62	88.46	91.89	93.75	94.92	94.44	95.18	95.74	94.44	2.40	83.58	100.00
	prox	100.00	100.00	100.00	100.00	95.74	96.55	97.14	97.53	96.77	97.14	-	88.09	99.12

Table 5.14: Performance evaluation of in-vocabulary queries with length = 7.

based spoken document retrieval system increases with growing query-length (Figure 5.11(e) and (f)). The *proximity* method outperforms the others. It can be observed, that even with a high maximal recall rate (100% by queries including 5 and 7 phones with the *conf*-method), the phone-3gram based SDR approaches cannot provide reliable retrieval results for out-of-vocabulary queries due to the poor mean average precision.

		0	10	20	30	40	50	60	70	80	90	100	mAP	max. RE
Certain	baseline	25.00	57.14	63.64	68.75	75.00	69.23	68.75	72.22	73.17	-	-	54.79	83.78
	tfidf	16.67	30.77	42.11	47.83	46.88	42.86	41.51	45.61	46.88	41.77	1.20	38.74	100.00
	conf	33.33	14.29	7.00	6.92	6.05	5.77	6.71	5.14	4.19	1.80	1.13	5.90	100.00
	prox	16.67	16.67	16.28	19.64	23.08	23.68	21.57	18.98	16.30	16.90	1.40	17.45	100.00
Costs	baseline	100.00	83.33	78.57	80.00	78.57	79.41	72.73	74.51	71.67	74.24	-	69.30	94.44
	tfidf	100.00	71.43	68.75	66.67	66.67	64.29	62.75	48.72	48.31	46.23	-	54.38	98.15
	conf	100.00	83.33	47.83	51.61	36.67	33.75	32.32	29.69	28.10	25.13	1.48	36.99	100.00
	prox	100.00	100.00	91.67	80.00	75.86	67.50	69.57	73.08	72.88	72.06	-	70.26	98.15
Foreign	baseline	100.00	100.00	75.00	72.22	78.26	78.57	74.29	73.81	-	71.43	72.73	69.63	95.45
	tfidf	100.00	80.00	64.29	56.52	56.25	53.66	57.78	56.36	58.33	37.38	-	52.06	97.72
	conf	10.00	6.15	5.63	3.82	3.15	3.17	2.85	2.90	2.69	2.80	0.93	3.41	100.00
	prox	100.00	66.67	29.03	22.41	23.08	25.58	18.84	19.38	19.89	21.05	-	24.59	97.73
Funds	baseline	100.00	100.00	100.00	92.86	89.47	84.00	80.65	80.56	82.93	79.17	79.25	86.89	100.00
	tfidf	100.00	100.00	50.00	59.09	65.38	53.85	51.02	46.03	43.59	41.30	35.59	54.59	100.00
	conf	100.00	80.00	61.54	44.83	42.50	46.67	49.02	44.62	43.04	42.70	42.42	49.73	100.00
	prox	100.00	57.14	72.73	61.90	65.38	67.74	67.57	70.73	73.91	76.00	77.78	69.09	100.00
Meeting	baseline	100.00	66.67	80.00	85.71	88.89	90.48	92.00	93.10	91.18	-	-	68.80	79.49
	tfidf	33.33	50.00	66.67	66.67	66.67	67.86	65.71	64.29	65.96	3.56	-	51.74	94.87
	conf	33.33	21.05	26.67	30.00	24.59	10.73	10.60	9.25	9.78	3.08	0.50	14.63	100.00
	prox	100.00	50.00	66.67	70.59	64.00	52.78	51.11	52.94	54.38	1.73	-	46.42	94.87
Stocks	baseline	50.00	50.00	66.67	76.92	81.25	80.00	83.33	79.31	78.79	78.95	-	67.52	90.91
	tfidf	100.00	50.00	70.00	55.56	46.43	51.61	51.28	50.00	45.61	40.54	-	46.10	96.97
	conf	20.00	18.75	7.78	8.33	6.84	7.51	8.44	9.13	9.56	10.31	0.55	8.72	100.00
	prox	50.00	60.00	75.00	76.92	65.00	69.57	71.43	74.19	74.29	75.00	-	64.14	96.97
Total	baseline	100.00	100.00	100.00	92.86	94.44	87.50	89.66	88.24	87.18	86.67	-	82.66	90.70
	tfidf	100.00	80.00	81.82	68.42	65.38	65.63	65.00	61.22	64.15	60.00	1.87	61.35	100.00
	conf	14.28	12.50	12.68	10.24	8.81	8.43	7.00	7.06	6.91	7.16	2.20	8.30	100.00
	prox	33.33	66.67	81.82	65.00	68.00	70.00	74.29	75.00	75.56	75.00	1.65	65.30	100.00
Value	baseline	14.29	45.45	62.50	70.00	70.37	75.00	76.32	79.07	76.00	68.25	65.75	68.87	100.00
	tfidf	12.50	41.67	55.56	63.64	67.86	68.57	70.73	65.38	62.30	62.32	54.55	61.26	100.00
	conf	100.00	38.46	31.25	30.43	33.33	36.36	40.28	44.16	45.78	45.26	45.28	39.06	100.00
	prox	100.00	62.50	76.92	82.35	79.17	82.76	80.56	82.93	74.51	76.79	69.57	76.81	100.00
Woman	baseline	100.00	100.00	100.00	100.00	100.00	100.00	96.29	93.75	94.44	92.86	-	87.73	93.02
	tfidf	100.00	100.00	100.00	100.00	100.00	95.45	92.86	93.75	91.89	90.70	0.87	86.55	100.00
	conf	50.00	5.06	6.16	5.73	4.72	4.34	4.30	3.95	3.70	2.28	0.68	4.09	100.00
	prox	100.00	30.77	27.27	28.89	28.33	28.77	32.91	30.61	24.11	20.21	1.17	25.30	100.00
Better	baseline	50.00	83.33	83.33	88.24	86.96	89.29	81.08	83.33	81.63	-	-	67.72	84.00
	tfidf	50.00	71.43	83.33	83.33	86.96	86.21	83.33	83.33	76.92	39.13	-	69.40	92.00
	conf	50.00	71.43	55.56	65.22	58.82	55.56	56.60	54.69	51.28	3.46	0.62	47.32	100.00
	prox	100.00	100.00	71.43	71.43	71.43	75.76	78.95	77.78	76.92	3.91	-	62.76	92.00

Table 5.15: Performance evaluation of in-vocabulary short queries with length < 5.

H_1	H_0	N		Mean		Std. Deviation		t	p-value
		N_1	N_2	M_1	M_2	S_1	S_2		
$RE_{short\_tfidf} > RE_{short\_baseline}$	$RE_{short\_tfidf} \leq RE_{short\_baseline}$	10	10	98.00	91.20	2.58	6.56	3.04	$< p_{0.005}$
$mAP_{short\_tfidf} > mAP_{short\_proximity}$	$mAP_{short\_tfidf} \leq mAP_{short\_proximity}$	10	10	57.7	50.1	12.56	21.01	0.98	$> p_{0.25}$
$mAP_{short\_baseline} > mAP_{short\_tfidf}$	$mAP_{short\_baseline} \leq mAP_{short\_tfidf}$	10	10	74.8	57.7	10.03	12.56	3.36	$< p_{0.005}$

Table 5.16: t-test statistics of statements about *tfidf* benefits short query (length < 5) retrieval task.

Conclusion

The results of the in-vocabulary query retrieval task, as shown in Table 5.5 and in Figure 5.7, confirm that the phone-3gram based spoken document retrieval system

	len < 5		len = 7		len = 8		len ≥ 10	
	mAP	maxRE	mAP	maxRE	mAP	maxRE	mAP	maxRE
baseline	74.80	91.20	91.70	96.90	88.30	96.40	94.60	98.70
tfidf	57.70	98.00	72.90	99.40	70.70	99.60	82.60	100.00
conf	17.70	100.00	45.10	100.00	39.20	100.00	51.00	100.00
proximity	50.10	98.00	83.90	99.40	73.80	99.60	91.60	100.00

Table 5.17: Performance evaluation of in-vocabulary queries with different lengths.

	len < 5		len = 7		len = 8		len ≥ 10	
	mAP	maxRE	mAP	maxRE	mAP	maxRE	mAP	maxRE
tfidf	0.40	44.00	1.40	78.60	2.30	58.20	1.20	58.20
conf	0.80	100.00	1.20	100.00	0.70	80.00	3.80	90.00
proximity	1.00	57.20	3.50	90.40	4.00	82.80	3.20	87.50

Table 5.18: Performance evaluation of out-of-vocabulary queries with different lengths.

achieves a reliable retrieval performance. The maximum retrieval recall is significantly improved. However, the mean average precision drops. In this work the new phone-3gram based *proximity* method labelled as *proximity_weighting* yields the best performance. The *proximity* method reaches a 69.94% mean average precision while keeping a 95.99% maximal recall. The best retrieval recall is achieved with the $conf$ method. The $conf$ method enables approximate matching between a query and the document terms, which results in too many false alarms being added into the retrieved result list. Consequently, the $conf$ method can only achieve 39.93% mean average precision.

As can be seen in Figure 5.8 and in Table 5.6, the phone-3gram based methods are not suitable for out-of-vocabulary query tasks. Even with the $tfidf$ method, a 92.83% maximal recall can be reached, but the mean average precision is very poor. The *proximity* labelled as *proximity_weighting* method yields the best mean average precision.

The real application experiments, as shown in Figure 5.9(b) and in Table 5.7, reflect that the proposed *proximity* method labelled as *poximity_weighting*) outperforms the baseline system. In comparison with the baseline system, this new method improves the maximal recall rate by about 15.34% and reaches a 61.08% mean average precision that is comparable to the baseline system. Adding multiple recognition hypothesis to the document representation in the form of the phone-3gram PSPL does not benefit the retrieval performance of a spoken document retrieval system.

Another major finding, as shown in Figure 5.10, is that the query-length has an impact on retrieval performance of a phone-3gram based spoken document retrieval system. Generally, the mean average precision increases with a growing query-length. The *proximity* method yields the best mean average precision for long query words. The best retrieval recall is achieved with the $conf$ approach, which is independent to the query-length. The mean average precision of the $conf$ system drops with decreasing query-length. These conclusions have been verified with additional statistical significance tests. Experiment results in previous experiments confirmed that phone-3gram base spoken document retrieval systems are generally not for the out-of-vocabulary query retrieval task.

5.4 Probabilistic Phone String Matching

In the last section, we explored the phone-3gram based spoken document retrieval approaches. The experimental results indicate that this kind of spoken document retrieval methods can detect some of the out-of-vocabulary words. However, the retrieval precision is low. The reasons for this are:

- The phone-3gram is still a kind of context-dependent index.
- The phone-3gram representation of a spoken document is directly extracted from its word-1best representation, provided by a speech recognizer.

The context-free index unit (very short index unit, e.g. phone) should provide more flexibility in the retrieval process. Therefore, in this section we focus on the investigation of context-free index unit-based spoken document retrieval methods. There are two context-independent index units that have been applied to the information retrieval application: grapheme and phone. In comparison with grapheme the phone index unit bears extra useful acoustic information. For this reason, we are going to investigate phone-based (monophone) spoken document retrieval methods in this section. Based on the previous experiments in Section 5.2, the monophone representation of a spoken document is constructed from its word-1best transcription. The main obstacle that the monophone-based spoken document retrieval methods have to deal with is the large number of transcription errors.

Wechsler et. al. ([115], [117]) tried to compensate for the transcription errors by means of exploiting the probabilities of the phone recognition errors. The notion 'probabilistic string matching' was defined by Wechsler as the retrieval methods, which are based on the spotting query (phone) sequences in a continuous and corrupted document (phone) sequences [115]. The string matching method can be used for both spoken document retrieval and text retrieval in a scanned document generated by an Optical Character Recognizer (OCR).

This method were evaluated using German information retrieval task with short queries [79]. 10 German city names were selected as queries. Three distinct sets of queries were built, queries in the text-form, clean spoken queries and noisy spoken queries that was recorded once in adverse conditions (low quality microphone, presence of background noise). Moreau and Jin reported that the clean spoken queries do not perform as well as text queries. The experiments also showed that the string matching SDR act more effectively in compare with phone-Ngrams based SDR methods. Further experiments on German retrieval tasks were conducted and summarized in Jin's Master thesis [56]. Also short 10 German city names were used as queries to evaluate different phone-based SDR methods in Jin's Master thesis. ([56]).

The string matching method applied for spoken document retrieval usually consists of three main components: slot detection, slot-probability estimation and retrieval status value (RSV) estimation (as shown in Figure 5.12).

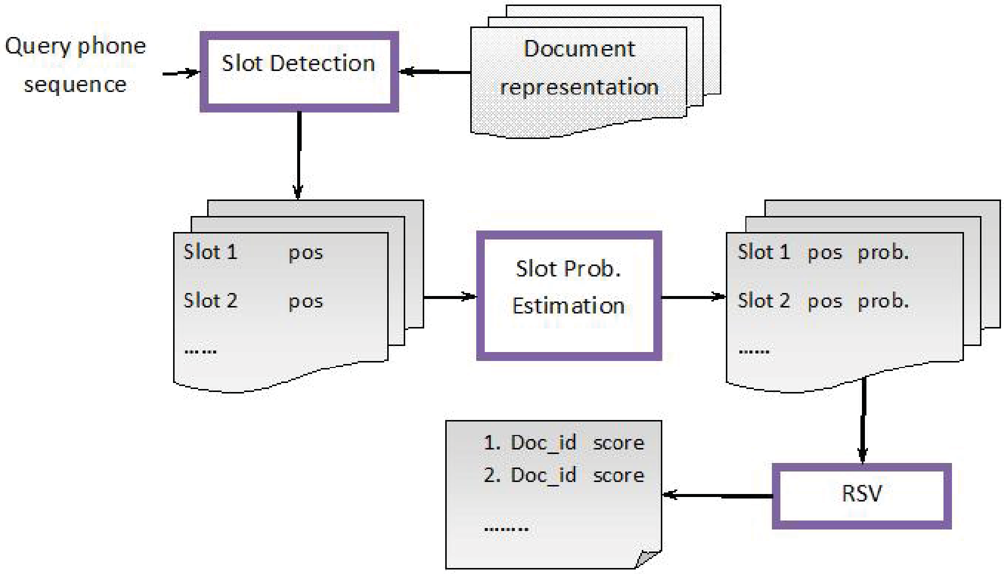


Figure 5.12: Structure of an SM-based SDR system.

Similar approaches were proposed by Amir et. al. [4] using metaphone (broad phone class) instead of monophones. In his paper, phones are grouped into 7 metaphones based on their confusion information. The monophone representation of a spoken document was updated into metaphone representation. Slight improvement in precision could only be confirmed in combining with word-based SDR methods. Therefore, we are going to focus on the information retrieval in monophone representation of spoken documents.

In this work, probabilistic string matching methods will further evaluated on English single-word query retrieval tasks. The query-set contains 200 in-vocabulary words and 30 out-of-vocabulary words with different length. The effect of query-length on retrieval performance will also be researched 5.4.4. The query phone sequence in this work is obtained via the text-to-phone tools (phone dictionary) or via a phone recognizer. In the slot detection phase, all possible slots that may contain the query phone sequence will be detected. Then a probability will be estimated for each detected slot. This slot probability describes the similarity between a detected document slot and the given query phone sequence. Finally, the retrieval status value (RSV) is estimated based on the slot probability. This retrieval status value is used to rank the documents in the retrieved list.

5.4.1 Slot Detection

The slot detection problem can be defined as the search of a set of slots $S(q, d)$ in the document mono-phone transcription d, according to a given query sequence q, so that each slot $S(q, d)$ is a possible occurrence of sequence q in d. The following terms are defined for expressing the quality of a slot detection method.

- Hit slot : a detected slot containing the query sequence.
- Miss slot : a slot in which a query feature is spoken but not detected.
- False alarm slot : a detected slot that does not contain the query sequence.

Because the phone sequences in a spoken document representation are often corrupted, the slot detection component can return false alarm slots or miss some slots containing the query phone sequences. In order to improve the slot detection performance, we must reduce the number of miss slots and false alarms. However, the miss and false alarm errors are correlated. Decreasing the number of miss slots usually increases the number of false alarms. Conversely, decreasing the number of false alarms always leads to an increase in the number of miss slots. As stated before, the errors in the spoken document representation can be classified by three errors: substitution, insertion and deletion. Wechsler has proposed three different slot detection methods to deal with this problem: slot detection via exact matching, substitution-tolerant slot detection and error-tolerant slot detection.

The **Exact matching method** detects the slots in a spoken document representation that are exact copies of the given query phone sequence. This method is very restrictive and does not suit the spoken document retrieval application very well, as the phone sequences in a spoken document representation are often corrupted.

The **Substitution-tolerant method** counts the number of common phones that occur at the same position within the query sequence and a document slot. Only slots whose number of matching phones is greater than a pre-defined threshold value are returned as detected slots in a document. This method only takes the substitution errors into account. All detected slots have a fixed length.

The insertion or deletion errors in a spoken document representation introduce a shift in the relative position of a phone. The substitution tolerant slot detection method is a fixed-length approach and is unable to handle these kind of errors. The **Error-tolerant method** is designed to cope with these errors. This method is robust against all types of recognition errors. Firstly, a final beginning score is derived for each document position k. This score reflects the probability of a candidate slot at the position k. Then the slots are ranked in decreasing order according to their final beginning scores. No overlapping is allowed between the detected document slots. The method consists of three steps: preparation, final beginning score calculation and the selection of slots.

Figure 5.13 presents the output of three different slot-detection methods with a simple example.

The task of this example is very simple: finding possible slots in a document $4o8c0e06$ given a query word 'BUT' with phone sequence 'B AH T'. As we can see in Figure 5.13, the *exact* method can only detect slots that are exactly the same as the given query phone sequence. In this example, this method is perfect; a 'hit' is found. However, as we mentioned before, the reference phone transcription of a spoken

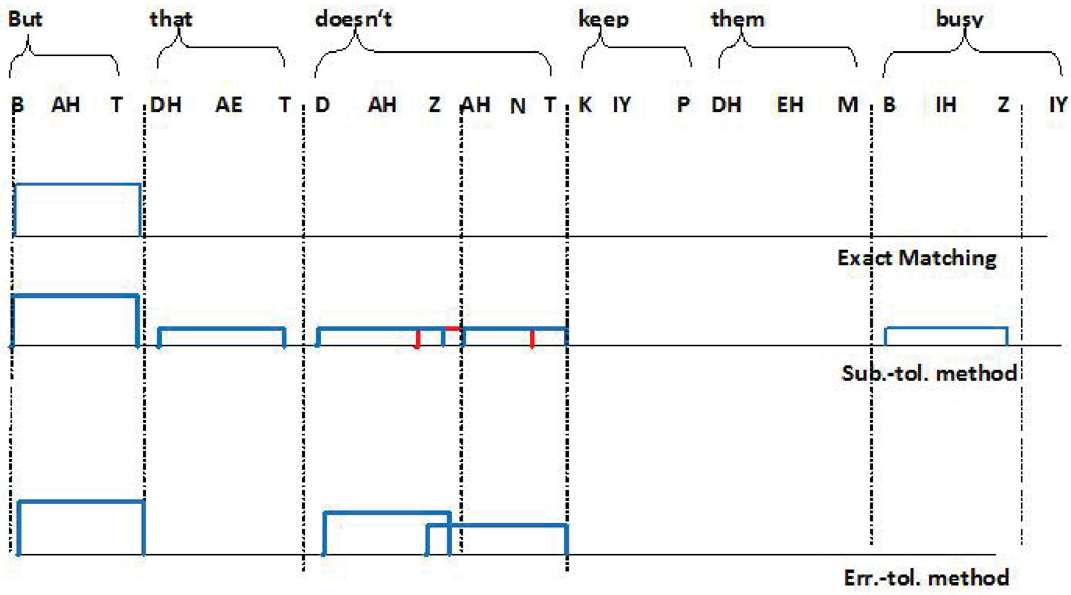


Figure 5.13: Example of slot-detection: looking for BUT (B AH T) in $4o8c0e06$.

document is often corrupt and contains a lot of errors. It will be very hard to find a slot that is a perfect *match* for the given query phone sequence. More slots that are similar to the query phone sequence are detected using the *substitution_tolerant* method. The overlapped slots are ignored. In order to reduce the number of false alarms, a threshold is fixed so that some un-important slots are eliminated. Let l be the length of the given query phone sequence. The threshold is set to $0.5 \cdot l$, which means at least half of the common phones in a document slot are placed in the same position in the given query phone sequence. As showed in Figure 5.13, the slot marked with red colour will be ignored. The advantages of using the $error - tolerant$ method is that it can deal with all kinds of errors in the phone representation of a spoken document. The slots shown in this example are detected using the $error - tolerant$ method at $threshold = 0.43$. A document slot with an insertion error at position 3 (phone 'N' is inserted into representation), can also be detected and prepared for further processing.

5.4.2 Slot-probability Estimation

In order to assess the relevance of a detected slot regarding the given query, a probability is assigned for each document slot. This slot-probability indicates that the query phone sequence is uttered in slot s. This probability reflects the slot relevance to the given query phone sequence and will be used in the next stage to form the retrieval status value. The symbol q is used to represent the query phone sequence. The symbol s indicates a slot detected in document d. The task of the slot-probability estimation stage is to estimate $P(q \in s)$. We are going to discuss two of the slot-probability estimation methods in this section.

Edit Distance based Probability Estimation

This section presents an approximate pattern matching method presented by [6] (pp.105–110). This method is based on the edit distance between a document slot and the given query phone sequence. The edit distance is defined as the minimum number of local edit operations required to transform one object into another [24]. For example, the distance between /aU k s b o k/ and /aU k s o t k/ is 2, which means two steps are needed to transform the first phone sequence into the second one. These are the deletion of 'b' and the insertion of 't'. The similarity measurement between two phone sequences is recursively obtained using similarity measurements between shorter substrings. The edit distance between two phone sequences is defined as the minimal number of transformations that are required to transform one phone sequence into another. With respect to the document, three types of transformations can be made: substitution, insertion and deletion.

- substitution : a phone in q is substituted by another phone in s;
- insertion : a phone not present in q is inserted in s;
- deletion : a phone in q is deleted in s;

This algorithm is based on the dynamic programming principle. The edit distance between query phone sequence q and document slot s is obtained recursively as:

$$\delta(s_u, q_v) = \infty \text{ when } u < 0 \text{ and } v < 0 \tag{5.21}$$

$$\delta(s_0, q_0) = 0 \tag{5.22}$$

$$\delta(s_u, q_v) = min \begin{cases} t_1 \\ t_2 \\ t_3 \end{cases} \tag{5.23}$$

where t_1, t_2 and t_3 are estimated as:

$$\begin{cases} t_1 = \delta(s_{u-1}, q_{v-1}); & when\ s[u-1] == q[v-1] \\ \begin{cases} t_1 = \delta(s_{u-1}, q_{v-1}) + 1; \\ t_2 = \delta(s_{u-1}, q_v) + 1; \\ t_3 = \delta(s_u, q_{v-1}) + 1; \end{cases} & else \end{cases} \tag{5.24}$$

s_u is the sequence of the first u phones in slot s; q_v is the sequence of the first v phones in query q. We assume that all phones have an equal acoustic distance of 1 between each other. This edit distance is used to estimate the probability $P(q \in s)$ of query sequence q occurs in document slot s, by taking the simple inverse normalized edit distance (INED):

$$P(q \in s) = 1 - \frac{\delta(s_{l_s}, q_{l_q})}{\max\{l_s, l_q\}} \tag{5.25}$$

However, this assumption poorly reflects the real acoustic distance between the slot and the query. The slot-probability would be more accurate if we take into account the error characteristics and substitution probabilities of a phone recognizer.

Integrating Confusion Information into Probability Estimation

An alternative method for slot-probability estimation was developed by Wechsler ([115]) using confusion information that statistically expresses the error-production characteristics of a phone recognizer. The slot-probability is then estimated as follows:

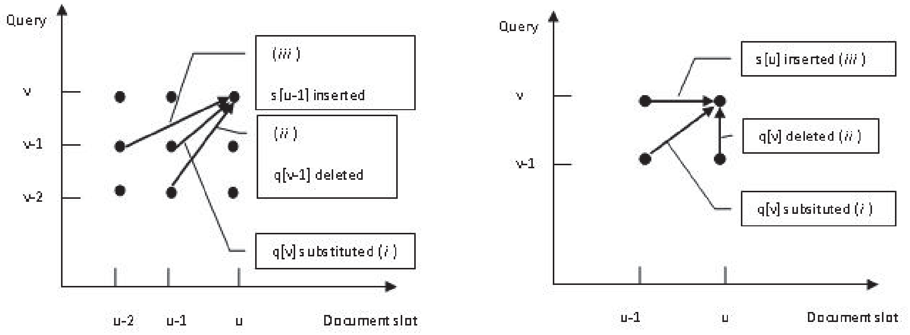
$$sim(s_0, q_v) \quad := \quad P_{sub}(q[v], s[0]); \tag{5.26}$$

$$sim(s_u, q_0) \quad := \quad P_{sub}(q[0], s[u]); \tag{5.27}$$

$$sim(s_u, q_v) = max \begin{cases} sim(s_{u-1}, q_{v-1}) + P_{sub}(q[v], s[u]); \\ sim(s_u, q_{v-1}) + P_{del}(q[v]); \\ sim(s_{u-1}, q_v) + P_{ins}(s[u]); \end{cases} \tag{5.28}$$

where S_u is the sub-string of the $u+1$ first phones of s (slot detected); q_v is the sub-string of the $v+1$ first phones of q (keyword phone sequence); $q[v]$ is the $v+1$ phone in q; $s[u]$ indicates the $(u+1)$ phone in detected slot; $P_{sub}(q[v], s[u])$ is the probability that phone $q[v]$ is substituted with phone $s[u]$; $P_{del}(q[v])$ indicates the probability that phone $q[v]$ is deleted; and $P_{ins}(s[u])$ is the probability that phone $s[u]$ is inserted.

The slot-probability estimation is implemented using dynamic programming (DP), as shown in Figure 5.14. The term $sim(s_u, q_v)$ denotes the string similarity between a document slot sub-string $\{s[0], ...s[u]\}$ and a sub-string $\{q[0], ...q[v]\}$ of query sequence. Three alternatives defined by the Equation 5.28 correspond respectively to the substitution, insertion and deletion. These three alternatives are illustrated as arrows in a two-dimensional grid defined by the document slot phone sequence (x-axis) and the query phone sequence (y-axis). Figure 5.14(b) shows a simpler and more straightforward recursive scheme used in the work described here.



(a) DP transitions used in Wechsler's work

(b) DP transitions used in this work

Figure 5.14: DP transitions

This method derives the slot-probability more accurately, as it takes into account the probabilities of all error types. But it requires more computation effort than the approach based on the edit distance.

5.4.3 Weighting and Scoring

Let d_j be jth documents in a collection and $q_i \in q$ be ith indexing feature of a query q. The similarity between document d_j and query q is obtained as the inner product between corresponding weighted term vectors. It was proposed by Wechsler ([116]):

$$Sim(q, d_j) = \sum_{q_i \in q} DW_{i,j} QW_i. \tag{5.29}$$

where $DW_{i,j}$ denotes the document weight of q_i in d_j and is defined by equation 5.30; QW_i denotes the query weight of q_i in q and is computed as follows:

$$\begin{aligned} DW_{i,j} &= \frac{1}{(1-\alpha)\bar{l} + \alpha l_{d_j}} \log(1 + eff(q_i, d_j)) && (5.30) \\ eff(q_i, d_j) &= \sum_{s_u \in d_j} sim(q_i, s_u) && (5.31) \end{aligned}$$

where the length of d_j is denoted as l_{d_j} and $\bar{l}$ denotes the average document length in the collection. s_u represents detected slots in document d_j. α [106] is the slope, set to 0.25.

$$\begin{aligned} QW_i &= (1 + log(ff(q_i, q)))iecf(q_i) && (5.32) \\ iecf(q_i) &= 1 + log(\frac{C_q + 1}{ecf(q_i) + 1}) && (5.33) \\ ecf(q_i) &= \sum_{d_j \in D} eff(q_i, d_j) && (5.34) \\ C_q &= max_{q_i \in q}(ecf(q_i)) && (5.35) \end{aligned}$$

Where the number of expected occurrences of q_i in a document d_j (denoted as $eff(q_i, d_j)$) is written as the sum of slot-probabilities [76]. $iecf$ represents inverse expected collection frequency, which is very similar to inverse document frequency. ecf represents the expected number of occurrences of a feature (term q_i). C_q is a query-specific constant. The slot-probabilities are estimated using the method introduced in the previous sections. $ff(q_i, q)$ denotes the number of the occurrences of q_i in q.

5.4.4 Experiment and Discussion

In this section, we are going to evaluate the monophone-based spoken document retrieval approaches using different slot-probability estimation methods and explore the impact of query-length on the monophone-based spoken document retrieval methods.

Comparison of Different Slot Probability Estimation Methods

In this study, we first evaluate the retrieval performance of the monophone-based spoken document retrieval approaches with the in-vocabulary query retrieval task. Figure 5.15 presents the experimental results. The label *baseline* represents our word-based baseline spoken document retrieval system in this work. The label $INED$ denotes the monophone-based spoken document retrieval system that assigns the inverse normalized edit distance between a document slot and the given query phone sequence to the slot probability. The label $SSPE$ (string similarity based probability estimation) indicates the monophone-based spoken document retrieval system, which integrates the confusion information into the slot probability (described in Section 5.4.2).

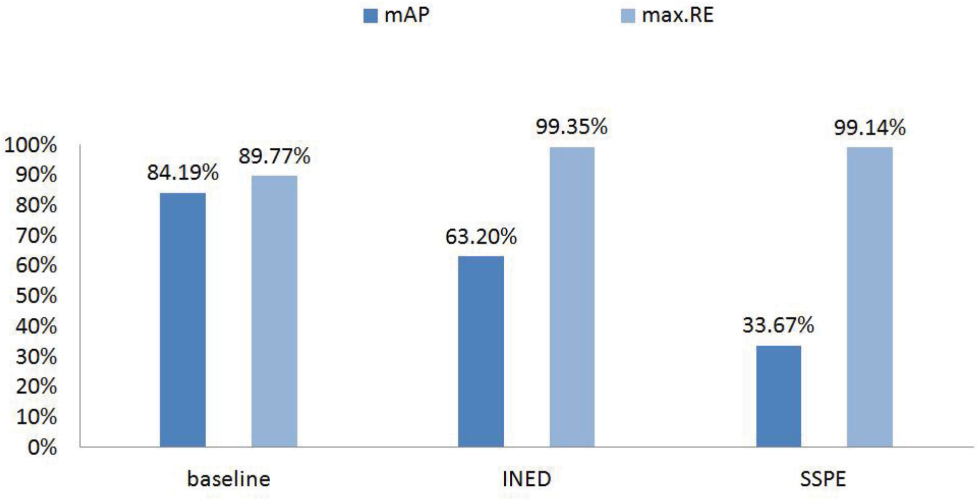


Figure 5.15: Performance (mAP) evaluation of different mono-phone based SDR systems with in-vocabulary query.

A high retrieval recall rate can be observed by the monophone-based spoken document retrieval system. In comparison with the baseline system, the $INED$ method improves the maximal recall by about 9.58%. The mean average precision yielded by the $INED$ method decreases by about 20.89%. The $SSPE$ method improves the maximal recall rate of SDR system (9.25%) compared with the baseline system. Because the $SSPE$ method performs approximate matching between the query phone sequence and a document slot, it introduces a number of false alarms into the retrieved result. Similar behavior can also be viewed in Table 5.19.

	baseline		INED			SSPE		
recall	INV	TOTAL	INV	OOV	TOTAL	INV	OOV	TOTAL
0	86.00	74.78	79.00	53.33	75.65	16.50	23.33	17.30
10	89.12	77.60	73.83	52.22	70.43	35.25	21.25	31.90
20	89.93	78.08	73.42	47.62	68.94	41.60	21.87	37.50
30	90.12	78.18	73.85	40.77	68.09	44.60	15.14	40.60
40	89.59	77.70	73.51	30.87	66.87	45.80	3.40	41.06
50	89.05	77.08	72.76	12.50	65.44	44.90	1.90	39.10
60	88.48	76.66	69.58	5.17	61.79	40.60	1.00	33.80
70	88.10	75.89	63.16	1.59	53.71	33.68	0.93	26.78
80	87.30	-	49.21	0.63	38.99	24.73	0.53	18.05
90	86.98	-	25.99	0.24	9.48	14.12	0.23	6.19
mAP	84.17	61.67	63.20	22.92	56.20	33.67	8.53	29.07
max.RE	89.77	78.06	99.35	91.72	98.35	99.14	91.05	97.90

Table 5.19: Performance of different monophone-based SDR systems in precision, recall, mAP and max. recall.

Next, we evaluate different monophone spoken document retrieval methods with the complete query set (200 in-vocabulary and 30 out-of-vocabulary queries). Figure 5.16 presents the experimental results. From the data in Figure 5.16, it is apparent that monophone-based spoken document retrieval method has significantly improved the maximal recall rate. Compared with the *baseline* system, a slightly improved retrieval precision (+0.87%) at lower recall level can be observed by the $INED$ method.

The Effect of Query-length

In this study, we are going to research the impact of query-length on the retrieval performance of a monophone-based spoken document system. Table 5.9 presents the words selected for this selection task. Again, separate experiments are made to evaluate the impact of the query-length on the in-vocabulary and out-of-vocabulary query retrieval performance.

The results of the in-vocabulary query retrieval task are presented in Figure 5.17 and Table 5.20. We can see that the monophone-based $INED$ and $SSPE$ methods outperform the baseline system when long queries are retrieved. For queries with a length of $\geq$ 10, a maximal recall of 100% is yielded by phone-based $INED$ or $SSPE$. The mean average precision achieved by phone-based $INED$ drops slightly in comparison with the baseline system. However, the mean average precison value

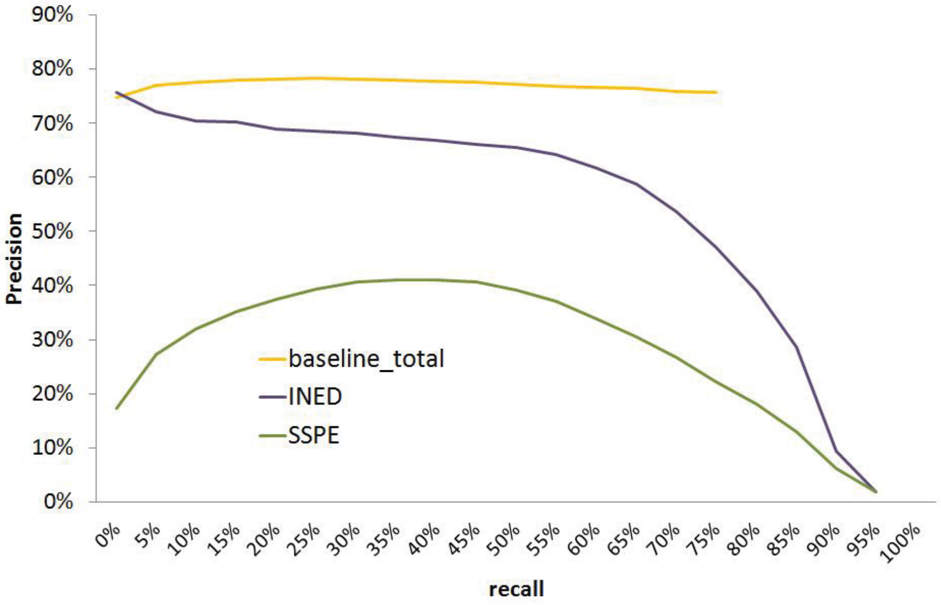


Figure 5.16: Precision/recall plot of different mono-phone based SDR systems with total query set.

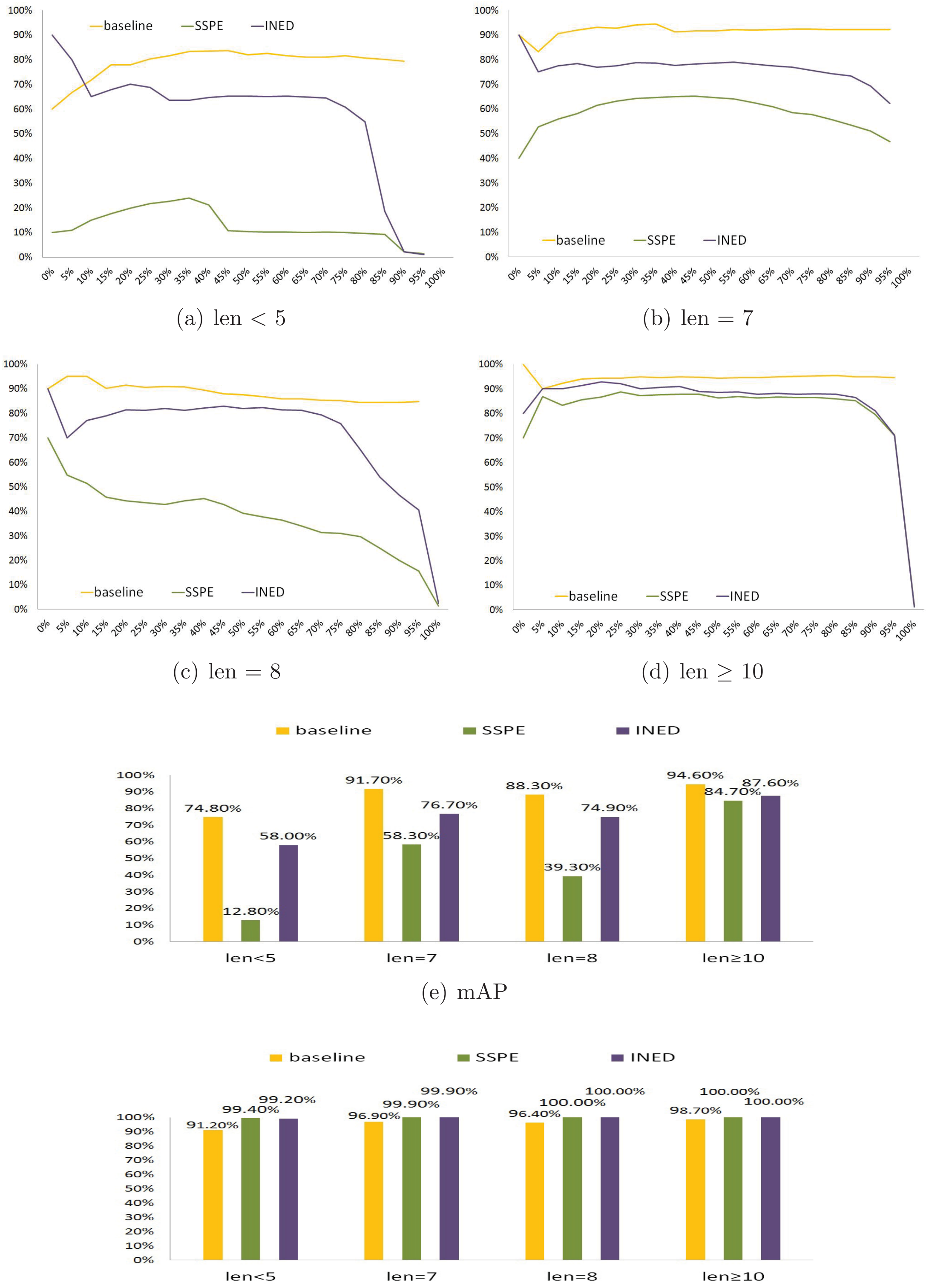
of phone-based methods drops dramatically with decreasing query-length. It can be observed that the $SSPE$ method is extremely sensitive to the query-length. For very short queries with a length of < 5, the $SSPE$ method yields only a mean average precision of 12.8%. The $INED$ method acts better in retrieving the short queries. For queries with a length of < 5, the $INED$ method reaches a mean average precision value of 58%. The $SSPE$ achieves comparable retrieval performance to the $INED$ method in retrieving long queries (length ≥ 10).

	len < 5		len = 7		len = 8		len ≥ 10	
	mAP	maxRE	mAP	maxRE	mAP	maxRE	mAP	maxRE
baseline	74.80	91.20	91.70	96.90	88.30	96.40	94.60	98.70
SSPE	12.80	99.40	58.30	99.90	39.30	100.00	84.70	100.00
INED	58.00	99.20	76.70	99.90	74.90	100.00	87.60	100.00

Table 5.20: Performance of retrieving in-vocabulary queries of different lengths using phone-based SDR.

In the case of out-of-vocabulary query retrieval, the experimental results are presented in Figure 5.18 and Table 5.21. As we can see, the mean average retrieval precision of both monophone-based approaches increases with growing query length. The $SSPE$ method yields poor precision when retrieving short queries.

We compare the monophone-based spoken document retrieval methods and the



(a) len < 5

(b) len $= 7$

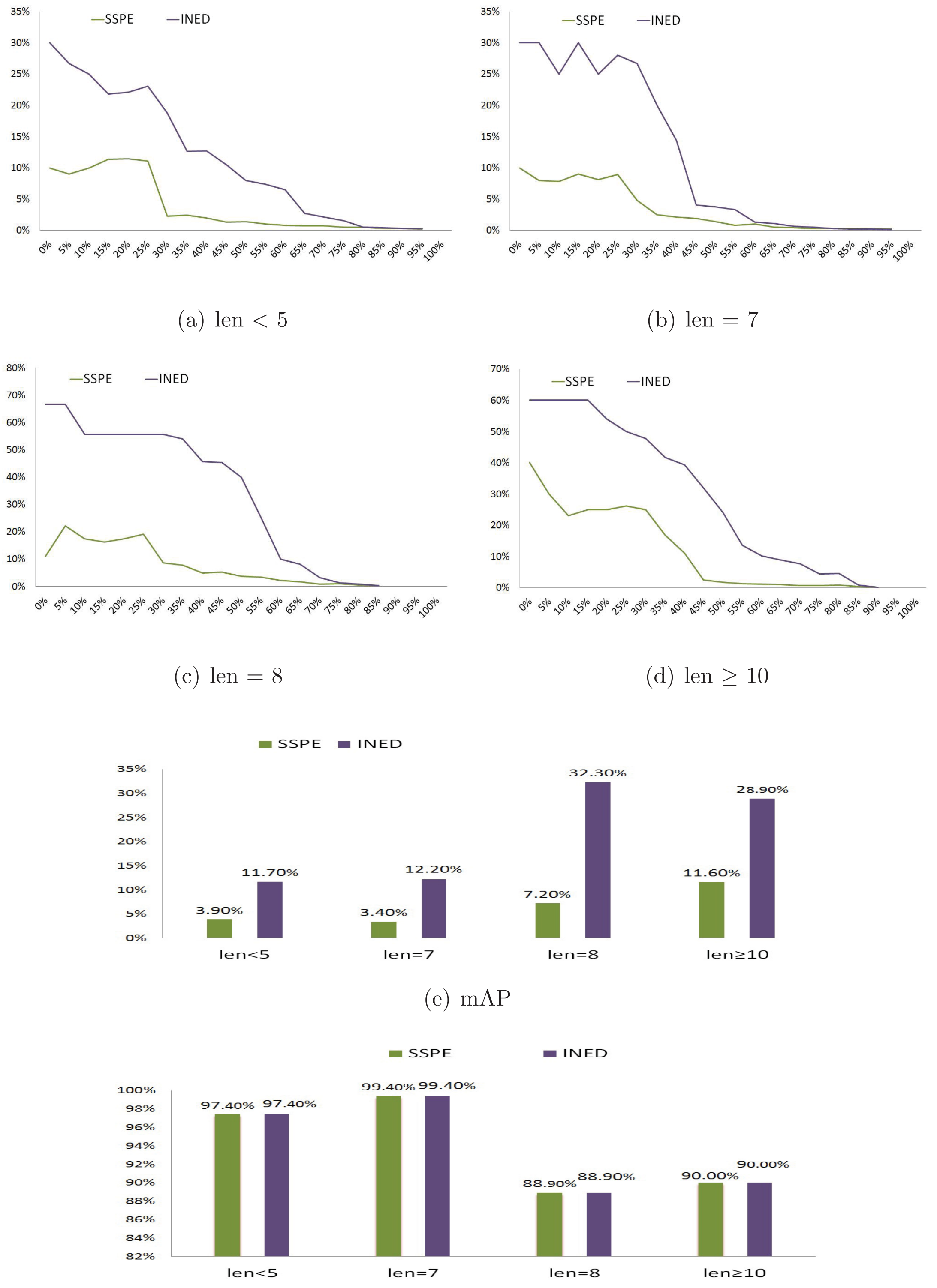
(c) len $= 8$

(d) len ≥ 10

(e) mAP

(f) max. recall

Figure 5.17: Performance (precision/recall plot) of retrieving in-vocabulary queries of different lengths using mono-phone based SDR method.



(a) len < 5

(b) len $= 7$

(c) len $= 8$

(d) len ≥ 10

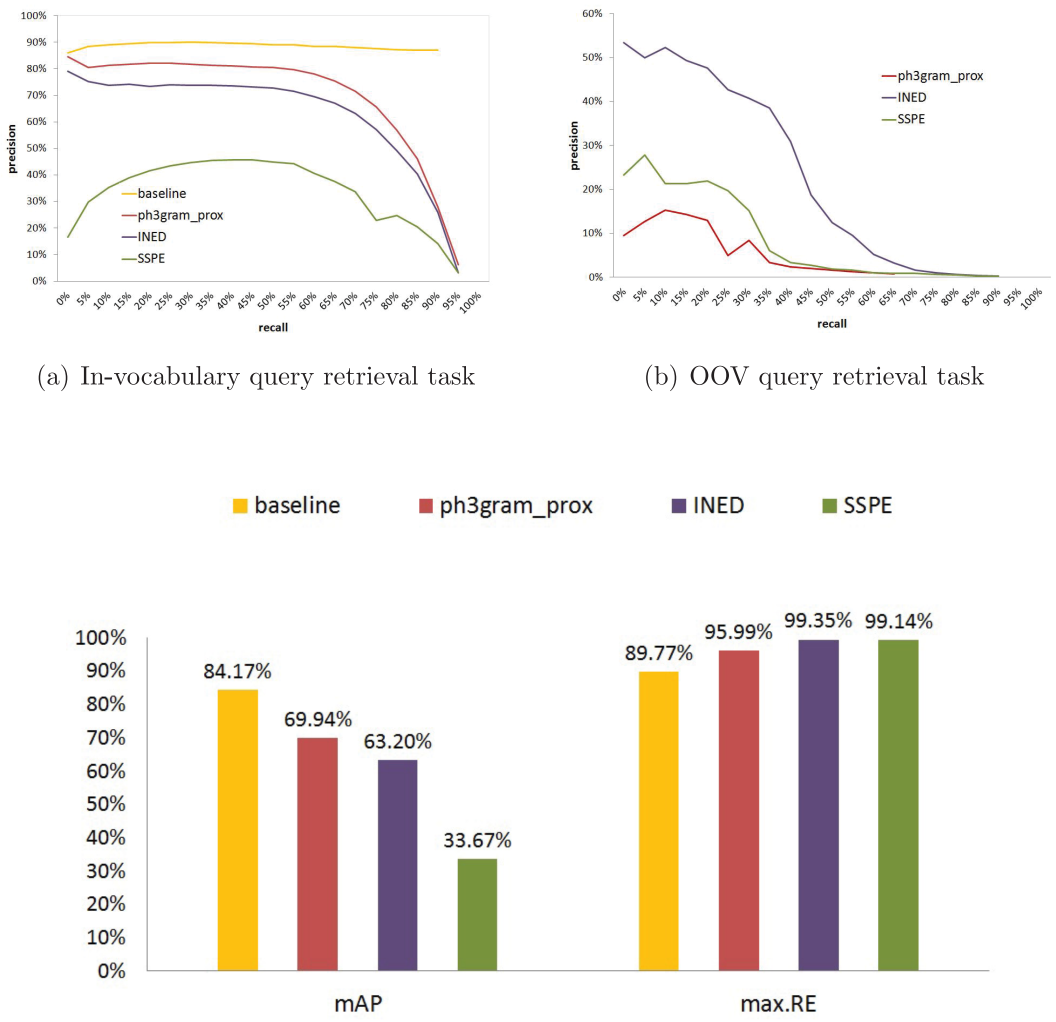
(e) mAP

(f) max. recall

Figure 5.18: Performance (precision/recall plot) of retrieving OOV queries with different length using mono-phone based method.

	len < 5		len = 7		len = 8		len ≥ 10	
	mAP	maxRE	mAP	maxRE	mAP	maxRE	mAP	maxRE
SSPE	3.90	97.40	3.40	99.40	7.20	88.90	11.60	90.00
INED	11.70	97.40	12.20	99.40	32.30	88.90	28.90	90.00

Table 5.21: Performance of retrieving OOV queries of different lengths using monophone-based SDR.



(a) In-vocabulary query retrieval task

(b) OOV query retrieval task

(c) INV query set

Figure 5.19: Comparison of different subword-based SDR approaches.

phone-3gram based *proximity* method (called *ph3gram_prox* in the following part of this section). The complete query set is selected. This set of queries includes 13% out-of-vocabulary words (similar to the real application scenario). Comparing the results shown in Figure 5.19 and Table 5.22, we can see that:

- The $INED$ and the $SSPE$ methods further improve the maximal recall rate. Compared with the *ph3gram_prox* method, the $INED$ method improves the maximal recall by about 3.36%.

- The $INED$ method yields a mean average precision slightly worse than the *ph3gram_prox* method, while the mean average precision of the $SSPE$ approach drops dramatically.

- The $INED$ methods show their advantage in dealing with out-of-vocabulary queries.

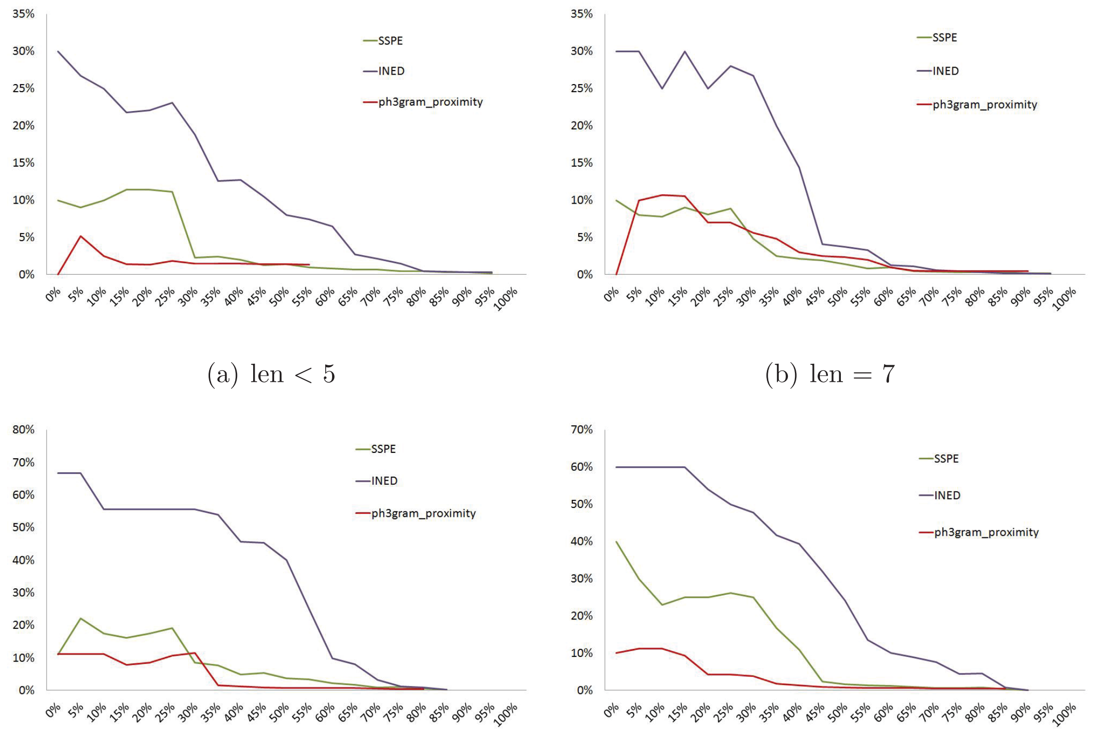
- From the data in Figure 5.20, it is apparent that subword-based spoken document retrieval methods can deal with out-of-vocabulary queries. The $INED$ and $SSPE$ methods outperform the *ph3gram_prox* method. The best performance of retrieving very short query (length < 5) is achieved using the $INED$ method.

	baseline		ph3gram_prox			INED			SSPE		
recall	INV	TOTAL	INV	OOV	TOTAL	INV	OOV	TOTAL	INV	OOV	TOTAL
0	86.00	74.78	84.50	9.50	77.38	79.00	53.33	75.65	16.50	23.33	17.30
10	89.12	77.60	81.30	15.30	75.06	73.83	52.22	70.43	35.25	21.25	31.90
20	89.93	78.08	82.22	13.00	75.70	73.42	47.62	68.94	41.60	21.87	37.50
30	90.12	78.18	81.75	8.40	75.04	73.85	40.77	68.09	44.60	15.14	40.60
40	89.59	77.70	81.05	2.40	74.17	73.51	30.87	66.87	45.80	3.40	41.06
50	89.05	77.08	80.40	1.60	73.02	72.76	12.50	65.44	44.90	1.90	39.10
60	88.48	76.66	78.09	1.00	69.60	69.58	5.17	61.79	40.60	1.00	33.80
70	88.10	75.89	71.49		60.16	63.16	1.59	53.71	33.68	0.93	26.78
80	87.30		56.98		42.21	49.21	0.63	38.99	24.73	0.53	18.05
90	86.98		27.91		9.74	25.99	0.24	9.48	14.12	0.23	6.19
mAP	84.17	61.67	69.94	4.54	61.08	63.20	22.92	56.20	33.67	8.53	29.01
max.RE	89.77	78.06	95.99	68.85	93.40	99.35	91.72	98.35	99.14	91.05	94.80

Table 5.22: Comparison of the phone-based spoken document retrieval methods and the word-based baseline system

Conclusions

The experimental results in this study confirm that the monophone-based $SSPE$ and $INED$ methods improve the retrieval recall significantly, compared with the baseline system. The $INED$ method achieves a better mean average precision for the short (with 4 phones) and long query (with ≥ 10 phones) retrieval tasks. The $SSPE$ method



(a) len < 5

(b) len = 7

(c) len = 8

(d) len ≥ 10

Figure 5.20: Comparison of ability of phone-based SDR methods to deal with OOV-queries.

is very sensitive to query length. The $INED$ approach provides more reliable performance for the in-vocabulary queries.In comparison with the $ph3gram\_prox$ method, the retrieval recall rate is further improved using the monophone-based methods. The monophone-based $INED$ method can yield a mean average precision that is comparable to the $ph3gram\_prox$ method.

5.5 Summary

This chapter focuses on an exploration of the subword-based spoken document retrieval methods. First, we provide an overview about different sub-word units such as morphemes, syllables (VCV feature), phone sequences and monophones. Then we carry out a feasibility study into applying different sub-word units for classical text information retrieval task. This study has shown that smaller index units like monophone provides better information coverage. After this study, the phone-based spoken document retrieval methods will be thoroughly investigated.

The study of the phone-Ngram based SDR method has shown that the phone-3gram based spoken document retrieval provides reliable retrieval performance. The recall has been significantly improved. A clear benefit of integrating phone confusion information into term weighting (the $confusion$ method) cannot be identified in this analysis. The best mean average precision is yielded by the *proximity* method. The experimental results show that phone-3gram based spoken document retrieval methods are not suitable for the out-of-vocabulary query-tasks. Adding multiple recognition hypotheses into the phone-3gram representation of a spoken document in the form of a position specific posterior lattice (PSPL) does not benefit the retrieval performance. The query-length makes an large impact on the retrieval performance. The mean average precision increases with a growing query-length. The best recall is provided by the $confusion$ method, which integrates phone confusion information into term-weighting. Additional statistical significance tests are conducted to ensure the reliability of the conclusions made about the effect of query-length. In this work, normally distributed t-test statistic is applied to check the statistical significance of the statements. Some other significance tests can also be applied to ensure the reliability of a statement, for example, the NIST significance tests, that are not based on the normally distributed data. The NIST significance tests [2] take the value range and the way of the origin of the results into account. However, it is out of the scope of this dissertation.

The investigation of the monophone-based spoken document retrieval methods has identified that compared with the baseline system, the retrieval recall rate is improved significantly with monophone-based methods. The $INED$ method achieves a better mean average precision than the $SSPE$ method. The $SSPE$ method is sensitive to query-length.

In comparison with the phone-3gram based spoken document retrieval method, the retrieval recall rate can be further improved using the monophone-based methods. The monophone-based $INED$ method can yield a mean average precision comparable to those achieved by the $ph3gram\_prox$ method.

Chapter 6

A new Robust Open-Vocabulary Hybrid Spoken Document Retrieval Method

According to the type of recognizer used to transcribe the spoken document, the current spoken document retrieval methods can be grouped into four main categories: word-based approaches, subword-based approaches, phone-based approaches and combined approaches. Word-based approaches [123] rely on a large vocabulary automatic speech recognition (ASR) system that transcribes speech into a word sequence. The text string matching algorithms can then be directly used to find the required information. It has been verified that the word-based spoken document retrieval system can achieve comparable performance to the text information retrieval, if the word error rate of the applied recognizer is not too high. However, the queries are restricted by the size of the recognizable vocabulary. Logan [67] reported that about 13% of user queries contain out-of-vocabulary (OOV) words. Moreover, out-of-vocabulary words pose a serious problem in a word-based spoken document retrieval system, particularly to the domains where new words appear frequently over a short period of time.

As we mentioned before, most of the current spoken document retrieval approaches are based on the word-1best representation of a spoken document. This word-1best representation is provided by an automatic word recognizer. The mis-recognition problem may lead to the loss of important information. In order to keep more useful information contained by a spoken document, multiple recognition-hypotheses are added to the transcription of a spoken document. We have thoroughly investigated this kind of word-based spoken document retrieval methods in the Chapter 4. The result of this study indicates that performing retrieval on the word confusion network transcription of a spoken document achieves maximum information coverage. Generally, the word-based spoken document retrieval methods cannot deal with out-of-vocabulary words. The representation expansion method enables the word-based spoken document retrieval methods to work with out-of-vocabulary words. However, there is no obvious improvement in retrieval performance (4% in precision and 1.4% in recall).

Chapter 5 focused on the investigation of another solution on how to deal with out-of-vocabulary words; namely, retrieving spoken document based on its subword representation. In this study, various subword-based spoken document retrieval methods are researched. The results of this study show that indexing spoken documents with the phones (monophone or phone-3gram) can achieve the maximum information coverage. The phone-based methods have significantly improved the retrieval recall. However, the mean average precision will drop. Integrating the proximity information into the document ranking score improves the mean average precision. The query length has a great impact on the retrieval performance of the phone-based spoken document retrieval methods. The experimental results indicate that monophone-based spoken document retrieval methods achieve a better performance on the out-of-vocabulary query retrieval task. The monophone-based $SSPE$ method, which weights the document slots with the phone confusion information, can deal with out-of-vocabulary queries but with restrictive performance. The monophone-based $INED$ method, which weights document slots with edit distance, provides better performance when a short query (< 5 phones) or a very long query (> 11 phones) is retrieved.

The results of the previous investigation show that the word-based spoken document retrieval methods provides the best performance for the in-vocabulary query retrieval task, while phone-based spoken document retrieval methods show their advantages in the out-of-vocabulary query retrieval task. These findings suggest that the fusion of different indexing units can take advantage of both phone-based and word-based spoken document retrieval methods. Several attempts have been made to combine different indexing sources for robust spoken document retrieval ([51], [83]). James [51] combined word and phone recognition in a complete recognition system, in which the phone recognizer is only used to spot out-of-vocabulary words. This combination improves the retrieval effectiveness of a spoken document retrieval system. However, a large amount of annotated training data and more effort is required to build two recognizers. Similar approaches proposed by Amir et. al. [4] using metaphone (broad phone class) instead monophone to improve the monophone-based retrieval performance. In his paper, phones are grouped into 7 broad metaphones based on their confusion information. The monophone representation of a spoken document was updated into metaphone representation. Amir et. al. reported slight improvement in precision compared to word-based approaches was achieved by the method combining word-based and metaphone-based SDR methods. Ng's investigation verified that different subword unit representations can capture different types of information, and combining different types of subword units indexing terms results in better performance.

In this chapter, we are going to introduce a novel efficient hybrid spoken document retrieval method, which has been developed based on our research results in the previous chapters. This novel spoken document retrieval system enables open-vocabulary query retrieval and is robust against the recognition errors in the transcription of a spoken document. Section 6.1 focuses on the selection of the potential fusion- candidates. We describe the novel spoken document retrieval method in Section 6.2. Finally, we evaluate the performance of this novel spoken document retrieval method on our spoken document test collection.

6.1 Information Fusion

6.1.1 Background

The concept of information fusion in the field of spoken document retrieval means combining different information resources. The fusion of indexing units like phone and word enables the spoken document retrieval system to take advantage of both phone- and word-based spoken document retrieval methods. In Chapter 4 and Chapter 5, we explored a number of word-based and phone-based spoken document retrieval systems. In order to achieve an efficient fusion, it is essential first to find the potential candidates for fusion. The potential candidates for fusion are selected by analyzing the retrieved lists. Lee's idea [64] is applied. Lee indicated that the overlaps between different result sets provided by different retrieval methods are an important factor. He defined the overlap ratio of relevant (R)/non-relevant(N) document in the result sets A and B as:

$$R = \frac{R_{common} \times 2}{R_A + R_B} \tag{6.1}$$

$$N = \frac{N_{common} \times 2}{N_A + N_B} \tag{6.2}$$

Where R_{common} indicates the number of relevant documents that occur in both result sets. N_{common} means the number of non-relevant documents that occur in both result sets. The results of Lee's study has confirmed that a well joined result set must have a high R and a low N. Vogtś experiments [114] on the TREC-5 data verified that an effective fusion can only be reached when:

- At least one result has high precision/recall.
- High overlap of relevant documents and low overlap of non-relevant documents.
- Relevance score is similarly distributed.
- Each system ranks the documents relevant to the given query, differently.

A number of methods have been published for the information fusion. Thompson believes that a retrieval system should be considered more preferably than others, when its prior performance is better than the others. He assigned each ranked list a variable weight based on the prior performance of the system. A similar idea was presented in Bartell's paper [7]. He determined the optimal scalars for a linear combination of results from the training data using numerical optimization techniques.

Fox [34] fused result sets by combining the evidence from the multiple retrieval runs. Document-query similarities in different sets are treated as the evidence from different retrieval runs. Five combining strategies were investigated in his work, $CombMAX$, $CombMin$, $CombSUM$, $CombANZ$ and $CombMNZ$. $CombMAX$ sets the combined similarity to the maximum of the individual similarities. On the other hand, the combined similarity of $CombMin$ is the minimum of individual similarities. The $CombSUM$ method sums up individual similarities to form a new score. $CombANZ$ normalizes combined similarity provided by $CombSUM$ to the number of non-zero similarities, while $CombMNZ$ multiplies the $CombSUM$ score with the number of non-zero similarities.

6.1.2 Selection of Potential Candidates for Fusion

The results from our previous experiments in Chapter 4 have showed that indexing spoken document with a word confusion network and weighting document term with term-frequency ($tfidf$) is the best word-based spoken document retrieval method. This method achieves a 95.23% recall, while maintaining a reliable mean average precision (64.92%). This word-based spoken document retrieval method is denoted as WCN in the following sections.

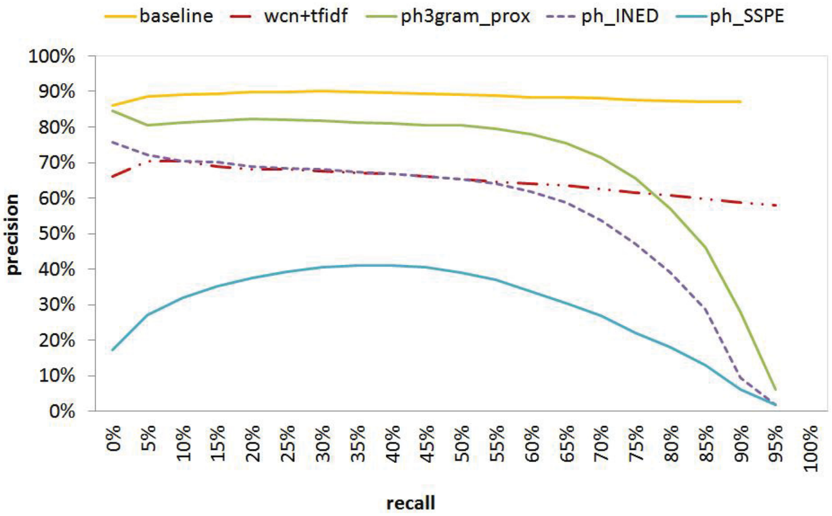
The studies in Chapter 5 have demonstrated the ability of phone-based spoken document retrieval methods to deal with out-of-vocabulary words. The results of these studies have also verified their ability to compensate the recognition errors in the phone transcription of a spoken document. We compare the phone-based spoken document methods using an in-vocabulary and out-of-vocabulary query set. Figure 6.1 presents the results of this study. The results of the in-vocabulary evaluation task are presented in Figure 6.1(a). We can see that the WCN method achieves a higher retrieval precision at a high recall level. The $ph3gram\_prox$ and ph_INED methods improve the retrieval precision at a lower recall level. The ph_INED and ph_SSPE methods show their advantages in dealing with out-of-vocabulary query words.

The phone-based and word-based spoken document retrieval methods estimate the document ranking score in different way. Therefore, according to Vogt's 4^{th} condition for a perfect fusion, by an effective fusion of the phone-based and word-based methods, the performance of a spoken document retrieval should be improved. We measure the overlap ratio of the non-relevant documents in the result sets provided using different phone-based and word-based spoken document retrieval methods, and look for the best fusion candidates. The results of this study are presented in Table 6.1.

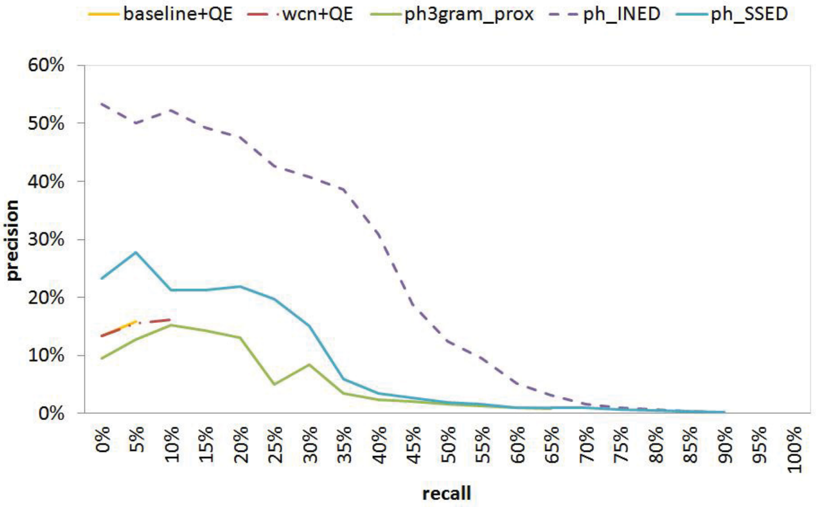
	wcn+tfidf	ph3gram_prox	ph_INED	ph_SSPE
wcn+tfidf	1.0000	0.8040	0.0818	0.0577
ph3gram_prox	0.8040	1.0000	0.1043	0.0760
ph_INED	0.0818	0.1043	1.0000	0.6940
ph_SSPE	0.0577	0.0760	0.6940	1.0000

Table 6.1: Overlap ratio of non-relevant documents retrieved by different spoken document retrieval system (in-vocabulary query set).

From the data in Table 6.1, supported by the Vogt's 2^nd condition for an effective fusion, the word-based WCN method and the phone-based ph_SSPE (denoted as $SSPE$ in the following part of this section) method are identified as the best potential



(a) Performance with in-vocabulary queries



(b) Performance with OOV queries

Figure 6.1: Precision/recall plot of potential candidates for fusion.

fusion pair, as the overlap ratio of non-relevant documents in the result sets has the minimal value. A similar overlap ratio of non-relevant documents in the result sets can also be observed by the fusion of WCN and ph_INED (denoted as $INED$ in the following part of this section) method.

In addition, the word-based WCN method achieves a high performance in precision/recall by the in-vocabulary query retrieval task, which satisfies the first requirement of the Vogt's theory for a potential 'effective' fusion.

6.2 A Hybrid Keyword Spotting Method for Spoken Document Retrieval

In the last section, we identified two effective fusion pairs. Base on this issue, we propose a novel hybrid spoken document retrieval methods that combines the word-based WCN method and the monophone-based methods ($INED$, $SSPE$) in a sufficient way for more robust spoken document retrieval. The fusion of word-based WCN and phone-based $SSPE$ method is referred to as WCN+$SSPE$ in the following sections. WCN+ $INED$ indicates the fusion of word-based WCN and phone-based $INED$ method. The novel hybrid spoken document retrieval is presented in Figure 6.2.

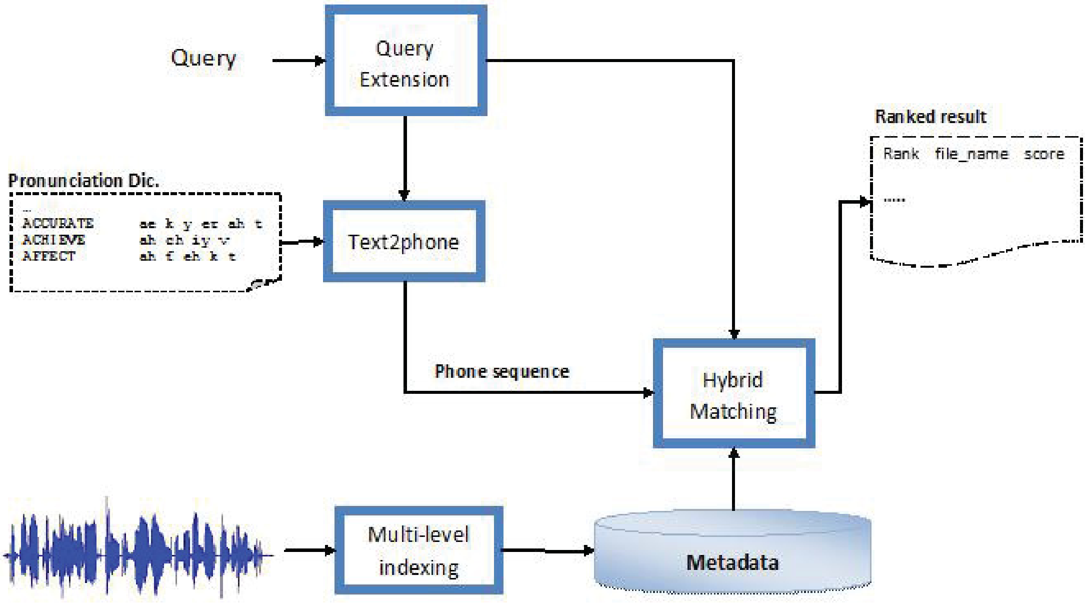
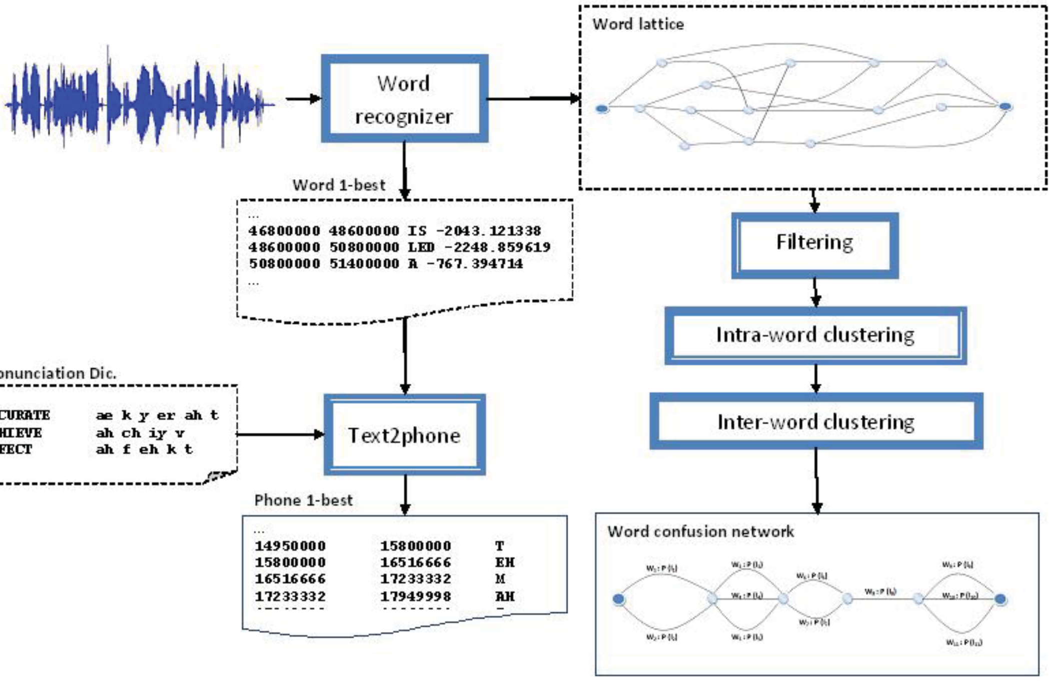


Figure 6.2: Structure of Hybrid SDR system

Figure 6.2 displays this novel hybrid spoken document retrieval system, which consists of three main components: multi-level indexing, query extension and hybrid matching. In the multi-level indexing stage, a multi-level representation is automatically produced for every spoken document in a collection. In the query extension stage, a monophone sequence for queries is prepared. The hybrid matching module performs query-detection in the multi-level representation of a spoken document. We describe this new hybrid spoken document retrieval method in more detail in the following sections.

6.2.1 Multi-level Indexing of a Spoken Document

Combining multi-level information (phone one-best, and WCN) should improve the performance of retrieving both in-vocabulary and out-of-vocabulary queries. The advantages of multi-level indexing of a spoken document have been verified in the recent works [48]. This paper reported that the combination of word and phone confusion networks is effective and yields high retrieval performance for both in-vocabulary and out-of-vocabulary queries. However, a large amount of annotated training data is required for building a reliable word and phone recognizer. In the last section, we discussed that a more robust spoken document retrieval can be achieved through the fusion of word-based WCN method and phone-based SDR methods($INED$,$SSPE$). Therefore, in the multi-level indexing stage, we will prepare two representations for a spoken document in a collection; one in form of a word confusion network and another one consisting of a monophone sequence.



```
...
46800000 48600000 IS -2043.121338
48600000 50800000 LED -2248.859619
50800000 51400000 A -767.394714
...
```

```
...
ACCURATE      ae k y er ah t
ACHIEVE       ah ch iy v
AFFECT        ah f eh k t
```

```
...
14950000       15800000      T
15800000       16516666      EH
16516666       17233332      M
17233332       17949998      AH
--------       --------      -
```

Figure 6.3: Hybrid system - Multi-level indexing stage

In order to build an efficient open-vocabulary spoken document retrieval system, a $20k$ word recognizer was built to transcribe the spoken information. As shown in Figure

6.3, with help of a pronunciation dictionary, the monophone representation of a spoken document is directly extracted from the 1best output of the word recognizer. The word confusion network representation of a spoken document is constructed by post-processing the lattice output of the word recognizer [71]. The word confusion network representation of a spoken document is constructed step by step. First, the posterior probability is computed for all links in the word lattice; Then the edges with posterior probability far below 1.0 are eliminated; The edges with the same word instance with time-overlap are grouped into one cluster. The cluster posterior probability is assigned to the sum of all clustered edges' posteriors. Finally, the edges with similar phonetic properties and different word instances, which compete around the same time interval with similar phonetic properties will be grouped together.

6.2.2 Hybrid Matching

Different matching strategies are used for the detection of queries in the multi-level transcription of a spoken document. As we discussed before, the monophone-based methods ($INED$, $SSPE$) are applied for the term matching at the phone level. The word-based WCN is applied for information retrieval at word level.

Modified Vector Space Model based Term Matching in a Word Confusion Network

We modify the term-frequency based weighting scheme ($tfidf$). In the work of Mamou ([70]), the term frequency $tf(t,d)$ is computed as follows:

$$tf(t,d) = \sum_{i=1}^{|occs(t,d)|} B_{rank(t|o_i,d)} \times P(t|o_i,d) \tag{6.3}$$

where $occs(t,d) = o_1, o_2, ...o_n$ denotes the sequence of all occurrences of t in d; $rank(t|o,D)$ denotes the rank of term t at offset o where all hypotheses at offset o are sorted in decreasing order according to their posterior probabilities; $B_{rank(t|o_i,d)}$ associates weight factor to each rank of the different hypotheses; Mamou ([70]) introduced tree configurations for the boosting vector B. The second configuration of the boosting vector proposed by Mamou is applied in this work, namely that the rank of a term is ignored by setting every element in B to 1. $P(t|o_i,d)$ is the posterior probability of a term t

at offset o in the word confusion network of document D. idf is re-estimated as:

$$idf = log\frac{O}{O_t} \tag{6.4}$$

where the number of occurrences of term t is represented by O_t:

$$O_t = \sum_{d \in D} \sum_{i=1}^{|occs(t,d)|} P(t|o_i, d); \tag{6.5}$$

O reflects the total number of occurrence in the corpus:

$$O = \sum_{t \in D} O_t \tag{6.6}$$

And the relevance score is computed with Equation 2.22.

Probabilistic Phone Matching

The probabilistic phone matching algorithm $SSPE$ is applied for query detection at the phone level. This matching method consists of slot detection and slot weighting. It is assumed that the most errors in the monophone transcription of a spoken document are substitution errors. Slots that have a sufficient conformity with the query phone sequence are selected. The conformity is measured as the number of common phones that is defined as the same phone occurring at the same position in the query phone sequence and the document slots. A slot is verified when its number of common phones is greater than the pre-defined threshold value. The slot probability is estimated using the method introduced in section 5.4.

Finally, the results provided by the phone-based SDR methods ($INED$, $SSPE$) and the WCN method are fused at the score level. The combined score will be used to rank documents in the retrieved result. Three combination strategies are applied for the estimation of the combined ranking of a retrieved spoken document. These are the $CombMax$, $CombSum$ and $CombANZ$ method, described in Section 6.1.1.

6.3 Experiments and Discussion

In this section, we evaluate the new hybrid spoken document retrieval method on our spoken document test collection. The objectives of these experiments are to investigate:

- whether performance can be improved by combining the word confusion network matching method and the monophone sequence based probabilistic string matching techniques;
- which combining strategy from $CombMax$, $CombSum$ and $CombANZ$ performs better.

We will compare the performance of our new hybrid spoken document retrieval methods ($WCN + INED$ and $WCN + SSPE$) to the performance of the WCN and the phone-based spoken document retrieval methods ($INED$, $SSPE$). Three combining strategies are also evaluated. We use the suffixes $\_CombSum$, $\_CombMax$ and $\_CombANZ$ to represent the combining strategy used by the hybrid system. As the relevance score of documents returned by the WCN and phone-based SDR methods lies in different range, the relevance scores are normalized to their maximum value before they are fused together. The experimental results are displayed in Figure 6.4, Table 6.2 and Table 6.3.

In comparison with the WCN method in Figure 6.4(b) in the in-vocabulary query retrieval evaluation task, the mean average retrieval precision is improved by combining the retrieval score provided by the WCN method and by the $SSPE$ and $CombMax$ methods. The recall rate is kept at the same level. Combining the WCN and $SSPE$ methods with $CombANZ$ improves the retrieval recall rate. However, the mean average retrieval precision drops.

The results of the in-vocabulary query retrieval task are presented in Table 6.2 and Figure 6.4(b). We can see that phone-based methods ($SSPE$, $INED$) reach a higher recall rate. The word-based WCN method achieves better precision at higher recall levels. This result can be observed in the $recall = 90\%$ level in Table 6.2. The benefits of the fusion of phone-based methods and word-based WCN method are identified. The best mean average precision value (65.01%) is reached by $WCN + SSPE\_CombMax$. The $WCN + INED\_CombANZ$ method reaches the highest recall rate (99.98%).

				WCN+SSPE			WCN+INED		
recall	SSPE	INED	WCN	CombMax	CombSum	CombANZ	CombMax	CombSum	CombANZ
0	16.50	79.00	66.00	66.00	66.00	66.00	66.00	66.00	66.00
10	35.25	73.83	70.30	70.30	68.65	70.51	69.75	68.54	69.75
20	41.60	73.42	68.12	68.93	67.28	69.10	67.76	66.32	67.76
30	44.60	73.85	67.58	67.58	66.00	67.71	66.71	64.82	66.71
40	45.80	73.51	66.78	66.78	65.30	66.87	65.58	64.05	65.58
50	44.90	72.76	65.35	65.52	63.87	65.24	64.30	62.89	64.17
60	40.60	69.58	64.18	64.31	62.67	63.03	62.97	61.52	62.03
70	33.68	63.16	62.54	62.54	60.79	58.18	61.45	59.81	58.49
80	24.73	49.21	60.78	60.79	59.13	49.67	59.79	58.16	51.35
90	14.12	25.99	58.77	58.78	57.15	33.17	57.71	56.15	38.29
mAP	33.67	63.20	64.92	65.01	60.70	58.89	63.95	62.53	59.77
max. RE	99.14	99.35	95.23	95.23	92.31	98.47	97.55	95.68	99.98

Table 6.2: Performance evaluation (in-vocabulary query-task) of hybrid system with different combination strategies.

				WCN+SSPE			WCN+INED		
recall	SSPE	INED	WCN	CombMax	CombSum	CombANZ	CombMax	CombSum	CombANZ
0	17.30	75.65	57.39	60.43	60.43	60.43	64.38	64.38	64.34
10	31.90	70.43	61.05	64.44	62.48	64.59	66.36	65.22	66.35
20	37.50	68.94	59.37	62.91	61.22	63.07	64.26	63.00	64.26
30	40.60	68.09	58.26	60.88	59.48	60.89	61.86	60.30	61.85
40	41.06	66.87	57.27	60.03	58.64	59.93	59.73	58.48	59.65
50	39.10	65.44	56.11	58.64	57.05	57.69	58.64	57.25	57.94
60	33.80	61.79	54.54	57.31	55.70	54.55	57.33	56.04	55.49
70	26.78	53.71	52.74	55.40	54.17	47.82	55.50	54.19	51.06
80	18.05	38.99	50.79	53.53	52.07	35.38	53.79	52.42	43.17
90	6.19	9.48				7.63	51.84		28.42
mAP	29.01	56.20	48.19	53.19	49.37	49.95	56.47	52.76	53.36
max. RE	94.80	98.35	82.80	86.85	84.32	97.34	91.08	89.37	98.97

Table 6.3: Performance evaluation (complete query set with 13% OOV words) of hybrid system with different combination strategies.

From Table 6.3, we can see that the results gathered by the complete application evaluation experiments show that the new hybrid spoken document retrieval method improves the retrieval precision at higher recall levels. Particularly at 80% recall, the $WCN + INED\_CombMax$ method reaches a precision value of 53.79%, which corresponds to an improvement in retrieval precision value of about 14.8%, compared with the phone-based $INED$ method. In comparison with the word-based WCN method, the retrieval precision at a recall level of 80% is improved by about 3%. The best mean average precision (56.47%) is yielded by the $WCN+INED\_CombMax$ method, whereas the highest retrieval recall 98.97% is achieved by the $WCN+INED_CombANZ$ method.

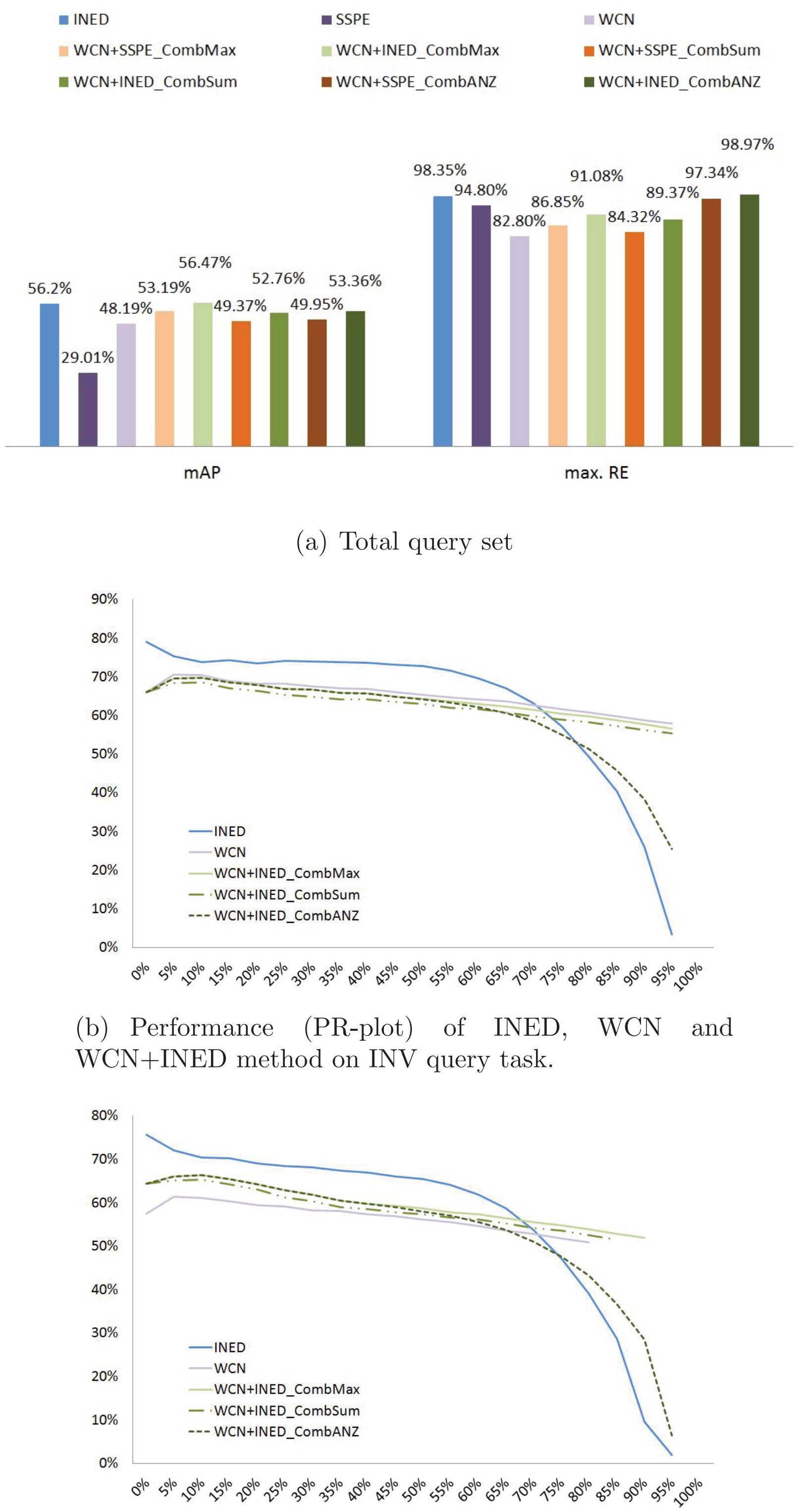
The following additional observations can be made from the experimental results summarized in Table 6.2 and 6.3 that include evaluation result with 200 in-vocabulary and total 230 queries:

- Even with a number of recognition errors in word transcription ($WER = 25\%$), the word confusion network based spoken document retrieval method achieves a mean average retrieval precision of about 64.92%.
- None of the out-of-vocabulary queries could be detected with the word confusion network based method. Consequently, in the evaluation experiments with a total query set (including 13% OOV queries), the word confusion network based spoken document retrieval system yields only a mean average precision of about 48.2%.
- The monophone sequence transcription includes a large amount of recognition errors. Therefore, the monophone sequence based algorithm yield degrading effectiveness compared to the word-based WCN method (in-vocabulary query task).
- In comparison with the word-based WCN method, the phone-based method $INED$ yields a better retrieval precision of 56.2% for the complete query set, including 13% OOV.
- The phone-based $SSPE$ yields only a 33.67% mean average retrieval precision for the in-vocabulary queries.
- The phone-based approaches could yield a better performance for in-vocabulary queries than the out-of-vocabulary queries. It is caused by the fact that the monophone transcription was directly extracted from word-1best recognition output. The chance of an exact match between a document slot and an in-vocabulary query phone sequence is much higher than for an out-of-vocabulary query.
- Combining the word-based WCN method and phone-based $INED$ with the $CombMax$ strategy (indicated by the symbol $WCN+INED\_CombMax$) yields the best mean average precision (56.47%) for the complete query set. The mean average precision is improved by about 8.27%, compared with the word-based WCN method. A recall rate of about 91.08% is reached.

6.4 Summary

In this chapter, we present a novel hybrid spoken document retrieval system. This hybrid spoken document retrieval system worked for the multi-level representation of a spoken document. The results of the investigation of the current information fusion techniques used in the text retrieval domain suggest that the multi-level representation of a spoken document consists of word confusion networks and monophone sequences, and is produced in a very efficient way. A $20k$ word recognizer is built for transcribing the spoken document. This system takes advantage of the word-based and phone-based spoken document retrieval methods and can perform an open-vocabulary query retrieval task.

We then investigated different combination strategies for the effective fusion of result sets. The fusion of result sets provided by the word confusion network based matching method and the phone-based $INED$ method with $CombMax$ method yields the best mean average precision (56.47%) for 'real application' (total query set). The mean average retrieval precision is improved by about 8.27%, compared with the result achieved by word confusion network based spoken document retrieval method. In comparison with the monophone sequence based probabilistic phone string matching $INED$ spoken document retrieval method, the mean average precision value is also slightly improved. Using the novel hybrid SDR system, the maximum retrieval recall reaches 91.08%.



(a) Total query set

(b) Performance (PR-plot) of INED, WCN and WCN+INED method on INV query task.

(c) Performance (PR-Plot) of INED, WCN and WCN+INED method on complete query set.

Figure 6.4: Performance of different combination algorithms

Chapter 7

A Video Indexing and Retrieving System

In this chapter, we propose a system for video indexing and retrieval. The spoken document retrieval methods investigated in the previous chapters are applied to retrieve the spoken content included in the audio stream of a video record. This system is developed in C++ and QT. The system structure is presented in Figure 7.1. We can seen in this diagram that this video retrieval prototype consists of two independent components: the video indexing section and the video retrieval section.

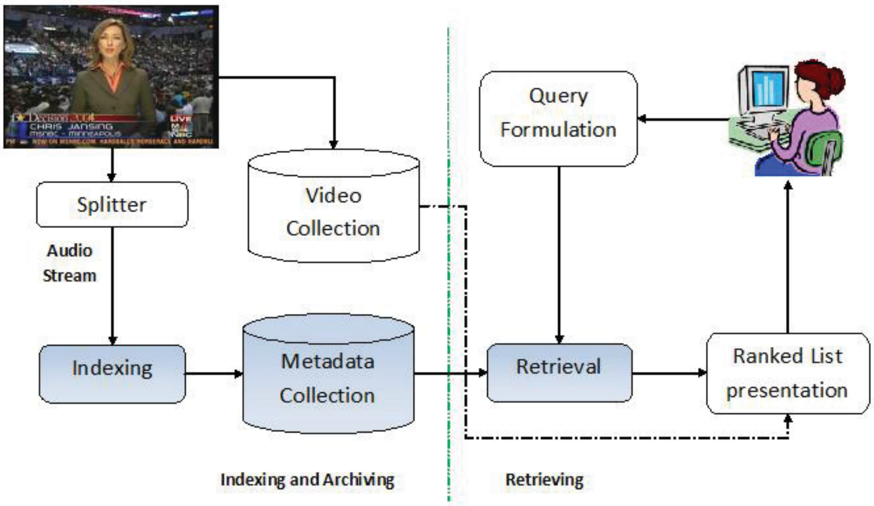


Figure 7.1: Video retrieving based on speech processing.

The video indexing and archiving part analyses new-coming videos and prepares them for information retrieval. This part of the system consists of both a splitter and the spoken content indexing tools. The audio stream is first separated from the video via the splitter and is then saved in waveform (mono, $16kHz$). The spoken document indexing methods is applied in the **Indexing** stage, to produce the transcription of spoken information included in the audio stream and to prepare it for further retrieval. The video retrieving part consists of the query formulation and the spoken content retrieval part. In our prototype, the **query formulation element** accepts natural language queries in text form. The given text query will be transcribed into a sequence of selected indexing units (phone, phone-3gram, word etc). The **retrieval element** will then guide the user to the important information required. The prototype user interface for our video indexing and retrieving system is shown in Figure 7.2.

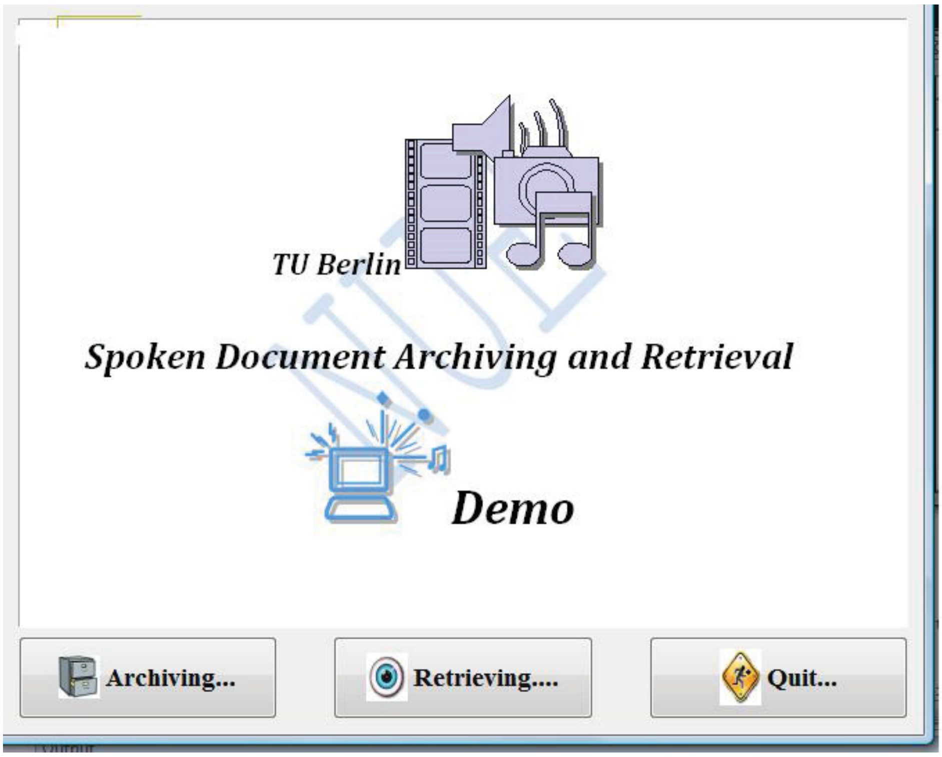
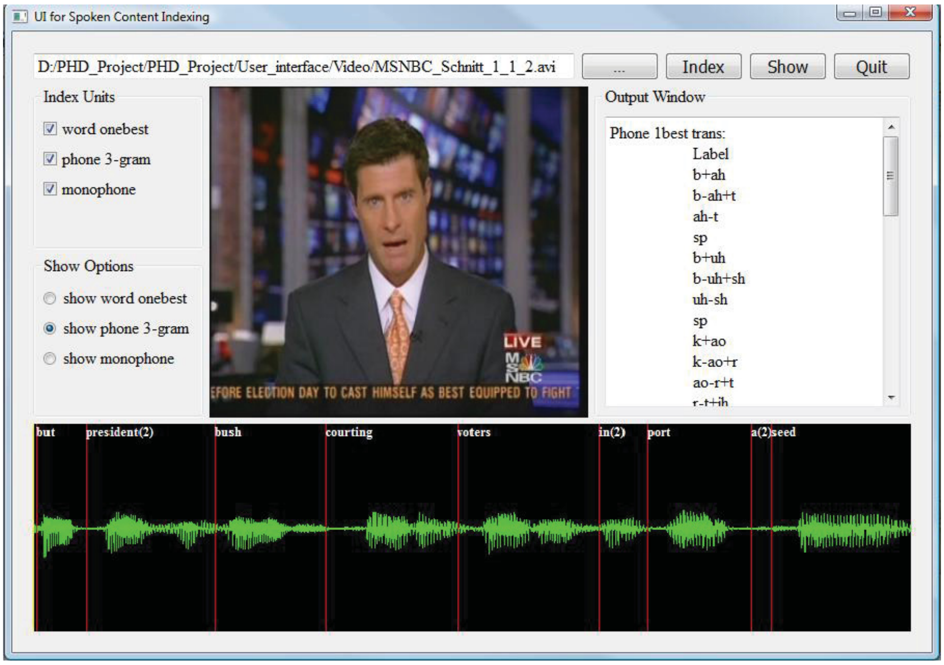


Figure 7.2: A prototype user interfaces for spoken content indexing and retrieving.

There are two sections in this chapter. In Section 7.1, we present the video indexing interface. The user interface or video retrieval is presented in Section 7.2.

7.1 Spoken Content Indexing Interface

The aim of the video archiving part is to process the spoken content of a video and prepare it for retrieval. Functions provided by the interface shown in Figure 7.3 include: preparing the multi-level representation for the spoken content of a video, displaying the result of indexing, and the playback function.



D:/PHD_Project/PHD_Project/User_interface/Video/MSNBC_Schnitt_1_1_2.avi

☑ word onebest
☑ phone 3-gram
☑ monophone

○ show word onebest
● show phone 3-gram
○ show monophone

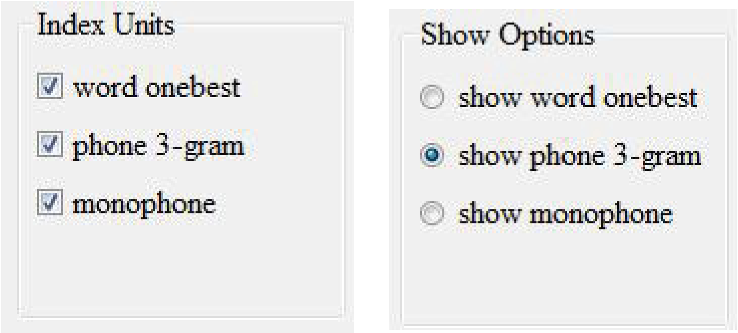
```
Phone 1best trans:
            Label
            b+ah
            b-ah+t
            ah-t
            sp
            b+uh
            b-uh+sh
            uh-sh
            sp
            k+ao
            k-ao+r
            ao-r+t
            r-t+ih
```

Figure 7.3: The user interface for spoken content indexing.

Sphinx open source speech software [63], developed by the Carnegie Mellong University, is used to build the 20k word recognizer for the indexing task. Once a video is selected, the audio stream of this video will be separated from this video and transformed into a mono 16kHz WAV file. Figure 7.4(a) shows that there are three types of index units available to date. The default option is set to preparing the word-1best representation for the spoken content of a selected video.

After clicking the **indexing** button, the spoken content of a video will then be transcribed into a text representation; a sequence of selected index units. As the indexing result, the representation will be presented in spoken indexing interface. We can see in the Figure 7.4(b) that there are three display functions in the **show options** block. For example, when the option **show word onebest** is selected, the word-1best representation will be displayed in the spoken indexing interface.

The spoken indexing interface will then present the selected representation to users.



Index Units

☑ word onebest

☑ phone 3-gram

☑ monophone

Show Options

○ show word onebest

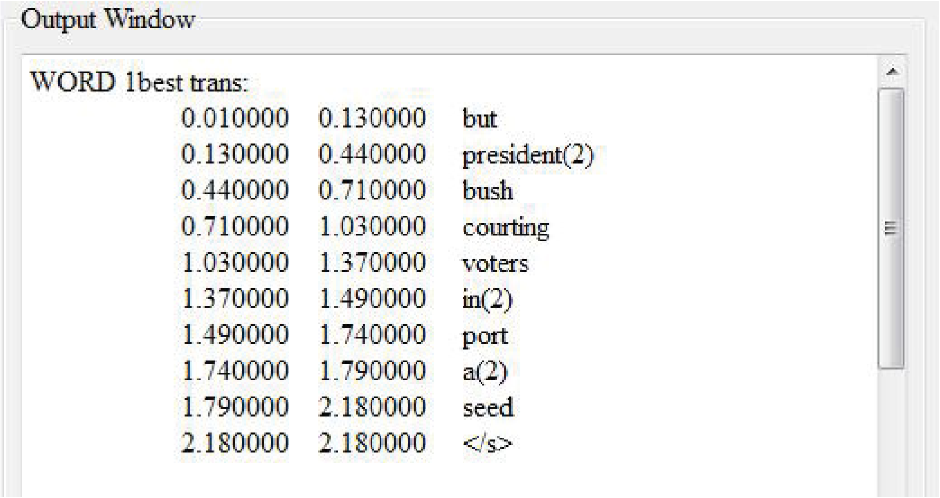
● show phone 3-gram

○ show monophone

(a) Selection of indexing units (b) Viewing transcription in different levels.

Figure 7.4: SCI interface option block.

As shown in Figure 7.5, a list of entries including attributes such as the start, and end time marks and label are displayed in the **Output Window**.



Output Window

```
WORD 1best trans:
                0.010000   0.130000   but
                0.130000   0.440000   president(2)
                0.440000   0.710000   bush
                0.710000   1.030000   courting
                1.030000   1.370000   voters
                1.370000   1.490000   in(2)
                1.490000   1.740000   port
                1.740000   1.790000   a(2)
                1.790000   2.180000   seed
                2.180000   2.180000   </s>
```

Figure 7.5: View spoken document transcription.

As shown in Figure 7.6, the temporal waveform views of audio stream will also be displayed in the interface. The axis denotes the time within a record. The entire record is presented in the waveform, labeled with the selected index units, in this interface.

7.2 Spoken Content Retrieving Interface

In this section, we describe the spoken content retrieval part of our video index and retrieval system. The task of the spoken content retrieval element is to guide the user

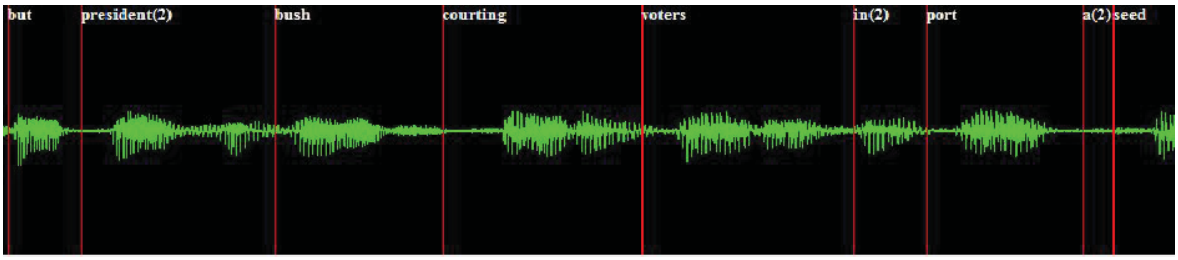
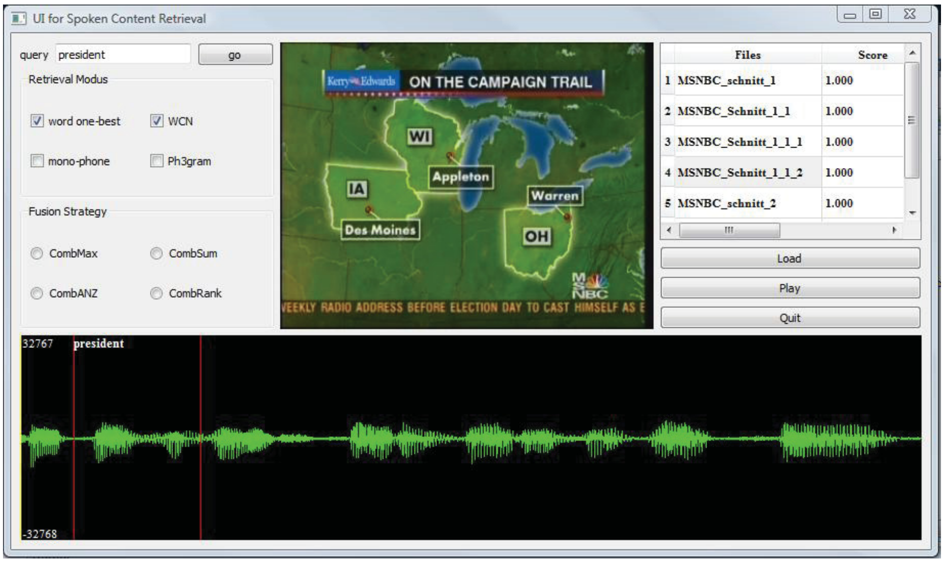


Figure 7.6: View of the labeled temporal audio stream.

to the information relevant to his specified request, as quickly as possible. In order to attain this goal, we need the efficiently implemented retrieval methods and an appropriate user interface. The user interface for the retrieval process should provide the following three main functions: query formulation; display of a ranked list of documents; document selection, display and playback functions. The Interface for Spoken content retrieval is shown in Figure 7.7. It consists of a block of query formulation, a retrieval Module block, a fusion strategy block, a video&audio viewing and a ranked document list view.


UI for Spoken Content Retrieval

query president

Retrieval Modus

☑ word one-best ☑ WCN

☐ mono-phone ☐ Ph3gram

Fusion Strategy

○ CombMax ○ CombSum

○ CombANZ ○ CombRank

	Files	Score
1	MSNBC_schnitt_1	1.000
2	MSNBC_Schnitt_1_1	1.000
3	MSNBC_Schnitt_1_1_1	1.000
4	MSNBC_Schnitt_1_1_2	1.000
5	MSNBC_schnitt_2	1.000

Figure 7.7: The user Interface for Spoken Content Retrieving.

The user forms his information request in the query formulation block. By clicking the 'Return' button, the query entry will be updated. At the same time, the user's text query will be translated into a sequence of selected retrieval basis (indexing units), with the help of a pronunciation dictionary. Query words that are not included in the pronunciation dictionary are not supported here.

The user can configure the retrieval system in the block by themselves by selecting the option in the block for the retrieval module selection and the block for the selection of different fusion strategies, which are shown in Figure 7.8.

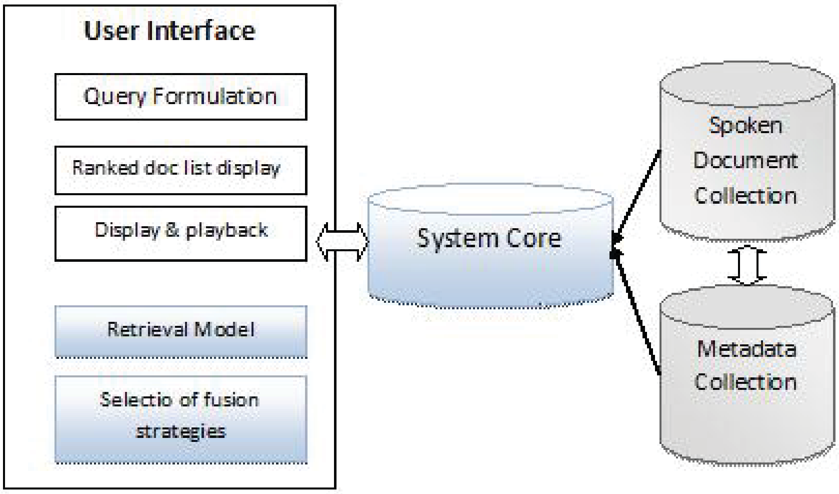
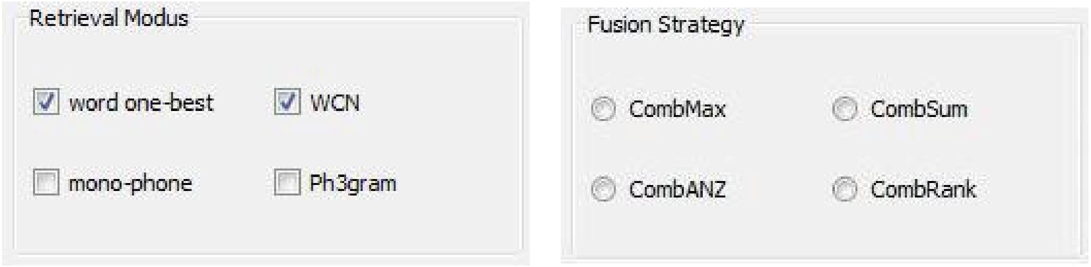


Figure 7.8: Basic Architecture of our Spoken document retrieval part.


Retrieval Modus

☑ word one-best ☑ WCN

□ mono-phone □ Ph3gram

Fusion Strategy

○ CombMax ○ CombSum

○ CombANZ ○ CombRank

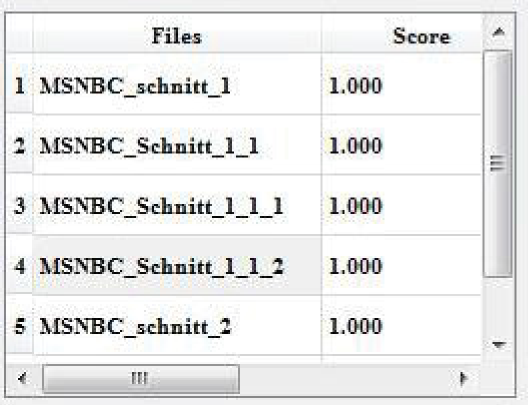
(a) Retrieval Modules selection block

(b) Combination strategies selection block

Figure 7.9: Function block for system core configuration.

In the retrieval model selection block in Figure 7.9(a), there are four different spoken document retrieval methods to be selected for the video spoken content retrieval task. These are the word-1best based method, which was introduced in Section 3.5, the word confusion network based method described in Section 4.3, the phone-3gram based method presented in Section 5.3.2 and the monophone-based probabilistic phone string matching discussed in Section 5.4.2. When more than one retrieval method is selected, the fusion strategy block is enabled. The user can then select one of the combination methods ($CombMax$, $CombSum$, $CombANZ$ and $CombRank$) for the fusion of the result sets returned by different retrieval models.

The user starts the retrieval process by clicking the 'go' button. A ranked list



	Files	Score
1	MSNBC_schnitt_1	1.000
2	MSNBC_Schnitt_1_1	1.000
3	MSNBC_Schnitt_1_1_1	1.000
4	MSNBC_Schnitt_1_1_2	1.000
5	MSNBC_schnitt_2	1.000

Figure 7.10: Results output window.

of documents is then returned. This retrieved list will be presented in the output Window, as shown in Figure 7.10. Each list entry consists of the document title and the document relevance score, corresponding to the given query.

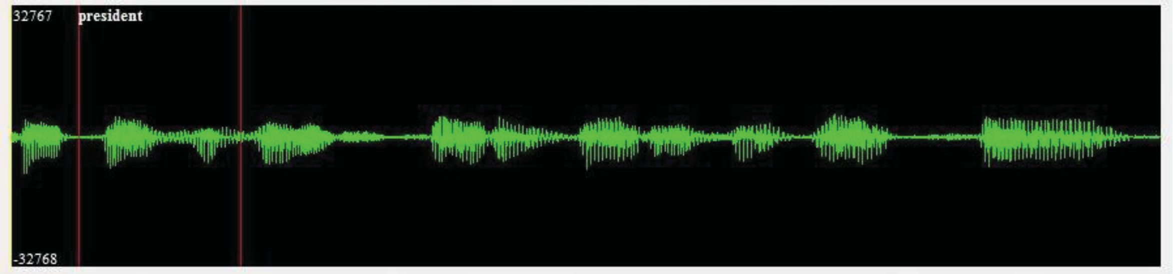


Figure 7.11: Audio view

To display the document selected from the results list the user will click the 'Load' button. We can see in Figure 7.11 that the video stream for the selected document will then load in the video display window, while the temporal view of the audio stream will be shown in the audio display window. The document slot that is the best match for the given query is highlighted. The user presses the 'play' Button to play the slot.

Chapter 8

Summary and Outlook

8.1 Summary

The performance of the current spoken document retrieval methods relies heavily on the quality of the transcription of a spoken document that is provided by an automatic speech recognition system. The performance of a spoken document retrieval system will drop with an increasing number of recognition errors in the transcription of a spoken document. Generally, the recognition errors in a spoken document transcription are often caused by mismatching between the training and test environment, different speakers, the occurrences of out-of-vocabulary words and so on. The main objective of this work is to explore effective and efficient spoken document retrieval methods that are robust against recognition errors caused by the out-of-vocabulary words and the misrecognition.

A $20k$ word recognizer has been built to transcribe a spoken document into the word sequence. The acoustic model of this recognizer was trained with the clean-speech part of the WSJ corpus, which is not used for the evaluation task. Our experiments have shown that most deletion and substitution errors are caused by out-of-vocabulary words. A lot of insertion errors occur in the spoken document transcription when there is a talking radio in the background.

In our work, different spoken document retrieval methods are explored. We have built a large test spoken document collection for the evaluation of these robust spoken document retrieval methods. This test spoken document collection includes 8158 spoken documents. A single-word query set is selected to simulate the real spoken document retrieval application scenario. This query set includes 200 in-vocabulary queries and 30 out-of-vocabulary queries. The portion of out-of-vocabulary queries is about 13%. We compare performance of the robust spoken document information methods with a baseline system. This baseline system work on the word-1best transcription of a spoken document and classical vector space retrieval model is applied to the retrieval of information contained in the spoken document.

We first investigated the word-based spoken document retrieval methods that improve performance by adding multiple recognition hypotheses to the word representation of a spoken document. Different spoken document retrieval methods based on N-best, the word confusion network or the lattice representation of a spoken document were then also evaluated. The results of this study verified that:

- In terms of the N-best representation based spoken document retrieval methods, we found out that the maximum recall rate increased with the growing amount of considered recognition hypothesis. The mean average precision (mAP) value reached the best 85.41% when 9-best recognition hypotheses are added into the spoken document representation. The document relevance score is estimated based on the document term weight. Weighting document term with probabilities could slightly improve the mean average precision, but the maximal recall drops by about 1.7%. In comparison with the baseline system, a better retrieval precision at the lower recall level has been achieved. With a growing number of considered recognition hypotheses, a lot of misrecognized words are also added into the document representation. This will lead to a drop in retrieval precision at a higher recall level.
- The word lattice representation of a spoken document includes far more recognition hypotheses than the N-best representation. It contains the most probable paths through the search space given models. A $DNLLR$ value was estimated for each link in the lattice. This value can be used to reduce the lattice search space. The $DNLLR$ threshold is fixed. The lattice links with a $DNLLR$ that is out of predefined value range will be eliminated. In our case, the best threshold for the $DNLLR$ value is $[-118, -90]$.
- The word confusion network is the most compact lattice. In comparison with the lattice, the search space is reduced by about 76.5%. The maximum achievable retrieval recall remains 95.23%. The retrieval precision at lower recall level is

improved when using the word posterior directly as the weight of a document term. The number of queries with its relevant document at first rank of the retrieved results list increase when we combine the term frequency ($tfidf$) and the posterior for document term weighting.

- Compared with the baseline system, the maximal improvement in the mean average retrieval precision is achieved by the 9-best based SDR (2.3%), and the most improvement in maximal retrieval recall rate is yielded by the word confusion network based approaches (6%).
- The query or document representation expansion enables the word-based spoken document retrieval methods to deal with out-of-vocabulary words. Our experiment results verified the improvement in yield when we replace the out-of-vocabulary queries with the acoustically similar words included in the recognizer vocabulary.

We also explored the robust spoken document retrieval methods that deal with out-of-vocabulary words by indexing the spoken document with subword units. We first researched at some of the sub word units (word, vowel-consonant-vowel features, phone-3gram and monophone). We then explored the feasibility of using subword units for the information retrieval task. The results of a text information retrieval experiments confirmed the feasibility of using subword units for the information retrieval task. This study identified that smaller subword units provide better information coverage. Therefore, the following studies in this chapter were focused on exploring phone-based spoken document retrieval methods. Our major findings are listed below:

- The phone-Ngram based spoken document retrieval method that uses term-frequency weighting has significantly improved the maximum retrieval recall compared with the baseline system. Weighting the document terms with the confusion information has enabled an approximate matching between the query representation and the document slots. The experimental result has verified that this $confusion$ method improves the retrieval recall. We proposed a new weighting method (*proximity* method) that integrated the position information into the document relevance score. This method showed its advantages in both mean average retrieval precision (69.94%) and retrieval recall (95.99%). Our experimental results have verified that the *proximity* method achieves a very good performance for the long out-of-vocabulary query retrieval task.

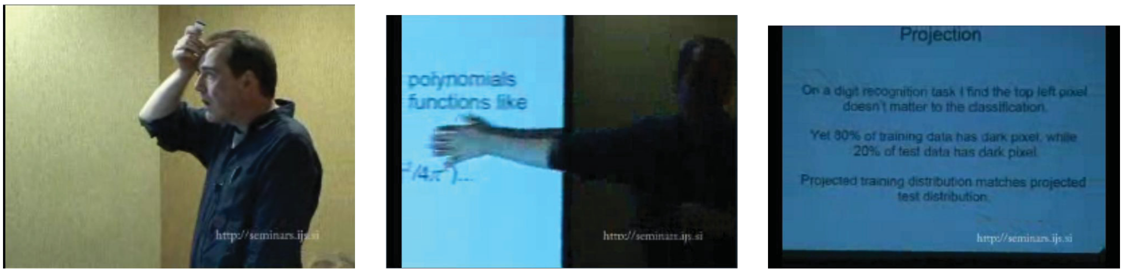
- The monophone sequence based probabilistic phone string matching ($SSPE$) method show their advantages in dealing with out-of-vocabulary queries. The monophone-based $INED$ method uses the edit distance between the query phone sequence and a document slot to form the document relevance score. The monophone-based $SSPE$ method integrates the confusion information into the document relevance score. In comparison with the phone-3gram based *proximity* method, the $INED$ and $SSPE$ methods improve the maximum retrieval recall by about 3.36%. The $INED$ method has achieved a retrieval average precision that is comparable to that of the phone-3gram based *proximity* method. The mean average retrieval precision of the $SSPE$ method drops dramatically with a decreasing query length.

- The results of these studies verify that the phone-3gram based *proximity* method and monophone-based methods ($INED$ method) provided reliable mean average retrieval precision for the long out-of-vocabulary query retrieval task.

The word-based spoken document retrieval methods provide a high retrieval precision for in-vocabulary queries. The phone-based spoken document retrieval methods show their advantages in dealing with out-of-vocabulary queries. We proposed a novel hybrid spoken document retrieval system [55]. This novel hybrid spoken document retrieval system is robust against recognition errors in the representation of a spoken document and can provide a reliable performance for the out-of-vocabulary retrieval task. In this hybrid spoken document retrieval system, the word confusion network based approach and monophone sequence based method are combined in a sufficient way. Only one $20k$ word recognizer is built for the production of the multi-level representation of a spoken document. The word confusion network representation of a spoken document is constructed by post-processing the recognition lattice output. The monophone representation of a spoken document is extracted from the word-1best output. A pronunciation dictionary is required. The experimental results have verified that this novel hybrid method ($WCN+INED\_CombMax$) provides the best precision (mAP = 56.47%) for the total query retrieval set. In comparison with the word confusion network based approach, the mean average precision is improved by about 8.27%. The maximum achievable retrieval recall that can be reached for this system is 91.08%. The results of these studies confirm that the hybrid matching system developed in this thesis provides a reliable retrieval precision rate while maintaining a high recall rate.

8.2 Outlook

There are a number of possible future studies in the field of spoken document retrieval. If the spoken document to be retrieved includes the video stream, then the precision and recall rate of the spoken information retrieval could be further improved if we take into account the video information. Figure 8.1 presents a potential application scenario. The useful video information includes the text shown in the video which could be collected using video OCR tools. We can integrate the output sequence of video OCR tools into the representation of a spoken document and prepare it for further information retrieval. In this case, the image information provided by video OCR can be combined with the Spoken information for a more accurate retrieval.


http://seminars.ijs.si
polynomials
functions like
http://seminars.ijs.si
Projection
On a digit recognition task I find the top left pixel doesn't matter to the classification.
Yet 80% of training data has dark pixel, while 20% of test data has dark pixel.
Projected training distribution matches projected test distribution.
http://seminars.ijs.si

(a) (b) (c)

Figure 8.1: Potential multi-modal IR scenario.

Another extension area could be to integrate semantic information into the retrieval process when we work with natural language queries that may contain more than 5 words. A latent Dirichlet allocation (LDA) model [10] could be built to capture the semantic information of the word transcription. Initial research has been conducted in this field. The LDA model is employed to estimate the topic distribution in queries and in word transcription of spoken documents. The matching is performed at topic level. The acoustic matching between the query phone sequence and the spoken document slots is realized using the phone-based matching algorithm. The experimental results presented in this paper [54] have verified that the results of the acoustic and topic level matching methods are complementary to one another.

In addition, research about extending applications of spoken content retrieval technologies may also be very interesting, such as the PRO-LIFE-LOG system proposed by Ziaei et. al. [124] which was designed for activities classification in daily audio streams, the lecture retrieval system introduced by Lee [62] and the news summarization system presented by Cai [12].

Appendix A

200 INV Query Set

Word	Occ	Word	Occ	Word	Occ	Word	Occ	Word	Occ
SAID	849	POINT	477	PERCENT	455	DOLLARS	396	COMPANY	383
NINETEEN	329	MILLION	312	YEARS	312	SEVEN	267	EIGHTY	227
MARKET	227	PRESIDENT	219	STOCK	201	THOUSAND	198	PEOPLE	196
COMPANIES	190	GOVERNMENT	183	FIFTY	176	OFFICIALS	163	THIRTY	162
BILLION	158	SALES	155	NINETY	152	AMERICAN	149	DOLLAR	149
UNITED	143	INDUSTRY	141	CLINTON	139	PRICES	139	STATES	136
BASED	132	BUSINESS	132	SHARE	126	STATE	122	MAJOR	121
SEVENTY	121	YESTERDAY	119	INCREASE	115	MONEY	115	ANNOUNCED	111
SHARES	111	CALIFORNIA	107	EXPECTED	105	PRICE	103	QUARTER	102
ACCORDING	100	GENERAL	99	HEALTH	99	FEDERAL	97	TODAY	97
TRADING	97	LARGE	96	NUMBER	96	CORPORATION	95	MONTH	94
NATIONAL	89	WORKERS	89	ECONOMY	88	RECENT	84	CALLED	82
CENTS	82	DON'T	81	FINANCIAL	80	SERVICE	79	ANALYSTS	78
GROUP	78	RATES	78	ECONOMIC	77	PLANS	77	COUNTRY	76
MONTHS	75	PROGRAM	75	TRADE	75	JAPANESE	74	PUBLIC	74
COMPANY'S	71	INCOME	71	PRODUCTION	71	HIGHER	70	ADMINISTRATION	69
ISN'T	67	POLICY	67	INTERNATIONAL	66	REPORTED	65	AVERAGE	64
CHANGE	64	EMPLOYEES	64	HOUSE	64	LITTLE	64	CONGRESS	63
STUDY	63	CHAIRMAN	62	LOWER	62	PRODUCTS	62	SCHOOLS	62
SOUTH	62	GOING	61	MEDICAL	61	REPORT	61	THINK	61
WHITE	61	AIRLINES	60	COSTS	60	BEGAN	59	ENVIRONMENTAL	59
INDEX	59	INVESTORS	59	SMALL	59	CHIEF	58	DIDN'T	58
INCREASED	58	BANKS	57	DEVELOPMENT	56	EXECUTIVE	56	YOUNG	56
AGREEMENT	55	AMERICANS	55	AREAS	55	CHANGES	55	DIFFERENT	55
RECORD	55	BOARD	54	EXCHANGE	54	INVESTMENT	54	NATION'S	54
BETTER	53	CASES	53	FORCE	53	RECENTLY	53	STUDENTS	53
VALUE	53	CAPITAL	52	MILITARY	52	POSITION	52	WORLD	52
EXECUTIVES	51	FOREIGN	51	CHILDREN	50	INCORPORATED	50	WOMEN	50
COUNTRIES	49	FIRMS	49	FUNDS	49	GROWTH	49	ISSUES	49
OFFICIAL	49	PROBLEMS	49	TIMES	49	EARLIER	48	SCHOOL	48

THEY'RE	48	ADDED	47	CLEAR	47	LARGEST	47	LOOKING	47
RESSURE	47	PROFIT	47	PROPOSAL	47	SECOND	47	SPOKESMAN	47
TRAVEL	47	UNION	47	ACTUALLY	46	BONDS	46	CHINA	46
COMES	46	COURT	46	DEPARTMENT	46	EARLY	46	NETWORK	46
RIGHT	46	SECURITIES	46	FOURTEEN	45	MEETING	45	POINTS	45
STORES	45	TELEVISION	45	CAUSE	44	CERTAIN	44	CORPORATE	44
COURSE	44	GROUPS	44	PRIVATE	44	SHORT	44	ADVERTISING	43
ANGELES	43	CLASS	43	CONSUMER	43	DECEMBER	43	DEFENSE	43
EFFORT	43	FUTURE	43	MARKETS	43	POSSIBLE	43	SOCIAL	43
STOCKS	43	TOTAL	43	CURRENTLY	42	EMPLOYMENT	42	INCLUDING	42

Appendix B

Evaluation Single-queries Retrieval on Lattice Transcription

recall	baseline	no_th	pr0004	pr0006	pr0008	pr0010	pr0020	pr0030	pr0040
0	86.00	93.00	93.00	93.00	92.50	93.00	92.50	92.50	89.50
10	89.12	90.70	90.70	90.70	90.70	90.70	88.50	89.00	86.10
20	89.93	89.00	89.00	89.00	88.90	88.50	86.70	87.40	81.00
30	90.12	87.00	87.00	87.00	86.90	86.70	83.70	83.70	75.70
40	89.59	84.90	84.90	84.90	84.90	84.50	80.20	80.40	68.00
50	89.05	82.10	82.10	82.10	81.90	81.50	75.40	75.30	61.30
60	88.48	77.90	77.90	77.90	77.90	77.20	70.20	69.80	-
70	88.10	72.70	72.70	72.70	72.50	72.00	63.90	63.20	-
80	87.30	67.10	67.10	67.10	67.00	66.30	58.50	-	-
90	86.98	60.80	60.80	60.80	60.80	60.30	-	-	-
mAP	84.17	79.60	79.60	79.60	79.50	76.20	66.20	63.20	42.40

Table B.1: Precision/recall and mAP value (in %) with different DNLLR-thresholdings

Appendix C

Abbreviations

ASR	Automatic Speech Recognition
DNLLR	Durationally-normalized Log Likelihood Ratio
DP	Dynamic Programming
HMM	Hidden Markov Model
idf	inverse document frequency
INED	Edit Distance based phone string matching
IPA	International Phonetic Alphabet
IR	Information Retrieval
LPCC	Linear Predictie Cepstral Coefficients
LVCSR	Large Vocabulary Continuous Speech Recognition System
MDL	Minimum Description Length
MFCC	mel-Frequency Cepstral Coefficients
mAP	Mean Average Precision Value
OOV	Out-Of-Vocabulary
PLP	Perceptual Linear Predictive
PSPL	Position Secific Posterior Lattices
SCI	Spoken Content Indexing

SDR	Spoken document retrieval
SSPE	Confusion information based probabilistic phone string matching
tf	Term Frequency
TREC	Text REtrieval Conference
VCV	Vowel-Consonant-Vowel feature
VSM	Vector Space Model
WCN	Word Confusion Network
WER	Word Error Rate

Bibliography

[1] Enginnering tables/student's t-distribution.

[2] The nist scoring toolkit sctk.

[3] Wiktionary:frequency list/tv/2006/explanation.

[4] A. Efrat A. Amir and S. Srinivasan. Advances in phonetic word spotting. In *tenth international conference on Infromation and knowledge management, CIKM*, 2001.

[5] A.Gandhe, F. Metze, A. Waibel, and I. Lane. Optimization of neural network language models for keyword search. In *ICASSP*, 2014.

[6] J. Aoe. Computer algorithms - string pattern matching strategies. *IEEE Computer Society press*, 1994.

[7] B.T. Bartell, G.W. Cottrell, and R.K Belew. Automatic combination of multiple ranked retrieval systems. In *Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 173–181 173–181, 1994.

[8] L.E. Baum. An inequality and associated maximization technique in statistical estimation for probabilistic functions of markov processes. *Inequalities*, 3:1–8, 1972.

[9] A. Berger and J. Lafferty. Information retrieval as staistical translation. In *Proceedings of the 22nd ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 222–229, 1999.

[10] D.M. Blei and A. Y. Ng an dM. I. Jordan. Latent dirichlet allocation. *Advances in Neural Information Processing Systems (NIPS)*, 14:601–608, 2002.

[11] M. Brent. An efiicient, probabilistically sound algorithm for segmentation and word discovery. *Machine Learning*, 34:71–105, 1999.

[12] X. Cai and W. Li. Ranking through clustering: An integrated approach to multi-document summarization. *IEEE Trans. Audio, Speech , Lang. Process.*, 21(7):1424–1433, July 2013.

[13] U.V. Chaudhari and M. Picheny. Improvements in phone based audio search via constrained match with high order confusion estimates. In *IEEE automatic speech recognition and Understanding workshop*, 2007.

[14] C. Chelba and A. Acero. Position specific posterior lattices for indexing speech. In *Proceedings of the 43rd Annual meeting of the ACL*, pages 443–450, 2005.

[15] C. Chelba, J. Silva, and A. Acero. Soft indexing of speech content for search in spoken documents. *Computer Speech and Language*, 21:458–478, 2007.

[16] G. Chen, C. Parada, and T.N. Sanath. Query-by-example keyword spotting using long short-term memory networks. In *ICASSP*, 2015.

[17] I-F. Chen and C.H. Lee. A resource-dependent approach to word modeling for keyword spotting. In *INTERSPEECH*, 2013.

[18] J. Choi, D. Hindle, J. Hirschberg, I. Magrinchagnolleau, C. Nakatani, O Pereira, A. Singhal, and S. Whittaker. An overview of the at&t spoken document retrieval. In *DARPA Broadcast News Transcription and Understanding Workshop*, 1998.

[19] F. Crestani. Comination of semantic and phonetic term similarity for spoke document retreival and spoken query processing. In *Proceeding of the 8th Conference on Information Processing and Management of Uncertanty in knowledge-based Systems (IPMU)*, pages 960–967, 2000.

[20] F. Crestani. Combination of similarity measures for effective spoken document retrieval. *Journal of Information Science*, 29(2):87–96, 2003.

[21] M. Creutz. Unsupervised segmentation of words using prior distributions of morph length and frequency. In *In Proc. ACL03*, pages 280–287, 2003.

[22] M. Creutz and K. Lagus. Unsupervised discovery of morphemes. In *In Proc. Workshop on Morphological and Phonological Learning of ACL02*, pages 21–30, 2002.

[23] M. Creutz and K. Lagus. Unsupervised morpheme segmentation and morphology induction from text corpora using morfessor 1.0. Technical Report Report A81, Helsinki University of technology, 2005. Publications in Computer and Infomation Science.

[24] M. Crochemore and W. Rytter. *Jewels of Stringology.* ISBN 981-02-4782-6. World Scientific Publishing Co. Pte. Ldt,, 2002.

[25] J. Cui, X. Cui, B. Ramabhadran, J. Kim, B. Kingsbury, J. Mamou, L. Mang, M. Picheny, T. Sainath, and A. Sethy. Developing speech recognition systems for corpus indexing under the iarpa babel program. In *ICASSP*, pages 6753–6757, 2013.

[26] S.J. Young D.A. James. A fast lattice-based approach to vocabulary independent wordspotting. In *Proceedings of IEEE ICASSP, Adelaide, Australia*, pages 377–380, 1994.

[27] K.H. Davis, R.Biddulph, and S.Balashek. Automatic recognition of spoken digit. *Journalof Acoustic. Soc. Am.*, 24:637–642, 1952.

[28] S.B. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. In *Readings in Speech Recognition*, pages 65–74. USA. Morgan Kaufmann Publishers, 1990.

[29] C.G. de Marcken. *Unsupervised language acquisition.* PhD thesis, MIT, 1996.

[30] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, 39(B):1–38, 1977.

[31] A. Ferrieux and S. Peillon. Phoneme-level indexing for fast and vocabulary-independent voice/voice retrieval. In *ESCA Tutorial and Research Workshop (ETRW), "Accessing Information in Spoken Audio", Cambridge, UK*, 1999.

[32] W. M. Fisher, G. R. Doddington, and K.M. Gaudi-Marshall. The darpa speech recognition research database: Specifications and status. In *Proceedings of DARPA Workshop on Speech Recognition*, pages 93–99, 1986.

[33] G. D. Forney. The viterbi algorithm. In *Proc. of the IEEE*, volume 61, pages 268–278, 1973.

[34] E. Fox and J. Shaw. Combination of multiple searches. In *Proceedings of the 2nd Text Retrieval Conferene (TREC2), NIST Special Publication 500-215*, 1994.

[35] C. Buckley G. Salton. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513–523, 1988.

[36] J.S. Garofolo, L.F. Lamel, W.M. fisher, J.G. Fiscus, D.S. Pallett, and N.L. Dahlgren. Darpa timit acoustic-phonetic continuous speech corpus cd-rom. Technical report, National Institute of Standards and Technology, NISTIR 4930, 1993.

[37] J. Glass, T.J. Hazen, L. Hetherington, and C. Wang. Analysis and processing of lecture audio data: Preliminary investigations. In *HLT-NAACL 2004 Workshop: Interdisciplinary Approaches to Speech Indexing and Retrieval*, pages 9–12, 2004.

[38] U. Glavitsch. A first approach to speech retrieval. Technical Report TR 238, Swiss Federal Institute of Technology (ETH) Zurich, 1995.

[39] U. Glavitsch and P. Schauble. A system for retrieving speech documents. In *In Proceedings of the Fifteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Language Processing,*, pages 168–176, 1992.

[40] J. Goldsmith. Unsupervised learning of the morphology of a natural language. *Computational Linguistics*, 27(2):153–198, 2001.

[41] David Graff. English gigaword. Linguistic Data Consortium, Philadelphia, 2003. ISBN: 1-58563-260-0.

[42] A.K. Halberstadt. *Heterogeneous acoustic measurements and multiple classifiers for speech recognition.* PhD thesis, Massachusetts Institute of Technology, 1998.

[43] D.K. Harman. Sixth text retrieval conference (trec-6). Technical report, Gaithersburg, MD, USA. National Institute for Standards and Technology. NIST-SP 500-240, 1997.

[44] A. Hauptmann and H. Wactlar. Indexing and searcgh of multimodal information. In *Proc. ICASSP '97*, pages 195–198, 1997.

[45] G. Heigold, H. Ney, R. Schlter, and S. Wiesler. Discriminative training for automatic speech recognition: Modeling, criteria, optimization, implementation, and performance. *IEEE Signal Processing Magazine*, 29(6):58–69, November 2012.

[46] H. Hermansky. Perceptual linear predictive (plp) analysis of speech. *J. Acoust. Soc. Amer.*, 1:1738–1752, 1990.

[47] G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhaucke, P. Nguyen, T. N. Sainath, and Brian Kingsbury. Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal Processing Magazine*, 2012.

[48] T. Hori, I.L. Hetherington, T.J. Hazen, and J.R. Glass. Open-vocabulary spoken utterance retrieval using confusion networks. In *Proc. ICASSP*, volume 4, pages 73–76, 2007.

[49] R. Hsiao, T. Ng, F. Grezl, D. Karakos, S. Tsakalidis, I. Nguyen, and R. Schwartz. Discriminative semi-supervised training for keyword search in low resource languages. In *ASRU*, 2013.

[50] X.D. Huang, A. Acero, and H.W. Hon. *Spoken language processing - a guide to theory, algorithm, and system development*, chapter 1, pages 1–15. Prentice Hall PTR, 2001.

[51] D. James. *The application of classical information retrieval techniques to spoken documents.* PhD thesis, University of Cambridge UK, 1995.

[52] D.A. James. A system for unrestricted topic retrieval from radio news broadcasts. In *ICASSP*, pages 279–282, 1996.

[53] F. Jelinek, editor. *Statistical methods for speech recognition.* MIT Press, 1998.

[54] S. Jin, H. Misra, T. Sikora, and J. Jose. Automatic topic detection strategy for information retrieval in spoken document. In *WIAMIS*, 2009.

[55] S. Jin and T. Sikora. Combining confusion networks with probabilistic phone matching for open-vocabulary keyword spotting in spontaneous speech signal. In *EUSIPCO*, 2009.

[56] Shan Jin. Methods for phone-based spoken document retrieval. Master's thesis, TU Berlin, 2005.

[57] B.H. Juang and L.R. Rabiner. Automatic speech recognition–a brief history of the technology development. In *Elsevier Encyclopedia of Language and Linguistics, Second Edition*, volume second edition. Elsevier, 2005.

[58] B. Kinsbury, J. Cui, X. Cui, M. Gales, K. Knill, J. Mamou, L. Mangu, D. Nolden, M. Picheny, B. Ramabhadran, R. Schluter, A. Sethy, and P. Woodland. A high-performance cantonese keyword search system. In *ICASSP*, 2013.

[59] M. Kurimo and V. Turunen. An evaluation of a spoken document retrieval baseline system in finnish. In *International conference on spoken Language Processing ICSLP*, 2004.

[60] M. Kurimo and V. Turunen. To recover from speech recognition errors in spoken document retrieval. In *9th European Conference on Speech Communication and Technology (Eurospeech)*, 2005.

[61] H.-Y. Lee, Y.-C. Li, C. T. Chung, and L. S. Lee. Enhancing query expansion for semantic retrieval of spoken content with automatically discovered acoustic patterns. In *ICASSP*, 2013.

[62] H.Y. Lee, S.R. Shiang, C. F. Yeh, Y.N. Chen, Y. Huang, S.Y. Kong, and L.S. Lee. Spoken knowledge organization by semantic structuring and a prototype course lecture system for personalized learning. *IEEE Trans. Audio, Speech , Lang. Process.*, 22(5):883–898, March 2014.

[63] K.F. Lee, H.W. Hon, and R. Reddy. An overview of the sphinx speech recognition system. *IEEE Transactions on Acoustics Speech, and Signal processing*, 38 No1:35–45, 1990.

[64] J.H. Leek. Analyses of multiple eidence combination. In *Proceedings of the 20th Annual International ACM-SIGIR conference*, pages 267–276, 1997.

[65] M.S. Lew, N. Sebe, and J.P. Eakins. Challenges of image and video retrieval. In M.S. Lew, N. Sebe, and J.P. Eakins, editors, *Lecture notes in coputer science*, volume 2383/2002, pages 1–6. Es. Heidelberg: Springer-verlag, 2002.

[66] B. Logan. Mel frequency cepstral coefficients for music modeling. In *In International Symposium on Music Information Retrieval*, 2000.

[67] B. M. Logan, T. Pedro, and J.-M. V. Whittaker. An exaperimental study of an audio indexing system for the web. In *ICSLP2000*, volume 2, pages 676–679, 2000.

[68] H. P. Luhn. A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of Research and Development*, 1957.

[69] W. Macherey. *Discriminative training and acoustic modeling for automatic speech recognition.* PhD thesis, RWTH Aachen University, 2010.

[70] J. Mamou, D. Carmel, and R. Hoory. Spoken document retrieval from call-center conversations. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2006.

[71] L. Mangu, E. Brilll, and A. Stolcke. Finding consensus in speech recognition: word error minimization and other aplications of confusion networks. *Computer Speech and Language*, 14:373–400, 2000.

[72] J.E. Markel and A.H. Gray. *Linear Prediction of speech.* Springer-Verlag, New York, 1976.

[73] David R.H. Miller, T. Leek, and R. M. Schwartz. A hidden markov model information retrieval system. In *SIGIR*, 1999.

[74] D.R.H. Miller, T. Leek, and R.M Schwartz. Bnn at trec7: using hidden markov models for information retrieval. In *Proceeding of the 7th Text REtriegval Conference*, pages 133–142, 1998.

[75] B. Milner. A comparison of front-end configurations for robust speech recognition. In *ICASSP*, volume 1, pages 797–800, 2002.

[76] E. Mittendorf, P. Schuble, and P. Sheridan. Applying probabilistic term weighting to ocr text in the case of a large alphabetic laibrary catalogue. In *ACM SIGIR conference on R&D in Information Retrieval*, pages 328–335, 1995.

[77] N. Moreau, H.G. Kim, and T. Sikora. Phone-based spoken document retrieval in conformance with the mpeg-7 standard. In *AES 25th Internationa Conference, London*, 2004.

[78] I. Mporas, T. Ganchev, M. Siafarikas, and N. Fakotakis. Comparison of speech reatures on the speech recognition task. *Journal of Computer Science*, 3(8):608–616, 2007.

[79] S. Jin N. Moreau and T. Si. Comparison of different phone-based spoken document retrieval methods with text and spoken queries. In *INTERSPEECH*, 2005.

[80] A. Nadas. A decision theoretic formulation of a training problem in speech recognition and a comparison of training by unconditional versus conditional maximum likelihood. *IEEE Trans. on Acoustics, Speech and Signal Processing*, ASSAP-31:814–817, August 1983.

[81] C. Ng and J. Zobel. Speech retrieval using phonemes with error correction. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 365–366, 1998.

[82] K. Ng. Towards robust methods for spoken document retrieval. In *Proceedings of the ICSLP*, 1998.

[83] K. Ng. Information fusion for spoken document retrieval. In *ICASSP 2000*, volume 4, pages 2405–2408, 2000.

[84] K. Ng. *Subword-based approaches for spoken document retrieval.* PhD thesis, Massachusetts Institute of Technology, 2000.

[85] J.J. Odell. *The use of context in Large Vocabulary Speech Recognition.* PhD thesis, Cambridge University Engineering Department, 1995.

[86] D. O'Shaughnessy. *Speech communication human and machine.* Addison-Wesley, 1987. mel scale form stamm von dises book page 5.

[87] S.E. Johnson; K. S. Jones P. Jourlin and P.C. Woodland. General query expansion techniques for spoken document retrieval. In *Workshop on Extracting Information from Spoken Audio ESCA*, 1999.

[88] Y.C. Pan, H.L. Chang, and L.S. Lee. Analytical comparison betwen position specific posterior lattices and confusion networks based on words and subword units for spoken document indexing. In *ASRU*, pages 677–682, 2007.

[89] A. Park, T. J. Hazen, and J. Glass. Automatic processing of audio lectures for information retrieval: vocabulary selection and language modeling. In *ICASSP*, pages 447–490, 2005.

[90] D. B. Paul and J. M. Baker. The design for the wall street journal-based csr corpus. In *Proceedings of the Workshop on Speech and Natural Language*, 1992.

[91] J.W. Picone. Signal modeling techniques in speech recognition. In *Proceedings of the IEEE*, volume 81, pages 1215–1247, 1993.

[92] B. Ribeiro-Neto R. Baeza-Yates. *Modern Information Retrieval.* ACM press New York and Addison Wesley Longman Limited, 1999.

[93] L.R. Rabiner and B.H. Juang. *Fundamentals of speech recognition.* Prentice Hall, englewood clitffs, NJ, 1993.

[94] L.R. Rabiner and B.H. Juang. *Statistical Methods for the Recognition and Understanding of speech.* Encyclopedia of Language and Linguistics, 2004.

[95] S. Renals, D. Abberley, D. Kirby, and T. Robinson. Indexing and retrieval of broadcast news. *Speech Communication*, 32:5–20, 2000.

[96] R.C. Rose and D.B. Paul. A hidden markov model based keyword recognition system. In *ICASSP*, number 129-132, 1990.

[97] F. Rudzicz S. Kazemian and G. Penn. A critical assessment of spoken utterance retrieval through approximate lattice representations. In *The 10th ACM International Conference on Multimedia Information Retrieval*, 2008.

[98] G. Salton and M. McGill. *Introduction to modern information retrival.* McGraw-Hill, New York, 1983.

[99] G. Salton, A. Wong, and C.S. Yang. A vector space model for information retrieval. *Communications of the ACM*, 18(11):613–620, 1975.

[100] G. Saon and J.T. Chien. Large-vocabulary continuous speech recognition systems: a look at some recent advances. *IEEE Signal Processing Magazine*, 29(6):82–97, November 2012.

[101] M. Saraclar, A. Sethy, B. RamB. Ramabhadran. Mangu, J. CuiJ. Cui. Cui, and B. KB. Kingsbury . Mamou. An empirical study of confusion modelling in keyword search for low resource languages. In *IEEE workshop Auto. Speech Recogn. Understand, (ASRU)*, 2013.

[102] P. Schaeuble and U. Glavitsch. Assessing the retrieval effectiveness of a speech retrieval system by simulating recognition errors. In *In Proc. ARPA Human Language Technology Workshop*, pages 370–372, 1994.

[103] C.D. Manning; P. Raghavan; H. Schtze. *An introduction to information retreival.* Cambridge University Press, 2009.

[104] J.E. Shoup. Phonological aspects of speech recognition. In *Trends in Speech recognition.* W.A. LEA, Ed. Englewood Cliffs: Prentice Hall, 1980.

[105] M.A. Siegler. *Integration of continuous speech recognition and information retrieval for mutually optimal performance.* PhD thesis, Carnegie Mellon University, 1999.

[106] A. Singhal, C. Buckley, and M. Mitra. Pivoted document length normalization. In *ACM SIGIR conference on R&D in Information Retrieval*, pages 21–29, 1996.

[107] F. Song and W.B. Croft. A general language model for information retrieval. In *Proceeding of the 22th ACM SIGIR Conference on Research and Development in Infomation Retrieval*, pages 216–221, 1999.

[108] S. Srinivasan and D. Petkovic. Phonetic confusion matrix based spoken document retrieval. In *23rd annual international ACM SIGIR conference on Research and development in information retrieval*, 2000.

[109] B. Teufel. *Informationsspuren zum numerischen und graphischen Vergleich von reduzierten natuerlichsprachlichen Texten.* PhD thesis, Swiss Federal Institute of Technology, 1989.

[110] TREC. Common evaluation measures. In *NIST, 10th Text REtrieval Conference*, pages A–14, 2001.

[111] V. Turunen and M. Kurimo. Using latent semantic indexing for morph-based spoken document retrieval. In *9th International Conference on Spoken Language Processing (ICSLP)*, 2006.

[112] V. T. Turunen and M. Kurimo. Indexing confusion networks for morph-based spoken document retrieval. In *ACM SIGIR*, 2007.

[113] Office of Technology Assessment (1982) US Congress. *MEDLARS and Health information policy.* ISBN 1428924248. U.S. Government Printing Office, Washington, D.C. 20402, 1982.

[114] C.C. Vogt. How much more is better? characterizing the effects of adding more ir systems to a combination. *Content-based Multimedia Information Access (RIAO)*, pages 457–475, 2000.

[115] M. Wechsler. *Spoken document retrieval based on phoneme recognition.* PhD thesis, Swiss Federal institute of Technology (ETH), 1998.

[116] M. Wechsler. New approaches to spoken document retrieval. *Information Retrieval*, 3:173–188, 2000.

[117] M. Wechsler, E. Munteanu, and P. Schuble. New techniques for open-vocabulary spoken document retrieval. In *SIGIR'98*, pages 20–27, 1998.

[118] P.C. Woodland and D. Povey. Large scale discriminative training of hidden markov models for speech recognition. *Computer Spee*, 16(1):25–48, 2002.

[119] P.C. Woodland and S.J. Young. The htk tied-state continuous speech recognizer. In *EUROSPEECH*, http://www.htk.eng.cam.ac.uk, 1993.

[120] X. Yi and J. Allan. A comparative study of utilizing topic models for information retrieval. In *31th European Conference on Information Retrieval Research*, 2009.

[121] Y.Zhang, E. Chuangsuwanich, and J. Glass. Extracting deep neral network bottleneck features using low-rank matrix factorization. In *ICASSP*, 2014.

[122] X. Zhang, J. Trmal, D. Povey, and S. Khudanpur. Improving deep neural network acoustic models using genegeneral maxout networks. In *ICASSP*, pages 215–219, 2014.

[123] B. Zhou and J.H.L. Hansen. Speechfind: an experimental on-line spoken document retrieval system for historical audio. In *Proc. ICSLP-2002*, pages 1969–1972, 2002.

[124] A Ziaei, A. Sangwan, L. Kaushik, and J. H. L. Hansen. Prof-life-log: analysis and classification of activities in daily audio streams. In *ICASSP*, 2015.

[125] V.W. Zue and S. Seneff. Transcription and alignment of the timit database. In *Proceedings of the second Meeting on Advanced Man-Machine Interface through Spoken Language*, 1988.

[126] E. Zwicker. Subdivision of the audible frequency range into critical bands. *Journal. Acoust. Soc.*, 33(2):248–248, 1961. Bark scale.

List of Figures

List of Tables