POST-SPEECH-RECOGNITION PROCESSING IN DOMAIN-SPECIFIC TEXT-CORPUS-BASED DISTRIBUTED LISTENING SYSTEM: ANALYSIS, INTERPRETATION AND SELECTION OF SPEECH RECOGNITION RESULTS

Except where reference is made to the work of others, the work described in this thesis is my own or was done in collaboration with my advisory committee. This thesis does not include proprietary or classified information.

Spencer Jaehoon Lee

Certificate of Approval:

Cheryl D. Seals
Assistant Professor
Computer Science and Software Engineering

Juan E. Gilbert, Chair
Associate Professor
Computer Science and Software Engineering

Gerry V. Dozier
Associate Professor
Computer Science and Software Engineering

Stephen L. McFarland
Dean
Graduate School

POST-SPEECH-RECOGNITION PROCESSING IN DOMAIN-SPECIFIC TEXT-CORPUS-BASED DISTRIBUTED LISTENING SYSTEM: ANALYSIS, INTERPRETATION AND SELECTION OF SPEECH RECOGNITION RESULTS

Spencer Jaehoon Lee

A Thesis

Submitted to

the Graduate Faculty of

Auburn University

in Partial Fulfillment of the

Requirements for the

Degree of

Master of Science

Auburn, Alabama
December 15, 2006

POST-SPEECH-RECOGNITION PROCESSING IN DOMAIN-SPECIFIC

TEXT-CORPUS-BASED DISTRIBUTED LISTENING SYSTEM:

ANALYSIS, INTERPRETATION AND SELECTION

OF SPEECH RECOGNITION RESULTS

Spencer Jaehoon Lee

Signature of Author

Date

THESIS ABSTRACT

POST-SPEECH-RECOGNITION PROCESSING IN DOMAIN-SPECIFIC TEXT-CORPUS-BASED DISTRIBUTED LISTENING SYSTEM: ANALYSIS, INTERPRETATION AND SELECTION OF SPEECH RECOGNITION RESULTS

Spencer Jaehoon Lee

Master of Science, December 15, 2006
(B.E., Konkuk University, August 2003)

68 Typed Pages

Directed by Juan E. Gilbert

Achieving usable recognition rates has been an almost never-ending quest in speech recognition research for more than three decades. Recently speech recognition rates have dramatically improved in conjunction with the rapid development of computer technology, but it has never been enough to satisfy human expectation. Many researchers tried to testify the benefit of using multiple speech recognizers in improving recognition rates. The fundamental idea supporting this research trend is that recognition results agreed upon by a majority of recognizers can be considered correct. This paper tries to break the old idea which may prevent multi-recognizer researches forever from achieving usable recognition rates, by revealing the existence of common misrecognition (CMR) results agreed upon by the majority. The common misrecognition results are classified into several categories (contraction, missed words, spoken stop words, homophone, and combined misrecognition) and treated according to their characteristics. A collection of sentences users may speak (simple text-corpus) is used in order to overcome very low sentence recognition rates of speech systems. It is suggested that composite information made out of multiple recognition results is enough to correctly find its actual target sentence among thousands of sentences in a specific domain. Overall the results (87% of sentence recognition rate) of experiments conducted in this research strongly support that the processes described in this paper can greatly improve speech recognition rates of multi-recognizer systems.

Style manual or journal used: Journal of SAMPE

Computer software used: Microsoft Word 2003

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

Speech is a very basic, but the most natural and effective communication method that humans possess. Though there are many other effective means of human communication such as writing, gesture, and facial expressions, communication through speech is preferred especially when people exchange information and establish relationships with each other. [Pinker 1994]

As computer technology rapidly grows, development of communication methods between human and computers has also grown greatly. Especially graphical user interfaces (GUI) using traditional input-output devices such as screen display, keyboard, mouse, and joystick are in the very mature stage of usability and their further developments are focusing on increasing fidelity in order to meet the highly increasing expectation of human users and ergonomic usability to reduce repetitive strain injury (RSI).

As an alternative means of human-computer interaction, speech recognition technology has been developed with a great deal of effort from researchers for more than 30 years. In spite of the very rapid progress of recent speech technology development prompted by advances in computer power, algorithm and memory capacity, there still exists a big usability gap between human expectation and accuracy of speech recognition. [Deng and Huang 2004]

The traditional GUI is not always optimal for human-computer communication. In limited situations with small devices such as cell phones and PDAs, or driving a car, the use of a GUI is very limited or almost impossible. Even in normal working environments, it is much quicker to dictate text to a computer than to type it by hand [Larsen et al.1992]. In contrast, speech interfaces can provide consistent usability across all computer devices. [Deng and Huang 2004]

In highly developed countries like the United States, the rapid increase of the aged population (over 65 years of age) is a very common. The percentage of this population in the United States in 2000 was 12.6%, and it is estimated that the percentage will rise up to 20% by 2030 [Gavrilov and Heuveline 2003]. Aged people are not very familiar with interacting with computers and gradually lose their mobility and dexterity more and more as they age. Therefore, the development of an alternative means of human-computer interaction to the traditional GUI is very encouraged and emphasized, because the easy use of computer technology in conjunction with robotics technology may make their elderly life much better or even allow them to work beyond traditionally expected retirement ages [Stevenson and McQuivey 2003]. The best alternative communication means to the traditional GUI can be speech interactions, because most people retain their speaking ability throughout life.

Figure 1 from [Deng and Huang 2004] shows that word recognition error rate under controlled environment has gradually decreased by an average of about 10% every year during the past two decades. After the recognition rate reached beyond 90%, two new big research trends came out, which are speech recognition under normal acoustic (noisy) environment and conversational (casual) speech recognition. [Deng and Huang 2004]

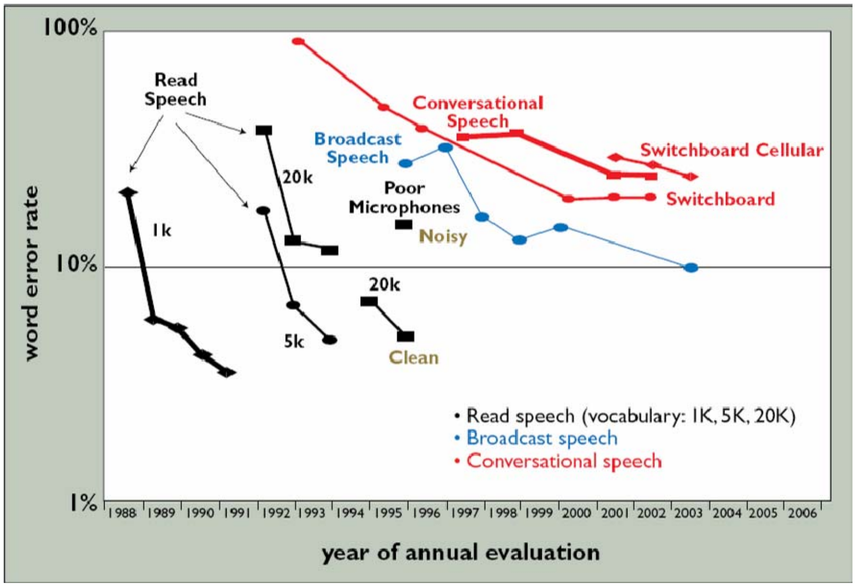


Figure 1: The progress of the development of speech recognition

Another research trend for achieving usable speech recognition is the use of multiple speech recognition engines to get a better overall speech recognition result. A sentence is a collection of words. Likewise speech recognition of a sentence is a result of continuous speech recognition of single words. Therefore successful recognition of a sentence depends on successful recognition of every single word in the sentence. Most speech recognition rates in speech recognition literatures refer to word recognition rates not sentence recognition rates. Even though word recognition rate reached beyond 90%, sentence recognition rate is still very poor.

The purpose of using multiple speech recognizers is to make a best (preferably a complete) composite output out of several incomplete recognition results by complementing each other's result. More specifically most multi-recognizer research efforts focus on how to select the right words in certain regions of recognition results where a majority of recognizers don't clearly agree.

This research is an extension of distributed listening [Gilbert 2005], which consists of multiple recognizers with their own individual microphones (called the 'listener") and one system (called the 'interpreter') that collects recognition results of listeners and tries to deliver a correct result. This paper will talk about the results of the preliminary experiment data analysis of distributed listening, suggest the interpreter design based on the analysis, and implement and test the suggested interpreter design. The data analysis revealed that it is very possible that a majority of recognition results returned by listeners contain a large number of the same incorrect words (common misrecognition) which make the interpreter unable to return positive results. The suggested interpretation design includes a normalization process to clean and reformat strings, a pattern matching process to examine and separate common information and uncommon information found in the multiple recognition results, several common misrecognition (CMR) treatments to resolve the common misrecognition problems found in the analysis part, and a selection process to choose a correct result when multiple matching results of sentence recognition were returned.

CHAPTER 2

LITERATURE REVIEW

[Barry et al. 1994] is one of the early researches using multiple speech recognizers. As shown in figure 2 one microphone input was shared by three computers (recognizers) and one master computer (RS-232) receives recognition results from each of the three recognizers, and then selects one with clear majority. The experiment was single-word speech recognition test and two sets of 20 and 25 words were used.

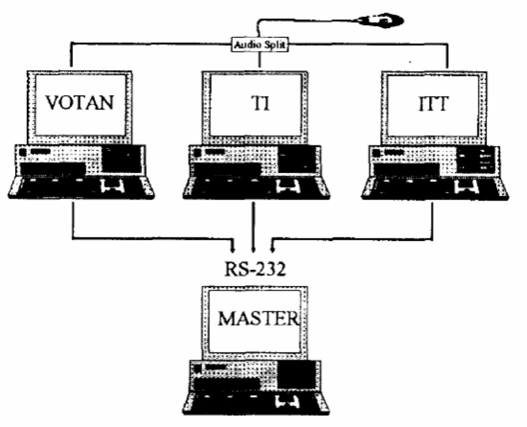


Figure 2: Hardware configuration used in the experiment of [Barry et. al. 1994]

The selection algorithm was very simple. When there is a word with clear majority, the word is selected as a correct recognition result. For example, if the master receives two identical words and one other, the common words is selected. If it receives one word and two invalid or no response, the word is selected. If there is no word with clear majority such as three different words, or no responses, the master received the second best words from the three recognizers, and then the new words participated in the competition with equal weight. Figure 3 is one of the experiment results and it clearly shows that the combined overall result (EMR) is better than any of the three individual recognizers.

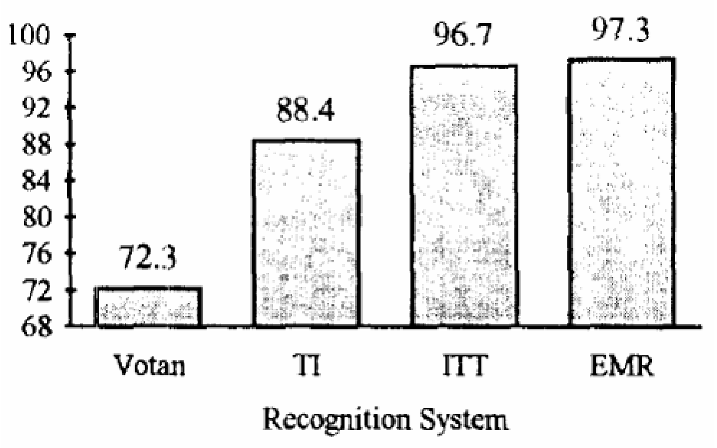


Figure 3: The result of thee recognizer system

As stated in the paper, the performance of ITT is much superior to the two other recognizers (Votan and TI) and the overall result (97.3%) is just slightly higher than the result of ITT (96.7%). So there is some possibility that the outstanding overall result is due to the superior performance of ITT. [Barry et al. 1994]

This early research testified that the use of multiple recognizers can improve the overall word recognition rate. The method of choosing a word with clear majority became the very fundamental idea. The algorithm to select a right word was frequently modified and improved upon to yield better results in the word level selection processes in many of the later research efforts using multiple speech recognizers.

Recognizer Output Voting Error Reduction (ROVER)

Recognizer Output Voting Error Reduction (ROVER) system [Fiscus 1997] used an alignment module to align multiple recognition results and then a voting module to form a composite output of words with higher voting scores as seen in figure 3.

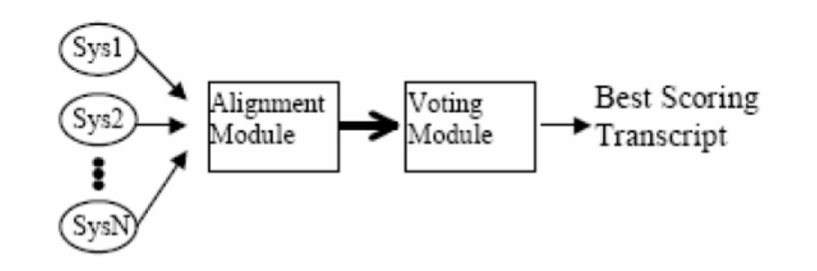


Figure 3: ROVER system architecture

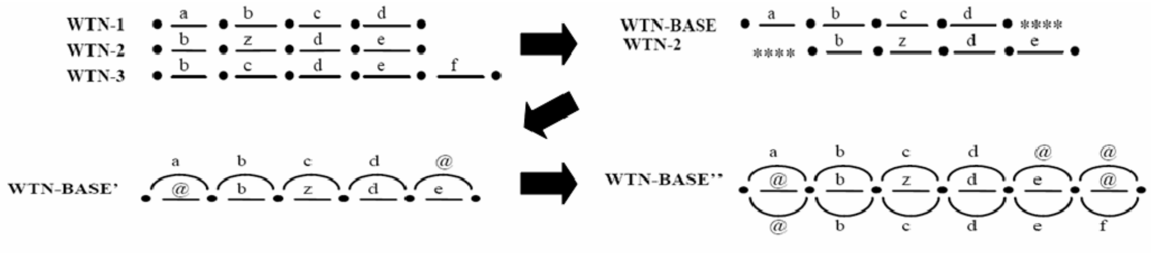
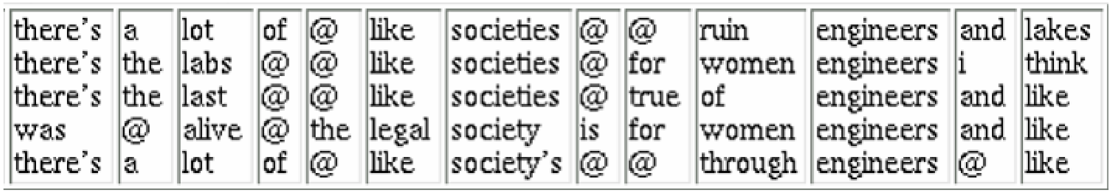


Figure 4: The progress of forming WTN

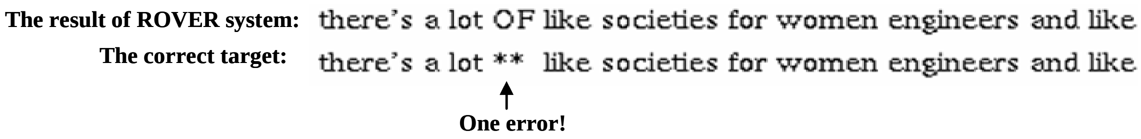
The alignment module aligns multiple speech recognition results into a single composite output called word transition network (WTN) by using dynamic programming (DP) method. Figure 4 shows the alignment progress. The first input (WTN-1) is chosen as a base WTN and then the second WTN is aligned with the base. The third picture of figure

4 shows a new base WTN. And this progress is repeated until there is no more speech recognition input to combine. As stated in the paper, the final form of the composite WTN is not always an optimal WTN because the combining order of WTNs affects the composite WTN and may result in different composite WTNs. [Fiscus 1997]



there's	a	lot	of	@	like	societies	@	@	ruin	engineers	and	lakes
there's	the	labs	@	@	like	societies	@	for	women	engineers	i	think
there's	the	last	@	@	like	societies	@	true	of	engineers	and	like
was	@	alive	@	the	legal	society	is	for	women	engineers	and	like
there's	a	lot	of	@	like	society's	@	@	through	engineers	@	like

Figure 5: An example of a composite WTN



The result of ROVER system: there's a lot OF like societies for women engineers and like

The correct target: there's a lot ** like societies for women engineers and like

One error!

Figure 6: The result of voting process

Figure 5 is an example of a composite WTN formed by aligning five outputs. Each column is an independent voting section and the word with highest score in each section is selected as a correct word. The final result produced from the composite WTN by ROVER system (Figure 6) shows that it has one error, which is much better than the least number of errors, 3, the best single system has.[Fiscus 1997] In the experiment using submissions of LVCSR 1997 Hub 5-E evaluation (administered by NIST), ROVER system reduced the word error rate from 44.9% (the rate of the best single system) to 39.4% [Fiscus 1997]. In the experiment using nine systems conducted by NIST, ROVER was able to reduce the word error rate from 13.5% to 10.6% [Pallett et al. 1999].

[Schewenk and Gauvain, Sept 2000] improved ROVER system by adding normalization/filtering process and using language model information. One of the problems the original ROVER system had was that the order of combining outputs in the alignment module affected the result of the aligning process. The paper suggested that it is better to use the output from the best recognition system first and then combine the rest in the descending order of recognition rate. Figure 7 shows that the error rate did not change much and sometimes even increased when 5 to 9 systems were combined. Therefore it was suggested that combining systems with high error rate may not improve the performance. [Schewenk and Gauvain, Sept 2000]

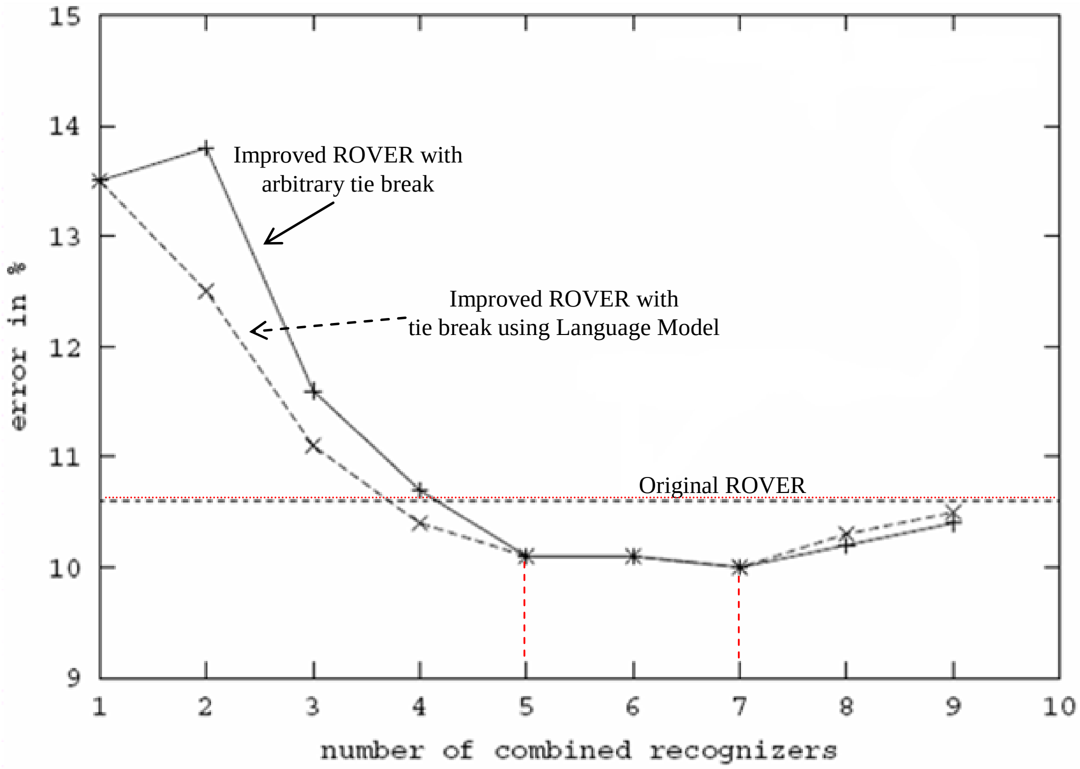
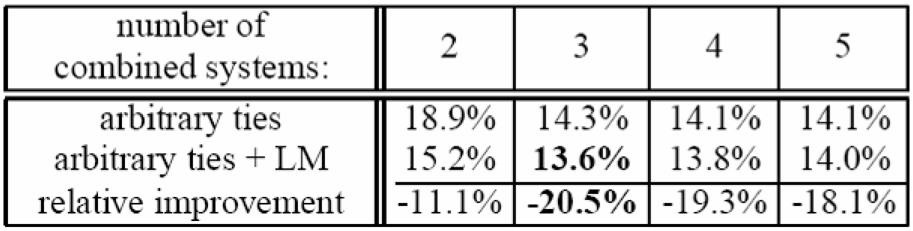


Figure 7: Word error rates in function of the number of combined system

The normalization/filtering process is a process to change phrases with alternative spellings to one common form. For example, “cannot” → “can not”, “child’s” → “child’s” or “child is” or “child has”. In the improved ROVER system, only simple one-to-one filtering was used because of technical difficulty, and a slightly improved result (10.1% to 10.0%) was produced when outputs of 7 recognizers were combined. Another problem of the original ROVER system is that some words had the same voting scores (called “ties”) after alignment and the ties were arbitrarily broken in the system. If the correct words are ideally selected every time a tie is encountered instead of breaking them arbitrarily, the word error rate may drop below 5%. So in the improved ROVER system, language model (LM) information was used as a tie breaker in order to select correct words in tie situations. Figure 7 shows that tie breaking using LM produced better results when combining up to the 4th recognizer, but almost same or even worse results when combing 5 to 9 recognizers. Table 1 from [Schewenk and Gauvain, Oct 2000] shows the relative improvement rate to the rate of the best recognizer achieved the highest (20.5%) when three best recognizers were combined. These result suggests that the use of LM for breaking ties can improve the word error rates but the number of recognizers combined may need to be restricted and recognizers with high error rates should be excluded. [Schewenk and Gauvain, Sept 2000]



number of combined systems:	2	3	4	5
arbitrary ties	18.9%	14.3%	14.1%	14.1%
arbitrary ties + LM	15.2%	**13.6%**	13.8%	14.0%
relative improvement	-11.1%	**-20.5%**	-19.3%	-18.1%

Table 1: Relative improvement rate with respect to the best single recognizer (17.1%)

ROVER is a very popular speech recognition system using multiple recognizers, which was frequently referenced in many other related research efforts to compare with their research results. ROVER introduced and tried many useful processes to improve recognition results, such as alignment, voting, normalization/filtering, and use of language model for tie break, and successfully proved that most processes can lower the error rates. It also proved that combining a restricted number (3 or 4) of the best recognition results is better than combining all recognition results including unreliable ones with high error rates.

[Cristoforetti et al 2003] used multiple speech recognizers to achieve robust speech interaction in vehicles. One input source (one microphone) was shared by recognizers as shown in figure 8 and each of the recognizers carried different sets of vocabularies and language model for specific narrow domains such as tourist information, geographic information, etc, instead of carrying one large vocabulary and language model replicated over each recognizer in order to reduce complexity. In contrast with ROVER system, the system selects one best recognition result with the highest possibility, instead of making a composite result out of all the recognition results returned by the recognizers. Another characteristic of this research is that a corpus of speech interactions which can happen in a vehicle was gathered in simulated driving situation applying the Wizard of Oz method (WOZ), and it was used to populate the sets of vocabularies and language models of each recognition unit. [Cristoforetti et al 2003] The result of the experiment in this research shows that word recognition rates were improved with respect to the rate of recognizers carrying information for a whole domain by 3.7% (58.5% → 62.2%) for closely located microphone inputs, and 2.3% (45.7% → 48.1%) for far located microphone inputs. Sentence recognition rates were improved by 5.6% (29.4% → 35.0%) and 6.9% (29.2% → 22.3%) respectively. Therefore this research revealed that there is a possibility that distributing vocabularies and language model among recognizers instead of making each one carry all of the information may greatly improve overall sentence recognition rates.

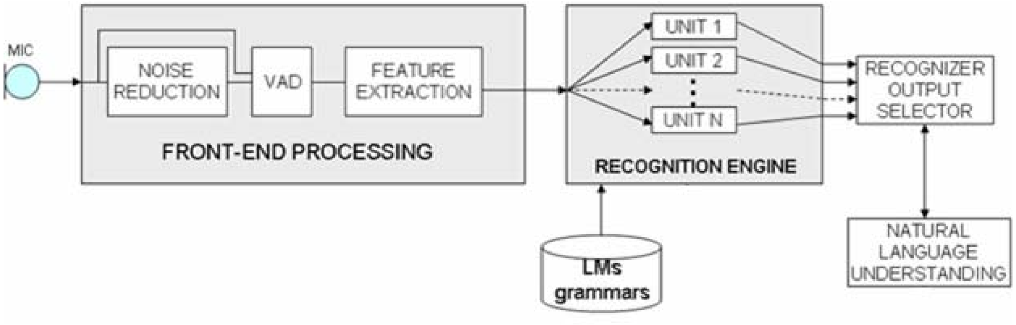


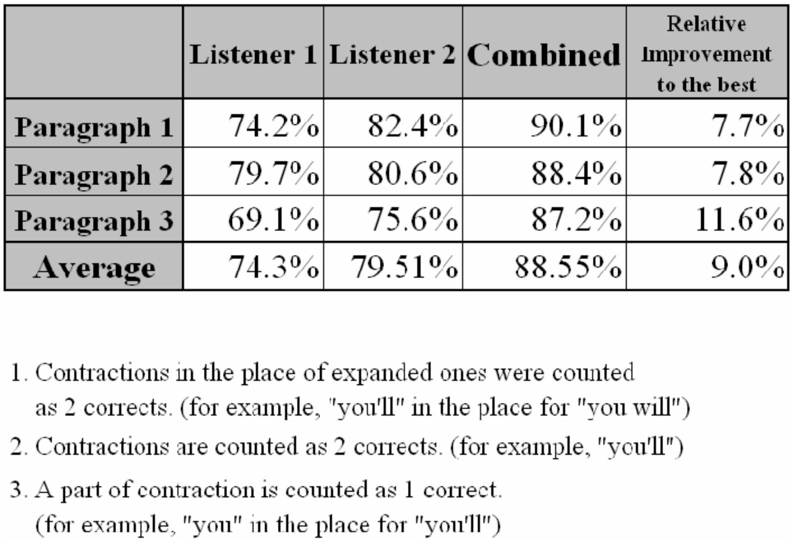
Figure 8: Parallel recognition units

Distributed Listening (DL)

[Gilbert 2005] introduced a detective story as a good analogue of multiple speech recognizer systems to develop a better architecture design in conjunction with their similarity. A crime is committed and it is witnessed by several people. A detective starts the crime investigation by interviewing the witnesses. Each of the witnesses gives the detective slightly different stories because they were in different places and perceived the event differently. The detective tries to develop one complete scenario based on the idea that common pieces of witnesses are more likely to be right. And incomplete parts are resolved by the experience and knowledge of the detective. This story can be restated in terms of speech recognition as described next.

An utterance is spoken and it is recognized by several recognizers. A process starts the speech examination by gathering information from the recognizers. Each of them gives slightly different results because they were in different places and recognized the speech differently. The process tries to develop one composite result based on the idea that common pieces in each recognition result are more likely to be correct. And incomplete parts are resolved by the intelligence and information the process has.

As an analogy of independent witnesses in crime scene, multiple speech recognizers with their own microphones were used and named “Listeners”. As a detective, a system examining recognition results returned by each listener is named “Interpreter”. It was suggested that it would be more beneficial for each listener to have its own input source because bad input will be given to all the listeners with only one shared microphone when a good signal isn’t received.



	Listener 1	Listener 2	Combined	Relative Improvement to the best
Paragraph 1	74.2%	82.4%	90.1%	7.7%
Paragraph 2	79.7%	80.6%	88.4%	7.8%
Paragraph 3	69.1%	75.6%	87.2%	11.6%
Average	74.3%	79.51%	88.55%	9.0%

1. Contractions in the place of expanded ones were counted as 2 corrects. (for example, "you'll" in the place for "you will")
2. Contractions are counted as 2 corrects. (for example, "you'll")
3. A part of contraction is counted as 1 correct. (for example, "you" in the place for "you'll")

Table 2: Word recognition rates of DL preliminary experiment

Table 2 shows the word recognition rates of the preliminary experiment with two listeners of distributed listening. 46 subjects (16 females, 30 males) participated in the experiment and one of them was excluded in the counting because it is an outlier. The tasks given to each subject was to dictate 3 2-sentence paragraphs taken from a book. Totally 135 pairs (45 x 3) of speech recognition results returned by two listeners were collected during the experiment. The combined result is the percentage of correct words when the recognition results of listener 1 and listener 2 are ideally combined. The average improvement with respect to the average rate of the best recognizer (listener 2) is 9.0% (79.51% → 88.55%). This improvement rate suggests that distributed listening systems may collect all the correct words in a sentence when the results of 4 or slightly more listeners are combined. And it may also closely support the number of recognizers suggested by [Schewenk and Gauvain, Oct 2000].

Many topics regarding multiple speech recognizer research were briefly described here. Most researchers tried to answer and resolve the following questions: Should we use one single input (microphone) or individual input source for each recognizer? Are normalization/filtering processes useful? Should we use same recognizers or different recognizers? Choose a best recognition result or make a composite one? How many recognizers need to be used? Is there any aligning method which is unaffected by combining order? Is a word (or phrase) agreed by a majority of recognizers always a correct one? Can a collection of sentences (corpus) users may speak be useful? Some of them were well proved but some are still in question.

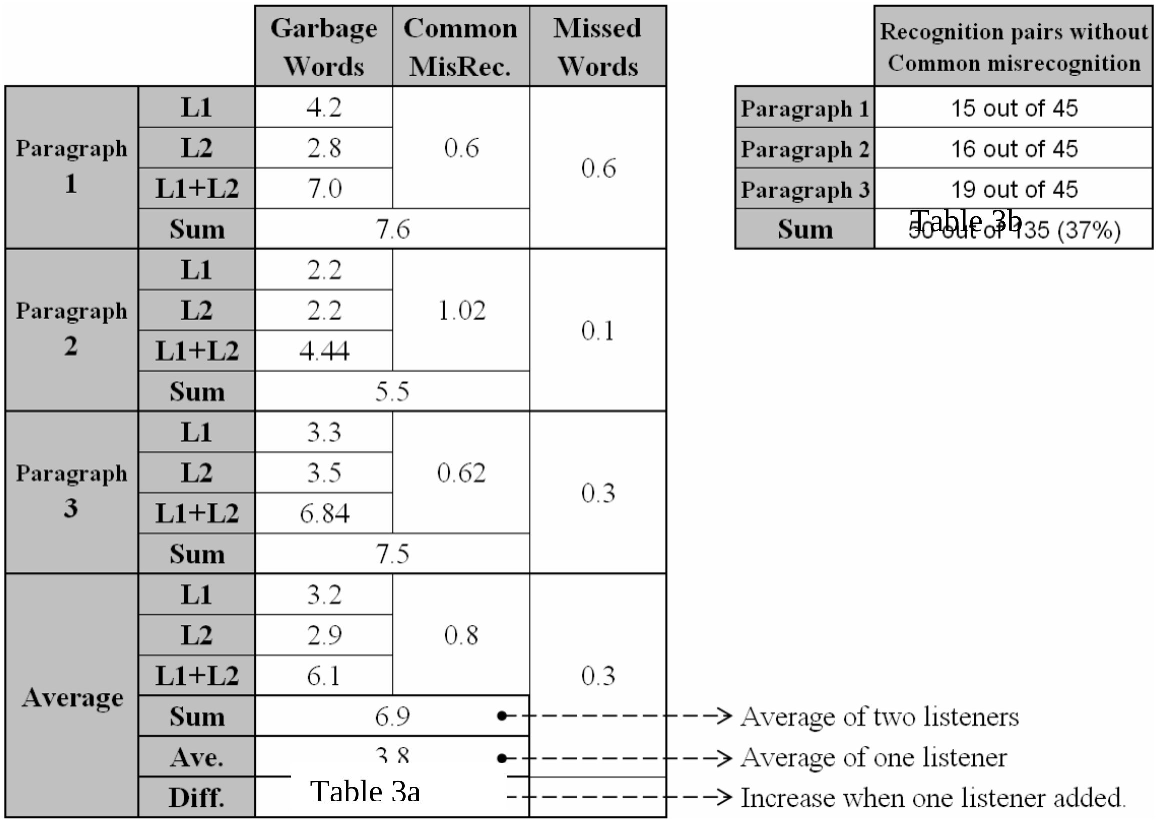
The ultimate objective of speech recognition research using multiple recognizers is to increase sentence recognition rates up to usable rates. Many researchers showed that there were increases in word recognition rates but sentence recognition rates are still very low though improved by combining multiple recognition results. The very fundamental idea supporting multi-recognizer research is that information agreed by a majority of recognizers is more likely to be correct. However, it may not always be true because there are so many similar words in English vocabularies and their pronunciations are very easily affected by surrounding noises and intonation of speakers. Therefore there is very high motivation to clarify the possibility of the existence of common misrecognitions, improve how to utilize the advantages of using multiple recognizers or find other ways to make usable multi-recognizer systems even when sentence recognition rates are very low.

CHAPTER 3

INITIAL STUDY

3.1 Analysis of preliminary experiment data of Distributed Listening (DL)

Table 3 shows the analysis results of the preliminary experiment data of Distributed Listening. Garbage words in table 3a are misrecognized words found in each result of listener 1 and 2, but not in both of them. Common misrecognition refers to misrecognized words commonly found in both results of listener 1 and 2. Missed words are commonly unrecognized words by both listeners. The average number of misrecognized words by both listeners is 6.9. This means that 2-listener distributed listening systems may have about 7 misrecognized words in every speech-recognition. More importantly it shows another finding that the number of commonly misrecognized words may increases by 3 when one listener is added to the system. 3 words are about 10% of the average number of words (29.7 words) the three paragraphs contain. [Schewenk and Gauvain, Oct 2000] suggested that combining 3 recognizers may give the best benefit of using multiple recognizers, and the analysis of the preliminary experiment of distributed listening showed that all the necessary correct words may be gathered when 4 or slightly more listeners are combined.



		Garbage Words	Common MisRec.	Missed Words
Paragraph 1	L1	4.2	0.6	0.6
	L2	2.8		
	L1+L2	7.0		
	Sum	7.6		
Paragraph 2	L1	2.2	1.02	0.1
	L2	2.2		
	L1+L2	4.44		
	Sum	5.5		
Paragraph 3	L1	3.3	0.62	0.3
	L2	3.5		
	L1+L2	6.84		
	Sum	7.5		
Average	L1	3.2	0.8	0.3
	L2	2.9		
	L1+L2	6.1		
	Sum	6.9		
	Ave.	3.8		
	Diff.			

Average of two listeners

Average of one listener

Increase when one listener added.

Table 3a

	Recognition pairs without Common misrecognition
Paragraph 1	15 out of 45
Paragraph 2	16 out of 45
Paragraph 3	19 out of 45
Sum	50 out of 135 (37%)

Table 3b

Table 3: Analysis results of preliminary experimental data

Another very important finding is almost 63% (table 3b) of recognition pairs returned by two listeners contain at least one commonly misrecognized word in both recognition results. And the average number of common misrecognitions per speech recognition is 1.1 (0.8 + 0.3 (common misrecognition + missed words)). This means that there is a very high possibility that every speech recognition result returned by 2-listener distributed listening systems may averagely contain at least one commonly misrecognized word. This finding strongly supports the theory that recognition results agreed by a majority of recognizers in multi-recognizer system may not be always correct. Therefore if this important fact is not seriously taken into account in designing multi-recognizer systems, the sentence recognition rates may never be able to meet the usable level.

3.2 Input Analysis

Input analysis in this research is the process which inspects recognition results received from each listener and then returns two important outputs used in later processes. One is Common Words and Structure (CWS), and the other is Different Words list (DW). CWS contains sets of common words found in the recognitions results and markers ("***" + index of each set) for uncommon words sets found between common words sets. DW contains the sets of the uncommon words and their index is same as the postfix index of place markers in CWS which indicate where they were collected. All the sets and markers in CWS and DW are in the order found. Figure 9 illustrates the actual structure of CWS and DW. Simply, CWS contains common words and markers in the order found, and the DW list contains uncommon words sets found in the region the markers of CWS indicate. In other words, CWS carries common information and DW carries uncommon information.

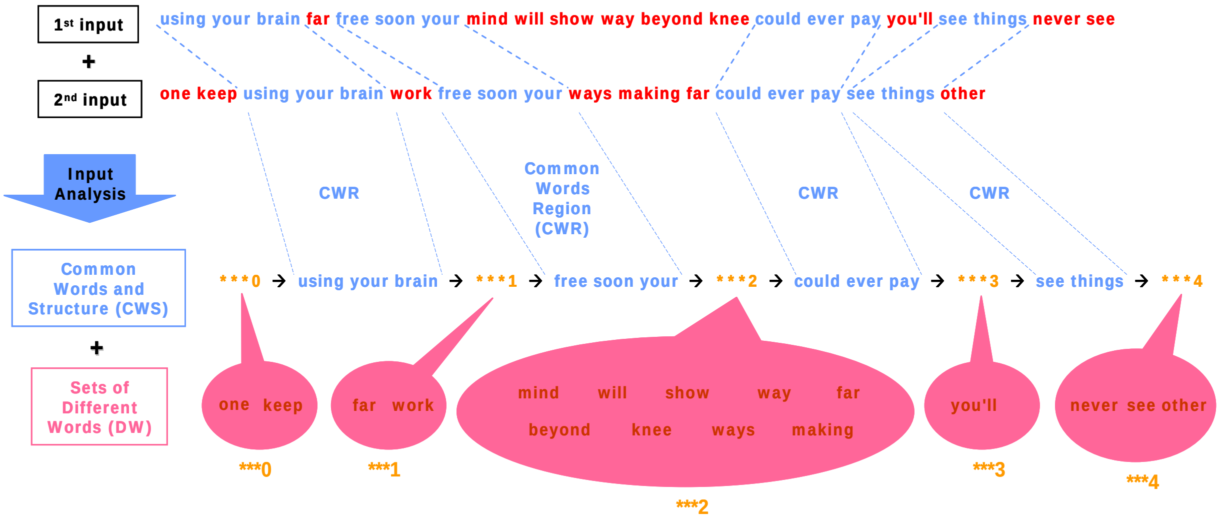

1st input
+
2nd input
using your brain far free soon your mind will show way beyond knee could ever pay you'll see things never see
one keep using your brain work free soon your ways making far could ever pay see things other
Input Analysis
CWR
Common Words Region (CWR)
CWR
CWR
Common Words and Structure (CWS)
+
Sets of Different Words (DW)
***0 → using your brain → ***1 → free soon your → ***2 → could ever pay → ***3 → see things → ***4
one keep
far work
mind will show way far beyond knee ways making
you'll
never see other
***0
***1
***2
***3
***4

Figure 9: Input Analysis

Later in this research, CWS is used for finding a target paragraph among all the candidate paragraphs. If a candidate paragraph contains all the common words sets in CWS in order, then the candidate paragraph is considered as a matching target paragraph. It is possible that the matching process returns multiple target paragraphs, so a process selecting a best target paragraph is necessary. Uncommon words sets of DW are used for counting the number of matching words in uncommon words regions in candidate paragraphs and calculating the final matching rate of the paragraph.

3.2.1 Pattern matching

During the input analysis process, a pattern matching algorithm was used in order to detect and collect common words in the two recognition results returned by two listeners and separately collect uncommon words between common words regions. The use of the pattern matching approach provides ease of changing and using various patterns, flexibility of accommodating inputs from more than two listeners and much better readability of program codes.

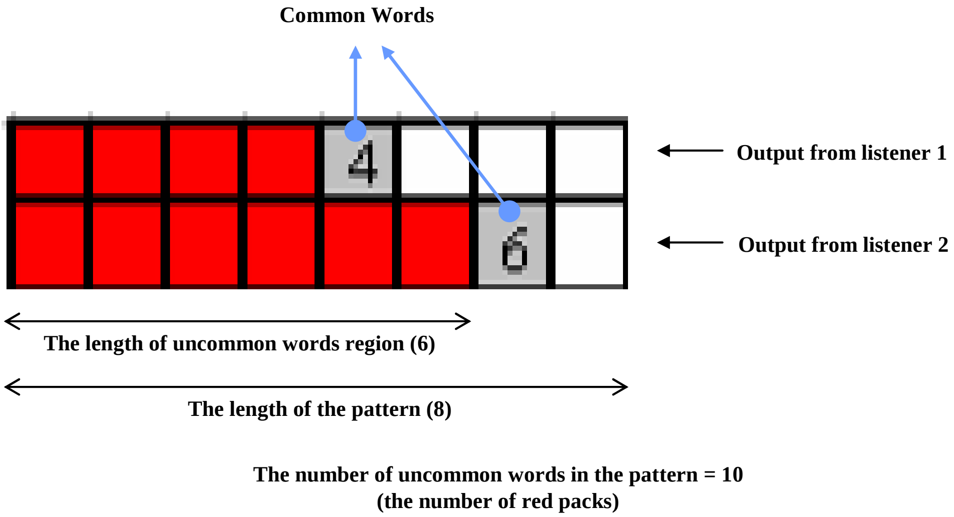

Common Words
4
6
Output from listener 1
Output from listener 2
The length of uncommon words region (6)
The length of the pattern (8)
The number of uncommon words in the pattern = 10
(the number of red packs)

Figure 10: 16-pack pattern for two inputs

2-input 16-pack patterns (figure10) which can accommodate 8 words from each of the recognition results were used for detecting common words in the pairs of recognition results. Each row receives words from each recognition results. The length of the pattern (8) was chosen based on the analysis result that the longest segment of uncommon words in all the recognition results of the preliminary experiment data is 7.

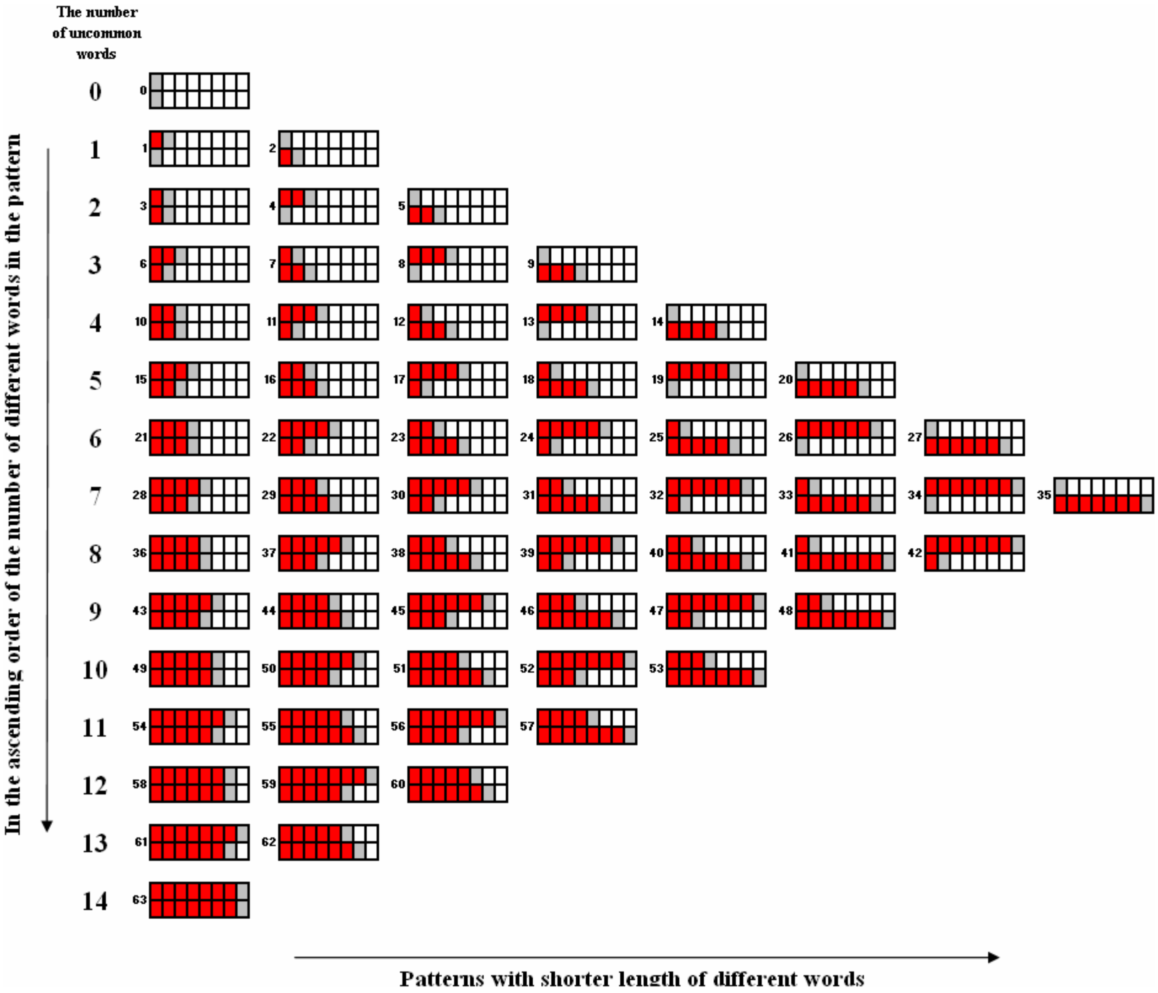


Figure 11: 64 patterns for input analysis

Figure 11 shows 64 patterns which are all the possible combinations that a pair of common words can be located in two 8-word phrases. The patterns are grouped by and then arranged in the order of the number of uncommon words in each pattern. And within each group, patterns are arranged in the ascending order of the length of uncommon words region. Therefore patterns with smaller number of uncommon words and shorter length of uncommon words region are used first in the pattern matching process.

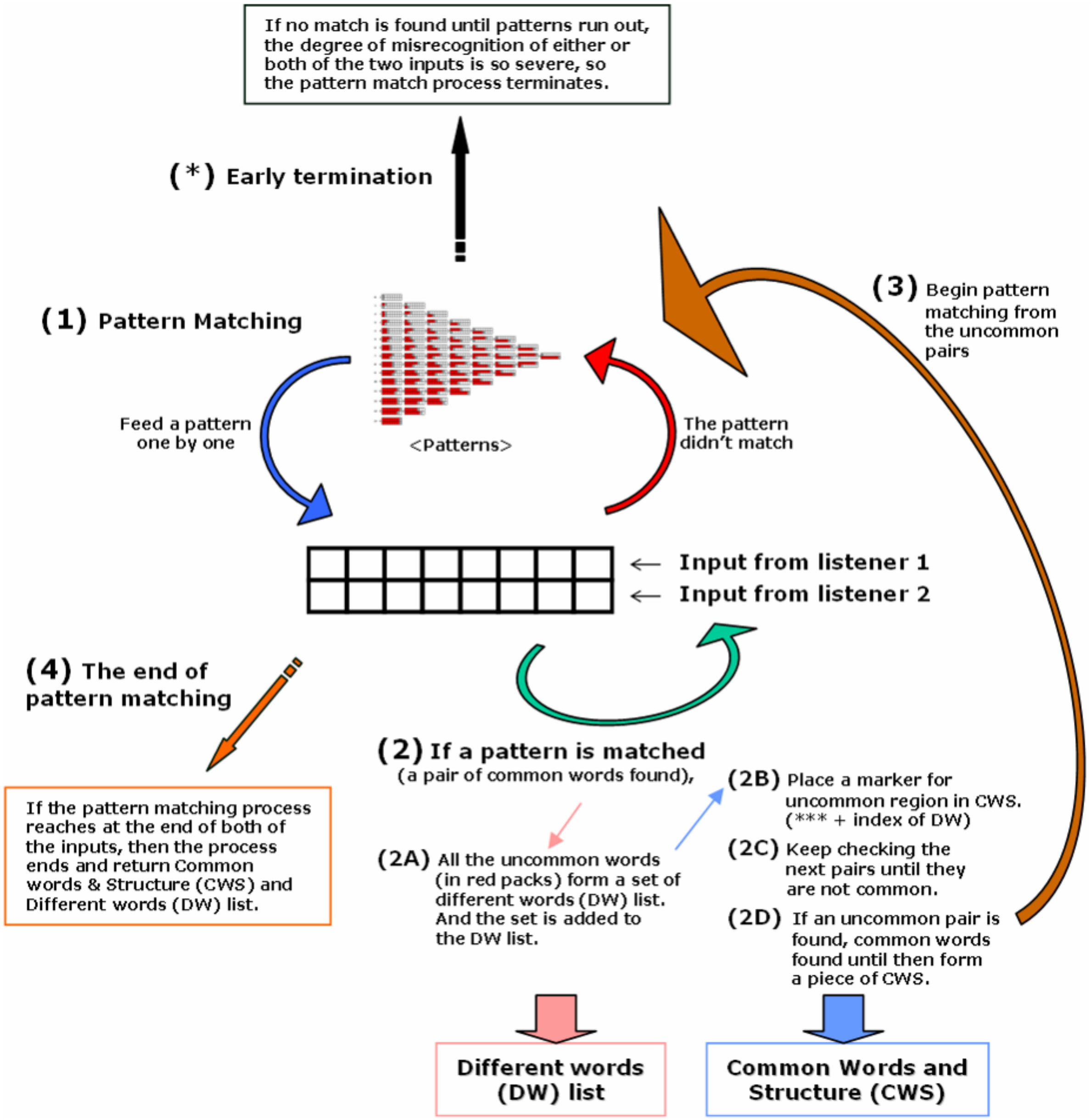

If no match is found until patterns run out, the degree of misrecognition of either or both of the two inputs is so severe, so the pattern match process terminates.
(*) Early termination
(1) Pattern Matching
Feed a pattern one by one
<Patterns>
The pattern didn't match
(3) Begin pattern matching from the uncommon pairs
← Input from listener 1
← Input from listener 2
(4) The end of pattern matching
If the pattern matching process reaches at the end of both of the inputs, then the process ends and return Common words & Structure (CWS) and Different words (DW) list.
(2) If a pattern is matched
(a pair of common words found),
(2A) All the uncommon words (in red packs) form a set of different words (DW) list. And the set is added to the DW list.
(2B) Place a marker for uncommon region in CWS. (*** + index of DW)
(2C) Keep checking the next pairs until they are not common.
(2D) If an uncommon pair is found, common words found until then form a piece of CWS.
Different words (DW) list
Common Words and Structure (CWS)

Figure 12: The progress of pattern matching

3.2.2 The progress of pattern matching

(1) Begin the pattern matching process with the first 8 words of each of recognition results (or next 8 words if comes from step 3) by comparing patterns one by one.

(*) If patterns run out without finding any match, then the pattern matching process terminates without returning any results. This means that the degree of misrecognition of either or both of the two inputs is severe.

(2) If a pattern is matched (a pair of common words found),

> (2A) All the uncommon words (red packs in patterns, if any) are added Different Words (DW) list as a set.
>
> (2B) Place a marker for uncommon word region (if uncommon words found) in Common Words & Structure (CWS). (*** + index of the set added to DW list)
>
> (2C) keep checking the next pairs of word by increasing indices of both inputs together one by one until an uncommon pair is found.
>
> (2D) if an uncommon pair is found, all the common words found until then form a piece of CWS and saved, and then go to step (3). If reached at the end of both inputs without an uncommon pair found, then all common words found until then form the last piece of CWS and saved, and the whole pattern matching process terminates.

(3) Go to step (1) and begin pattern matching process from the first pattern.

(4) If the pattern matching process reaches the end of both of the inputs, the pattern matching process ends. If there is (are) uncommon word(s) found until then, then do step (2A) and (2B). Finally the pattern matching process returns Common Words & Structure (CWS) and Different Words (DW) list.

3.3 Domain-specific Simple text-corpus database

A corpus is a big collected set of written or spoken texts normally used for linguistic research. A corpus is usually electronically processed and saved in order to make it more useful and add computerized efficiency for research.[Wikipedia 2006] The corpus used in this research is the collection of sentences or paragraphs speakers (users of distributed systems) may utter. The tasks given to the users during the preliminary experiment were to dictate 3 2-sentence paragraphs randomly collected from chapter 2 of a book [Kiyosaki 2000]. Therefore the domain of the corpus will be the book and the corpus will consists of all the sentences (paragraphs) found in chapter 1 and 2 of the book.

Target paragraphs in this research are the 3 2-sentence paragraphs the speakers actually spoke during the preliminary experiment. Candidate paragraphs are all other sentences (or paragraphs) in the corpus. In the early experiments in this paper, it will be tested to see how many common words & structures (CWS) can match their corresponding target paragraphs. The matching condition is true if a 2-sentence paragraph contains all pieces of a CWS in order. And then in later experiments, it will be tested how many of CWSs are able to correctly find their target sentences in the corpus (the targets are mixed with all other candidate paragraphs). If multiple target sentences are found for one CWS, then a selection process will compute a matching rate to determine the best matching target.

The corpus is a domain-specific simple text-corpus because the texts are collected from a specific domain which is a collection of written (not spoken) texts from a book, and electronically saved without any further processing.

The reason why simple text-corpus database is used in distributed listening system is because sentence recognition rate is still too low even after combining results of multiple recognizers. Only one listener (0.7%) was able to fully recognize a paragraph correctly during the preliminary experiment and only 7 combined results (5.2%) contain all the words in their target paragraphs. Moreover, these recognition rates are possible only when it is assumed that the contracted and expanded phrase are considered same (for example, “you’ll” = “you will”) and the recognition results are ideally combined. Therefore the sentence recognition rate is almost zero.

Therefore in this research, it is suggested and tested that the combined recognition results are compared with sentences in the collection of possible sentences users may speak. And the best matching sentence is selected as a right recognition result (target). If there are multiple matching sentences, then a selection process will be used to calculate matching rates of the sentences to choose the best matching sentence.

3.4 Normalization

The analysis of 135 pairs of speech recognition results collected from the preliminary experiment of Distributed Listening [Gilbert 2005] shows that there are many redundant white spaces, unnecessary special characters, and case sensitivity problems in the text of the speech recognition data. The existence of redundant spaces between words and same words with different cases (lower or upper), for example, “I’ll” and “i’ll” might result from minor mistakes in design of grammar or improper setting of return text values for the corresponding recognized words. And unexpected special characters like a period or a single quotation mark in the place of an apostrophe might result from minor mistakes while moving recognition data for analysis.

These mistakes can be considered very minor because the numbers are relatively small. But some kind of cleaning process is still required before the recognition results are sent to post-speech-recognition process, because they will negatively affect the string analysis and matching processes. An automated process is recommended for future experiments when transferring data from one medium to another. Furthermore, this result of the observation led to the conclusion that a same process must be performed for candidate (or target) paragraphs in a text-corpus database (a collection of candidate paragraphs) before they are involved in a matching process, because string analysis and matching must be done between words with same format. The format of text paragraphs saved in the database is more likely to be human-friendly (written in normal writing format), or has a possibility that it is not well formatted.

In this research, the cleaning and reformatting process is called “normalization”. Normalization changes all characters to lower case, removes all the redundant white spaces between words, remove all special characters except an apostrophe in contraction forms, for example, “they're” and replace a single quotation mark with an apostrophe.

These kinds of unexpected minor character problems will be very likely especially when the speech applications are multimodal systems which accept inputs from both speech and keyboard. For accurate and efficient string processing, a similar kind of normalization which makes text inputs well formed is very necessary for applications using speech recognition and natural language processing.

3.5 Experiment 1

In this experiment (illustrated in Figure 13), the pairs of speech recognition results returned by two listeners will be processed through normalization before sending them to the input analyzer (pattern matching) process. And then after input analysis, it is noted how many of the CWSs match their corresponding target paragraphs. Target paragraphs here mean the paragraphs the speakers actually spoke. If all the pieces of a CWS returned by the input analyzer are found in a candidate in order, the candidate is considered as the matching target. All of the 135 pairs of speech recognition results from the preliminary experiment will be used in the experiments.

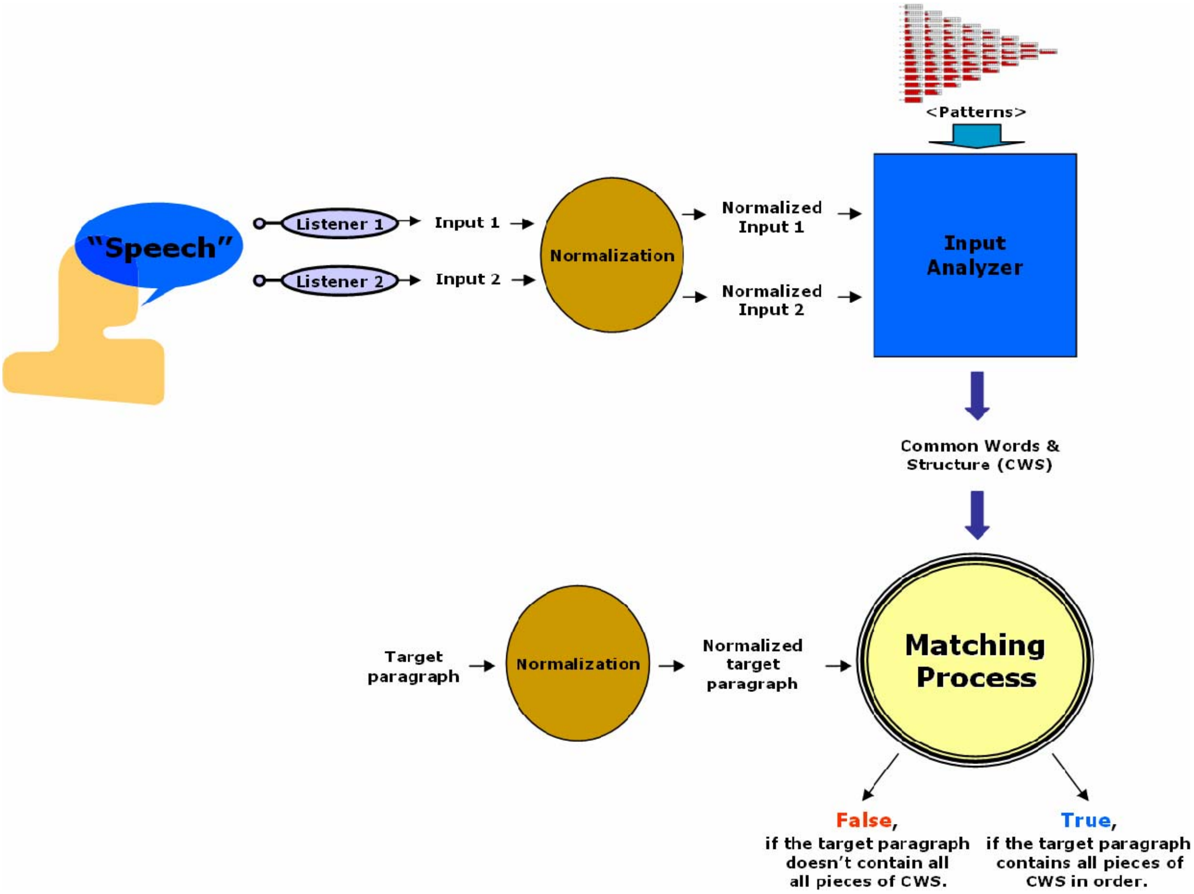

<Patterns>
"Speech"
Listener 1
Listener 2
Input 1
Input 2
Normalization
Normalized Input 1
Normalized Input 2
Input Analyzer
Common Words & Structure (CWS)
Target paragraph
Normalization
Normalized target paragraph
Matching Process
False, if the target paragraph doesn't contain all all pieces of CWS.
True, if the target paragraph contains all pieces of CWS in order.

Figure 13: Experiment 1 setup

3.5.1 The result of experiment 1

After adding the normalization process into the post-speech-recognition (PSR) process, 32 (27.8%) out of 135 CWSs were able to correctly detect their target paragraphs. The treatments included in the normalization seem to be very basic and trivial but they are very essential processes for correct and efficient string matching processes which will be performed later. When the normalization process is turned off, even though advanced CMR treatments were on which will be followed in this paper, the matching process simply returned 0% of matching rate. Therefore it can be said that the normalization process increased zero recognition rate up to 27.8%.

CHAPTER 4

IMPLEMENTATION AND EXPERIMENTS

4.1 Common misrecognition (CMR)

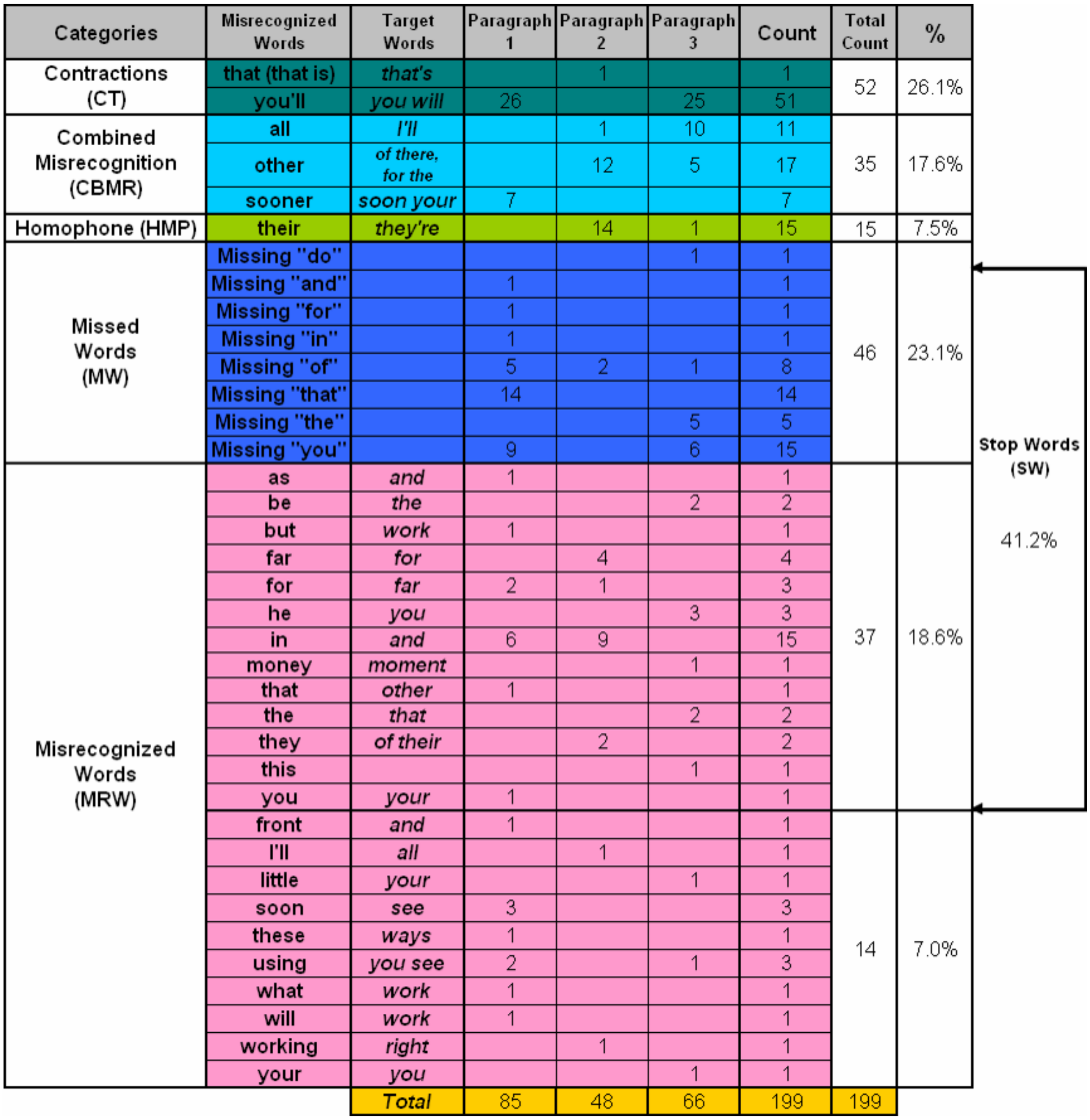
Common misrecognition (CMR) is commonly misrecognized words by both (all) listeners. The CMRs found in the analysis of the preliminary experiment can be classified into 6 categories: Contractions (CT), general Misrecognized words (MRW), Missed words (MW), Stop words (SW), Homophones (HMP), and Combined misrecognitions (CBMR) as seen in Table 4.

4.1.1 Contractions (CT)

4.1.1.2 Contractions in English

A contraction is a reduced form of usually two (rarely three) words. Mostly the first word is a pronoun and the second word is an auxiliary verb, or combination of auxiliary verb (or be verb) and negation (not). For example, "you'll" is the contraction form of "you will", and “aren’t” is the contraction form of “are not”. An apostrophe (') is inserted between the words of contraction forms. Contractions form a different word from their former words but they are same things. Contractions are more likely to be used in spoken language than in written language. [Wikipedia 2006]

.



Categories	Misrecognized Words	Target Words	Paragraph 1	Paragraph 2	Paragraph 3	Count	Total Count	%	
Contractions (CT)	that (that is)	*that's*		1		1	52	26.1%	
	you'll	*you will*	26		25	51			
Combined Misrecognition (CBMR)	all	*I'll*		1	10	11	35	17.6%	
	other	*of there, for the*		12	5	17			
	sooner	*soon your*	7			7			
Homophone (HMP)	their	*they're*		14	1	15	15	7.5%	
Missed Words (MW)	Missing "do"				1	1	46	23.1%	Stop Words (SW) 41.2%
	Missing "and"		1			1			
	Missing "for"		1			1			
	Missing "in"		1			1			
	Missing "of"		5	2	1	8			
	Missing "that"		14			14			
	Missing "the"				5	5			
	Missing "you"		9		6	15			
Misrecognized Words (MRW)	as	*and*	1			1	37	18.6%	
	be	*the*			2	2			
	but	*work*	1			1			
	far	*for*		4		4			
	for	*far*	2	1		3			
	he	*you*			3	3			
	in	*and*	6	9		15			
	money	*moment*			1	1			
	that	*other*	1			1			
	the	*that*			2	2			
	they	*of their*		2		2			
	this				1	1			
	you	*your*	1			1			
	front	*and*	1			1	14	7.0%	
	I'll	*all*		1		1			
	little	*your*			1	1			
	soon	*see*	3			3			
	these	*ways*	1			1			
	using	*you see*	2		1	3			
	what	*work*	1			1			
	will	*work*	1			1			
	working	*right*		1		1			
	your	*you*			1	1			
		Total	85	48	66	199	199		

Table 4: Common misrecognition collection (CMR)

4.1.1.3 Contractions in CMR

26.9% of CMRs found in the whole CWS collection are related to contraction problem. “you will” was recognized as “you'll” in 51 pairs of recognition results (26 from paragraph 1, 25 from paragraph 3), and “that’s” was recognized as “that is” in one recognition results (paragraph 2). This is the second largest category of CMR collection. The manner in which this problem is treated can contribute to significant increase of recognition results

4.1.1.4 Contraction or expansion?

It is difficult and sometimes ambiguous to expand contraction forms to their original forms (“you’ve” → “you have”), because many contraction forms are shared by two different combinations of words. For example, the contraction form of “I shall” and “I will” is “I’ll” and the contraction form of “he has” and “he is” is “he’s”. To expand contractions which have two possible expended forms is not simple. The first step to expand them can be to check the following words and choose the right one according to predefined rules. For example, “he’s going home” can be easily transformed to “he is going home”. But “you’ll have your money” can be either “you shall have your money” or “you will have your money”. The former is a promise and the later is a simple prediction. [The American Heritage 1996] To correctly expand a contraction form to either of two possible expanded forms, an advanced AI process which understands the context of situation or paragraph is required and sometimes it can be ambiguous when done even by humans.

Contrastively, it is much easier to contract words. By using a complete list of contraction forms and their expanded forms, two or three words can be simply replaced to their corresponding contraction form. Therefore, in this research the later method was applied and added as a later part of normalization process in order to increase the positive matching rate between CWSs and target paragraphs (or candidate paragraphs). All the normalized recognition results and normalized target or candidate paragraphs are processed through the contraction treatment which replaces all the word pairs found in the contractions list to the corresponding contraction forms.

4.1.2 Misrecognized words (MRW)

Misrecognized words are words commonly misrecognized by all listeners which do not belong to contractions, homophone, combined misrecognition, and missed words. Simply they are misrecognized words not belonging to any category. 25.6% of CMR collection is related to the general misrecognized words. 23 words are classified as general misrecognized words and many of them are commonly known stop words.

4.1.3 Missed words (MW)

Missed words are words commonly unrecognized by all listeners. So the listeners didn’t return anything for the speech utterances. 23.1% of CMR collection is related to the missed words. “do”, “and”, “for”, “in”, “of”, “that”, “the”, and “you” were commonly unrecognized (missed) by two listeners. The interesting thing is that all of them (98%) except one, “do”, are commonly known stop words.

Missed words problem is handled by stop words treatment described in the following section. Missed words problem is resolved by removing stop words in target or candidate paragraphs. Because the words are missed and not present in recognition results, the removal of stop words in candidate paragraphs may increase the chances of positive matching results between CWSs and paragraphs.

4.1.4 Stop words

Stop words is a term usually used in the search engine area. It means insignificant words whose occurrence in a certain domain such as a database system, articles, etc is so frequent that they are excluded in search operations in order to save space and speed up the operations. For example, “the”, “is”, “of”, “and”, “a” are very common stop words. [Wikipedia 2006] [Sullivan 2006] Every domain has its own stop words list because the context and words it has are different from others.

The reason why stop words is introduced here is because the observation of CMR collection led to the inference that most of Missed words and Misrecognized words in CMR collection are commonly known stop words. The union of missed words and misrecognized words is the largest portion of CMR collection. The way in which these words are treated in conjunction with the similarity between the two different word collections (missed and misrecognized) and stop words may greatly improve the overall result of this research.

[Bray 2003] introduced how to statistically determine stop words of a certain domain. [Table 5b] is the stop words list brought from [Bray 2003]. 42.3% of missed words and misrecognized words are stop words in the list. But the frequency of the occurrence of the 42.3% takes up 74.2% of the total frequency of missed words and misrecognized words. If a large portion of the words matches the stop words list from other domain, there is a high possibility that the percentages will go up when the stop words list from its own domain is used for the counting. Therefore a new stop words list was made using the statistical method presented in [Bray 2003].

The three target paragraphs used for the preliminary experiment of distributed listening [Gilbert, 2005] are three pairs of sentences brought from chapter 2 of “Rich Dad, Poor Dad” [Kiyosaki 2000]. The book contains 57702 words comprising 4832 unique words. [Table 5a] is the stop words list of [Kiyosaki 2000] which was made using the simple statistical method described in [Bray 2003]. The 23 words in the list are just 0.5% of all the unique words but they take up about 33% of the total word occurrence of the book. 42.3% of missed words and misrecognized words are found in the new stop words list and they take up 76.3% of the total frequency of missed words and misrecognized words. This is a smaller increase than expected.

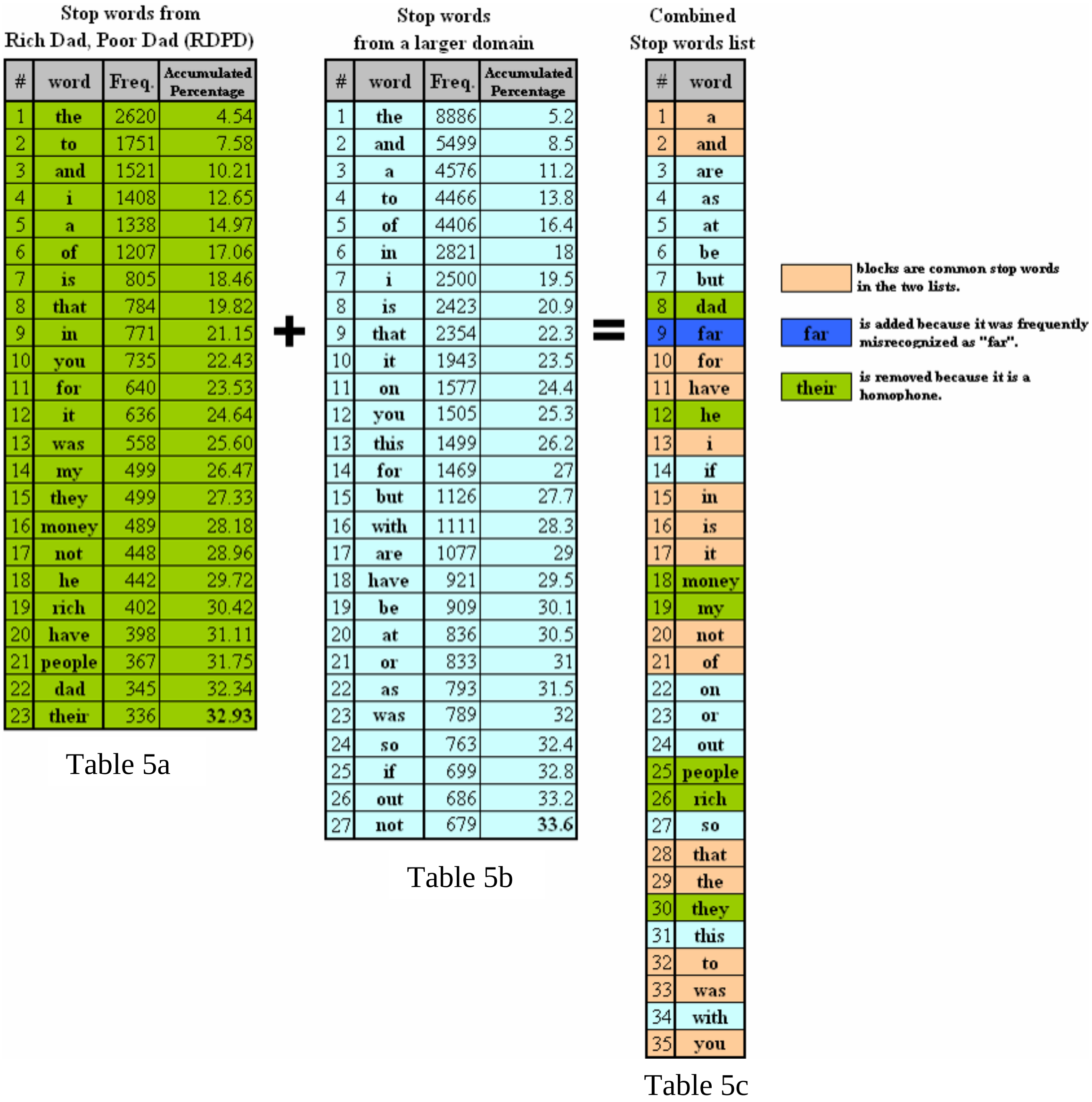
The stop words list (Table b) in [Bray 2003] is about three times larger in domain than [Kiyosaki 2000]. If a domain is much larger, its stop words list may contain more general and insignificant words which can be used in general. So in this research the union of the two stop words list was determined to be used for the experiment. It can be said that the new mixed stop words list [Table 5c] is the union of a general stop words list and a domain-specific stop words list. The new list contains 35 words (8 from [Kiyosaki 2000] list, 12 from [Bray 2003] and 15 in common). 57.7% of missed words and misrecognized words are in the new mixed stop words list and they take up 84.5% of the total word occurrence of missed words and misrecognized words. Later in this research “their” will be removed from the list because it is a combined misrecognition word (CBMR) which will be explained later, and “far” is added because it was frequently misrecognized for “for” as an example of spoken stop words.

4.1.4.1 Stop words treatment

The treatment for the missed and misrecognized words determined to be used in this research is “removal” of all the stop words as they were all ignored during search operations. All the stop words found in CWSs, DW lists and candidate paragraphs were removed before they are sent to matching process.

4.1.4.2 Spoken stop words

Stop words are not perfectly relevant to speech recognition research, but because of their strong similarity with common misrecognition words found in preliminary data collected from [Gilbert 2005], they were chosen to be used in this research as substitutes for common misrecognition words because there are no such data collections known to the public.



Stop words from Rich Dad, Poor Dad (RDPD)

#	word	Freq.	Accumulated Percentage
1	the	2620	4.54
2	to	1751	7.58
3	and	1521	10.21
4	i	1408	12.65
5	a	1338	14.97
6	of	1207	17.06
7	is	805	18.46
8	that	784	19.82
9	in	771	21.15
10	you	735	22.43
11	for	640	23.53
12	it	636	24.64
13	was	558	25.60
14	my	499	26.47
15	they	499	27.33
16	money	489	28.18
17	not	448	28.96
18	he	442	29.72
19	rich	402	30.42
20	have	398	31.11
21	people	367	31.75
22	dad	345	32.34
23	their	336	**32.93**

Table 5a

+

Stop words from a larger domain

#	word	Freq.	Accumulated Percentage
1	the	8886	5.2
2	and	5499	8.5
3	a	4576	11.2
4	to	4466	13.8
5	of	4406	16.4
6	in	2821	18
7	i	2500	19.5
8	is	2423	20.9
9	that	2354	22.3
10	it	1943	23.5
11	on	1577	24.4
12	you	1505	25.3
13	this	1499	26.2
14	for	1469	27
15	but	1126	27.7
16	with	1111	28.3
17	are	1077	29
18	have	921	29.5
19	be	909	30.1
20	at	836	30.5
21	or	833	31
22	as	793	31.5
23	was	789	32
24	so	763	32.4
25	if	699	32.8
26	out	686	33.2
27	not	679	**33.6**

Table 5b

=

Combined Stop words list

#	word
1	a
2	and
3	are
4	as
5	at
6	be
7	but
8	dad
9	far
10	for
11	have
12	he
13	i
14	if
15	in
16	is
17	it
18	money
19	my
20	not
21	of
22	on
23	or
24	out
25	people
26	rich
27	so
28	that
29	the
30	they
31	this
32	to
33	was
34	with
35	you

Table 5c

🟧	blocks are common stop words in the two lists.
far	is added because it was frequently misrecognized as "far".
their	is removed because it is a homophone.

Table 5: Stop words lists

The characteristics of stop words are insignificance in search operations, very high frequency in given domains, mostly meaningless in contexts but words with grammatical roles. Additionally for speech recognition research, one more characteristic can be added, which is “misrecognition-prone”. In English, words carrying very minor or no meaning but grammatical roles such as prepositions, articles, etc are usually spoken very short and unstressed to emphasize other important words such as nouns and verbs, etc in intonation. [Celik 2001] This factor may make speech recognizers easily miss or misrecognize such words.

Consequently, a new term is introduced in this research, which is “spoken stop words”. Spoken stop words can be defined as insignificant and misrecognition-prone words with very high frequency in a domain. Stop words make search operations stop, spoken stop words likewise make speech recognizers stop (misrecognize).

It will be very helpful in increasing performance and accuracy of distributed listening systems to remove information like stop words which seem unnecessary or are incorrect. But if too much information is removed, there will be some trade-off. While the removal process increases the accuracy of earlier processes, it may damage the accuracy of later processes because the removal process also eliminates correct information. Therefore it is very important to collect correct spoken stop words lists for specific domains or recognizers and restrict the amount of information to be removed.

4.1.5 Homophone (HMP)

Homophones are words with the same pronunciation but different spellings and meanings. For example, “write”, “rite”, “right”, and “wright” are homophones. 7.8% of CMR collection is related to homophone problem. “they’re” was recognized as “their” which has same pronunciation as “they’re” in 15 CWSs (14 from paragraph 2 and 1 from paragraph 3). “they’re” has same pronunciation as “there” and “their”, therefore it is very likely that speech recognizers recognize “they’re” as one of the two words.

4.1.5.1 HMP treatment

In this research, when a piece of CWS doesn’t match with a region of a candidate paragraph during the matching process, a homophone treatment process is initiated to check the presence of homophones in the unmatched region. If the region has one or more homophones, the homophone treatment process brings the corresponding homophone set(s) from the homophone list, and then generates alternative phrases by replacing the word in the piece of CWS with its homophone siblings. The alternative phrases are used to match the previously unmatched region of a candidate paragraph. If no homophone is found or none of the alternative phrases matches the region, then combined misrecognition (CBMR) treatment process is initiated. A homophone list [Cooper 2001] consisting of 706 homophone sets (1529 homophones) were used to populate the HMP database.

4.1.6 Combined misrecognition (CBMR)

Combined misrecognitions (CBMR) are single words recognized for a phrase consisting of two words. 17.6% of CMR collection is related to CBMR problem. “i’ll” and “soon your” were misrecognized as “all” and “sooner” respectively in 18 CWSs. And “of there” and “for the” were misrecognized as “other” in 16 CWSs. An observed characteristic of CBMRs is that the pronunciations of partial parts of CBMRs are similar to the pronunciations of partial parts of the corresponding target phrase. For example, “sooner”

(CBMR of “soon your”) has a full pronunciation of “soon” and the last part (“r”) of the pronunciation of “your”.

4.1.6.1 CBMR treatment

In the similar manner with the homophone (HMP) treatment, when a piece of CWS doesn’t match with a region of a candidate paragraph even after HMP treatment during matching process, a CBMR treatment process is initiated to check the presence of CBMR in the unmatched region. If a possible CBMR is present in the region, the CBMR treatment process fetches the corresponding CBMR set(s) from CBMR list, and then generates alternative phrases by replacing the CBMR in the piece of CWS with its CBMR sibling(s). And then the alternative phrases involve in matching process. If none of the alternative phrases matches the region during CBMR process, another process is invoked which makes all the possible combinations of homophones and CBMRs gathered in the previous processes and then produces another list of alternative phrases by replacing homophone or CBMR words with the corresponding siblings in each combination.

If the region didn’t match during the previous three processes or contains neither a homophone nor a CBMR, the region of a candidate is considered unmatched. Then the matching process shifts to the next region by skipping the first word and adding the word right next to the region to the end. The shifting continues until the matching process finds a matching region in a candidate paragraph or the process reaches the end of the paragraph. If a region is matched, the skipped words until then form an uncommon words region (UWR).

4.2 Matching process

During matching process, all the pieces of a CWS are compared with candidate sentences (paragraphs). If a candidate sentence (paragraph) contains all the pieces of a CWS in order, then the candidate is considered as its corresponding matching target. Figure 14 illustrates the matching process. Regions matching pieces of a CWS is called a common words region (CWR) and all other regions of a candidate (or target) are called uncommon words region (UWR). Words in UWR are compared with words in a corresponding set of different words (DW) list later during selection process, when there are multiple target sentences (paragraphs) returned by matching processing.

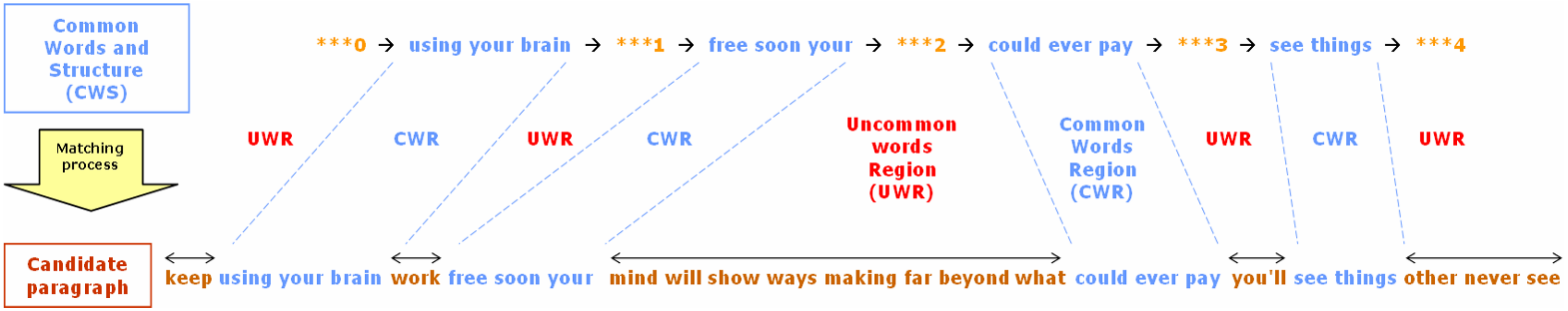

Common Words and Structure (CWS)
Matching process
Candidate paragraph
***0 → using your brain → ***1 → free soon your → ***2 → could ever pay → ***3 → see things → ***4
UWR
CWR
UWR
CWR
Uncommon words Region (UWR)
Common Words Region (CWR)
UWR
CWR
UWR
keep using your brain work free soon your mind will show ways making far beyond what could ever pay you'll see things other never see

Figure 14: Matching process

4.3 Experiment 2

Figure 15 illustrates the setup of experiment 2. In experiment 2, it is tested that how many of CWSs can correctly match their target paragraphs after adding common misrecognition (CMR) treatments (contraction (CT), stop words removal (SW), homophone (HMP) and combined misrecognition (CBMR) treatments) to the experiment 1 setup. Contraction treatment is added as a later part of normalization and stop words removal processes are added right after input analysis of recognition results and right after normalization of candidate paragraphs. HMP and CBMR treatment processes are added to the matching process.

A stop words removal process for speech recognition results is added after the input analysis because it helps preserve more information. If stop words are removed before the input analysis, it affects the result of input analysis. More specifically it changes the structure of CWSs in some cases. When words in a set of different words (DW) list are all stop words, they will all be removed during the removal process. If they are removed before the input analysis, the two adjacent pieces of CWSs become one piece because there are no words between them. If they are removed after the input analysis, the position of the set of DW is still preserved even though all the words in it are removed and the two adjacent pieces of CWS stay separated. The result of experiment 2 proves that stop words removal performed after input analysis affected about 11% of positive matching result.

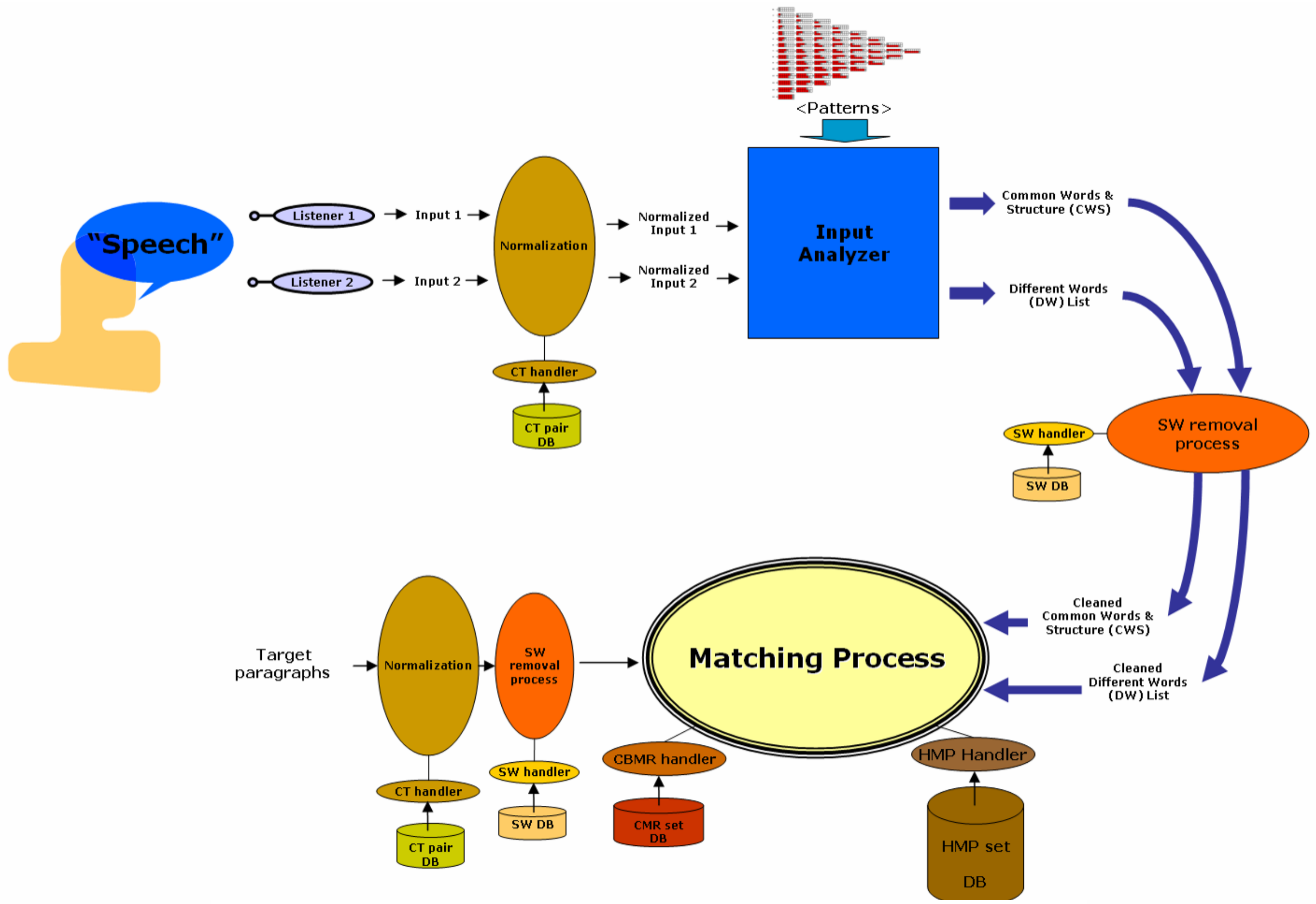
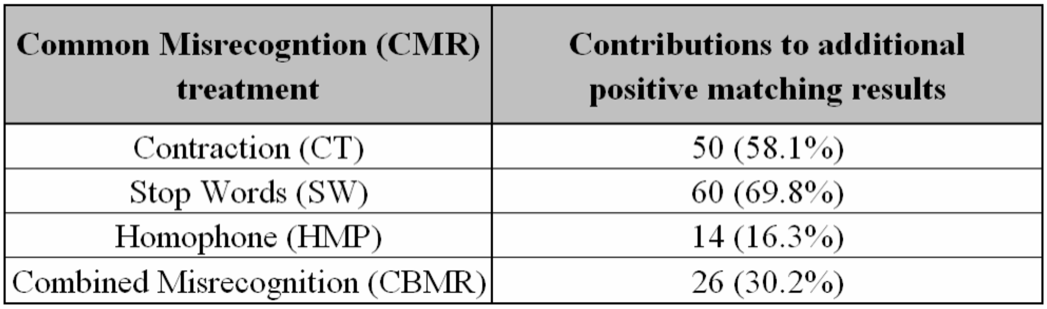


Figure 15: Experiment 2 setup

4.3.1 The result of experiment 2

118 out of 135 CWSs (87.5%) were able to correctly match their target paragraphs and none of them matched any other paragraphs as their target. This is 369% improvement with respect to the result (32) of experiment 1. 86 more CWSs were able to match their target with assistance of advanced CMR treatments (contractions, stop words, homophones, and combined misrecognitions). Their contribution rates are listed in Table 6.



Common Misrecogntion (CMR) treatment	Contributions to additional positive matching results
Contraction (CT)	50 (58.1%)
Stop Words (SW)	60 (69.8%)
Homophone (HMP)	14 (16.3%)
Combined Misrecognition (CBMR)	26 (30.2%)

Table 6: Contribution rates of each CMR treatment

The highest contribution rate (about 70%) of stop words treatment proves that speech recognition systems greatly suffer from spoken stop words problem, and how they are dealt with will significantly improve recognition rates of their systems. The very high contribution rate of contraction treatment indicates that contraction problem is very common in speech recognition systems and the treatment should be used as one of essential processes in speech recognition. And the rates of homophone (HMP) and combined misrecognition (CBMR) treatment show that the treatments are very useful though their improvement rates are not as great as rates of CT or SW treatments. HMP and CBMR treatments can be considered as more positive and advanced process than other processes because no information is removed like SW process and incorrect information is corrected based on cumulated information.

4.4 Experiment 3 (Overall design of interpreter)

Figure 16 illustrates the setup of experiment 3. The corpus database and a selection process are added to the setup of experiment 2. The corpus database will feed candidate paragraphs to the matching process and the selection process will select one target paragraph by computing matching rates when matching process returns multiple target paragraphs. The candidate paragraphs are supplied to the matching process through normalization (+ contraction) treatment and stop words removal processes.

In experiment 3, two sets of paragraph collections will be used. The first one is a collection of 2-sentence paragraphs collected from chapter 1 and 2 of [Kiyosaki 2000]. First, single sentence are extracted and then too short (< 4) or too long sentences (> 45) are excluded. 4 is rounded half of the number of words in the shortest sentence (7) of the three target paragraphs and 45 is the sum of the number of words in the longest paragraph (35) and the difference between the longest paragraphs (35) and the shortest one (25). The order the sentences were written in is preserved because they are all semantically related and the written order of the three paragraphs are also preserved. 420 single sentences are collected including sentences of the three target paragraphs. They are paired up to make 2-sentence paragraphs. The sentence in the order of “A B C D E G H …” forms 2-sentence paragraphs of “A+B B+C C+D D+E E+F F+G G+H …”. So finally, 839 2-sentence paragraphs are created. The second collection is 1000 2-sentence paragraphs collected from [Sykes and MeGregor 2001] in the same way used for the first collection of 2-sentence paragraphs. The second collection will be added to the corpus database as additional candidate paragraphs in later test. First it will be tested how many CWSs can correctly match their targets and how many of them match multiple paragraphs, when the targets are mixed with the first corpus (836 paragraphs), and then the same experiment will be repeated after adding the second corpus (1000 paragraphs) as a paragraph collection from other domain to the first one. Therefore the three target paragraphs are mixed with 1836 other paragraphs.

The selection process calculates the matching rates of paragraphs to distinguish the best matching target when multiple targets are returned by the matching process. Figure 17 illustrates the calculation of matching rate. Common misrecognition (CMR) processes are also involved in the matching rate calculation of the selection process. When there are words unmatched in uncommon words region (UWR) of target paragraphs even after word matching, it is checked whether or not the remaining words are homophone (HMP), or a part of contraction. If the remaining word is a homophone, the HMP treatment process checked if there is a sibling of the homophone in the corresponding set of different words (DW) list. If its HMP sibling is found, the remaining word is considered matched. If the remaining word is a part of contraction (for example, “you” is a part of “you’ll”) and there is its corresponding contraction in the set of DW list, then it is counted as one matched.

4.4.1 The result of experiment 3

The result of experiment 3 using the first corpus is exactly same as the result of experiment 2. This means that even though the three target paragraphs were mixed with 836 other paragraphs, exactly the same 118 CWSs were able to correctly find their target and none of them matched any other paragraphs as their target (no multiple targets). Therefore the selection process was never used. And the second test after adding the second corpus yielded exactly same result as the first test of experiment 3.

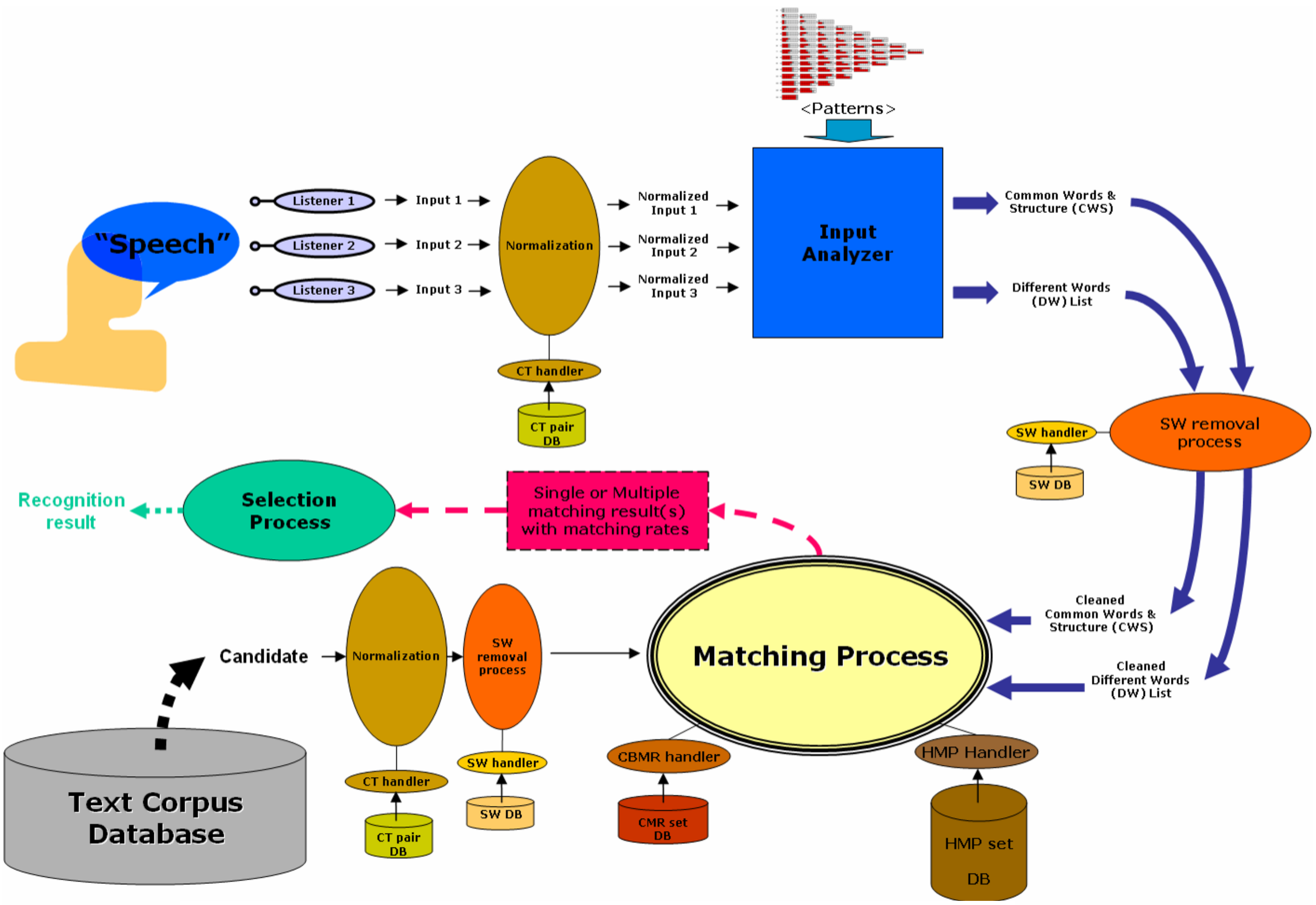
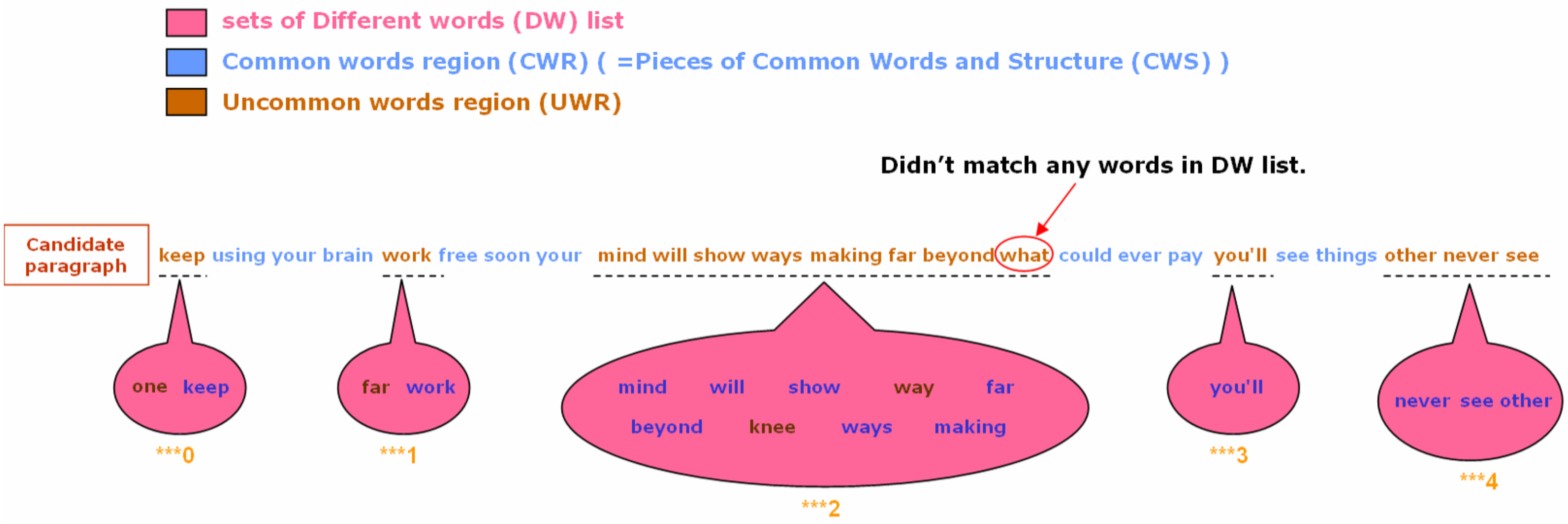


Figure 16: Experiment 3 setup (Overall design of interpreter)


sets of Different words (DW) list
Common words region (CWR) (=Pieces of Common Words and Structure (CWS))
Uncommon words region (UWR)
Didn't match any words in DW list.
Candidate paragraph
keep using your brain work free soon your mind will show ways making far beyond what could ever pay you'll see things other never see
one keep
***0
far work
***1
mind will show way far beyond knee ways making
***2
you'll
***3
never see other
***4

*Word matching happens locally.

$$**\text{Matching rate} = \frac{\text{Total number of words in CWS + Total number of words matched words in DW list}}{\text{Total number of words in candidate phrase}}$$

$$= (11+13)/25 = 24/25\ (96\%)$$

Figure 17: Matching process

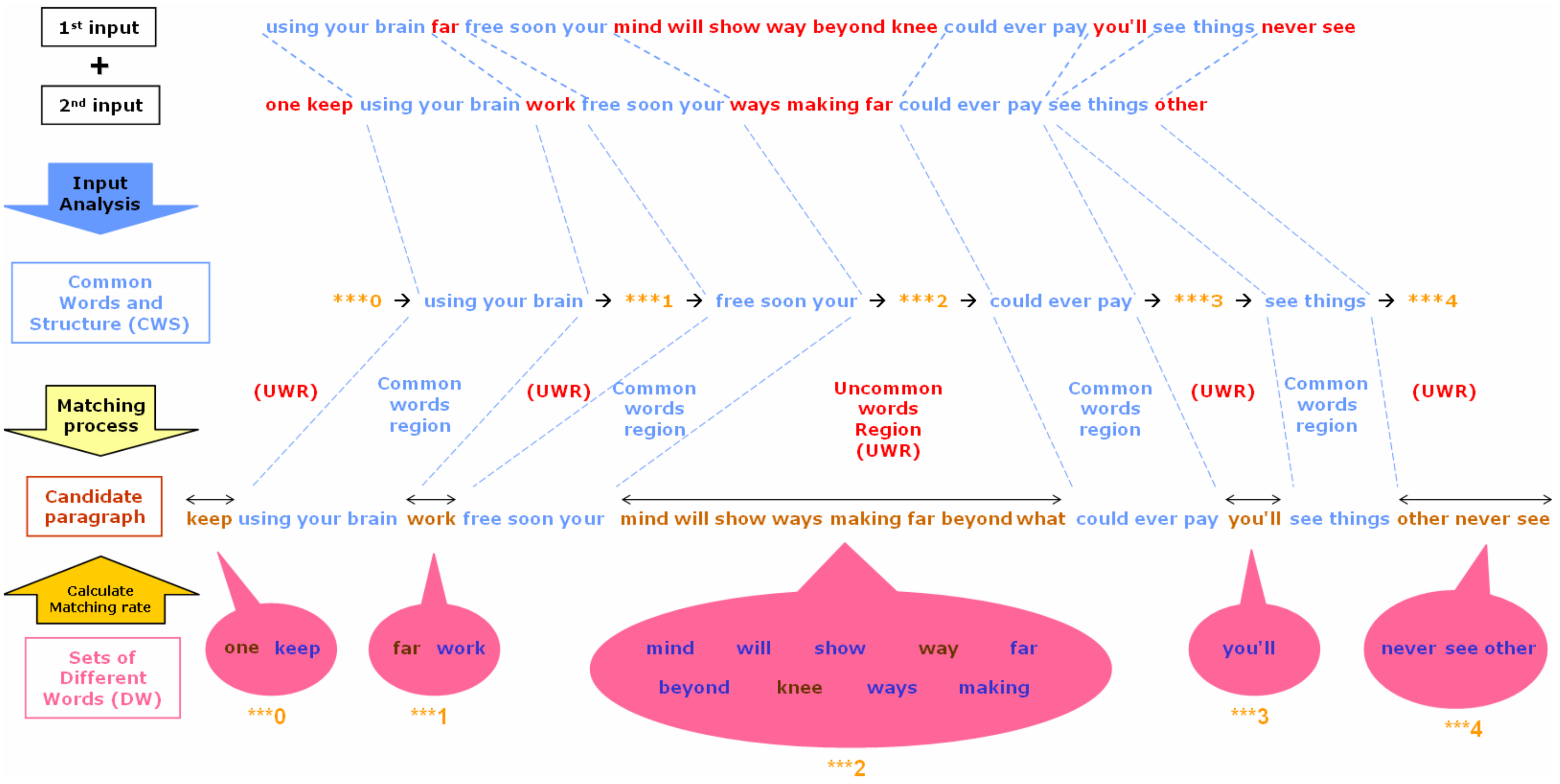

1st input
+
2nd input
using your brain far free soon your mind will show way beyond knee could ever pay you'll see things never see
one keep using your brain work free soon your ways making far could ever pay see things other
Input Analysis
Common Words and Structure (CWS)
***0 → using your brain → ***1 → free soon your → ***2 → could ever pay → ***3 → see things → ***4
Matching process
(UWR)
Common words region
(UWR)
Common words region
Uncommon words Region (UWR)
Common words region
(UWR)
Common words region
(UWR)
Candidate paragraph
keep using your brain work free soon your mind will show ways making far beyond what could ever pay you'll see things other never see
Calculate Matching rate
Sets of Different Words (DW)
one keep
***0
far work
***1
mind will show way far beyond knee ways making
***2
you'll
***3
never see other
***4

Figure 18: Overall interpretation progress

CHAPTER 5

CONCLUSIONS

For more than a decade, a significant number of research efforts were made to improve speech recognition rates by using multiple speech recognizers. The purpose of using multiple speech recognizers is to make a best composite output out of several incomplete recognition results by complementing each other’s result. Much of the research resulted in positive improvement in word and sentence recognition rates but the improved rates are still below usable level.

The analysis of the preliminary experiment data of distributed listening revealed that all the correct words of a speech may be collected when 4 or slightly more listeners are combined, and proved that words agreed by all listeners (speech recognizers) are not always correct. About 63% of combined recognition results contain at least one common misrecognition.

In this research, several treatment processes were used in order to overcome the common misrecognition (CMR) problems in multi-recognizer systems, and a corpus database was used as an example of a collection of sentence users may speak in order to overcome low sentence recognition rates by searching the best matching sentence in the database.

The highest contribution rate (about 70%) of stop words treatment proves that speech recognition systems greatly suffer from spoken stop words problem, and removing the insignificant words can greatly improve speech recognition rates. The very high contribution rate (58%) of contraction treatment indicates that the contraction problem is very common in speech recognition systems. And the rates (16% and 30%) of homophone (HMP) and combined misrecognition (CBMR) treatment show that the treatments are very useful. HMP and CBMR treatments are positive and advanced process than other processes because no information is removed like SW process and incorrect information is corrected based on cumulated information.

The normalization process which removes redundant information and reformats string information is very essential, because it enhances accuracy and efficiency of string comparison and matching tasks in post-speech recognition processes. The use of simple text-corpus, which is a collection of sentences (paragraphs) user may speak, is a very crucial part in this research. It was proved that common words & structure (CWS), which is a composite of information made out of multiple recognition results, was able to correctly match their corresponding target paragraphs even when mixed with about 800 other paragraphs from the same domain and 1000 paragraphs from other domain. This proves that the composite information is enough to distinguish its correct target information. Therefore, these results strongly support that the use of a collection of sentence users may speak is a good way to overcome low sentence recognition rate in speech recognition systems.

The pattern matching approach to combine multiple recognition result was used in this research. The approach effectively collected and separated common (CWS) and uncommon (Different Words) information out of multiple recognition results. Only 3 (2.2%) of the CWSs were affected by the combining order of recognition results and the changed structure didn’t affect the overall matching result.

In addition to the detective story introduced in distributed listening [Gilbert 2005], CSI (Crime Scene Inspector) teams which assist the detective with advanced skills and Scenarios built upon based on their collective experience and knowledge can be added to the story as analogies of the CMR (common misrecognition) treatments and the simple text-corpus.

Mainly normalization, pattern matching, common misrecognition (CMR) treatment and simple text-corpus were used to improve sentence recognition rates of multi-recognizer systems in this research. Overall the results (87% of sentence recognition rate) of the experiments strongly support that the processes can greatly improve speech recognition rates of multi-recognizer systems.

BIBLIOGRAPHY

Barry, T., Solz, T., Reising, J. and Williamson, D., **The simultaneous use of three machine speech recognition systems to increase recognition accuracy**, In Proceedings of the IEEE 1994 National Aerospace and Electronics Conference, vol.2, pp. 667 - 671, 1994.

Bray, T. **On Search: Stopwords** [Online] Available http://www.usa.net/~vinced/home/better-writing.html, July 2003

Celik, M., **Teaching English Intonation to EFL/ESL Students** [Online] Available http://iteslj.org/Techniques/Celik-Intonation.html, The Internet TESL Journal, Vol. VII, No. 12, December 2001.

Coopers, A., **Alan Cooper's Homonyms** [Online] Available http://www.cooper.com/alan/homonym.html, Alan Cooper, 2001

Cristoforetti, L., Matassoni, M., Omologo, M. and Svaizer, P., **Use of parallel recognizers for robust in-car speech interaction**, In Proceedings of the IEEE International Conference on Acoustic, Speech, and Signal Processing [ICASSP 2003], Hong-Kong, 2003.

Deng, L. and Huang, X., **Challenges in adopting speech recognition**, Communications of the ACM, vol. 47, no. 1, pp. 69-75, January 2004.

Fiscus, JG., **A post-processing system to yield reduced word error rates: Recognizer Output Voting Error Reduction (ROVER)**. Proc IEEE Workshop on Automatic Speech Recognition and Understanding, p 347–354, 1997.

Gavrilov, L. A. and Heuveline, P., **Aging of Population** [Online] Available http://longevity-science.org/Population_Aging.htm, The encyclopedia of population, New York, Macmillan reference USA, 2003.

Gilbert, J. E., **Distributed Listening** Research. In SpeechTEK West: Proceedings of AVIOS Speech Technology Track (pp. 1 – 10). San Francisco, California, 2005.

Kiyosaki, R. T. and Lechter, S. L., **Rich Dad, Poor Dad:** What the Rich Teach Their Kids about Money - That the Poor and Middle Class Do Not, Warner business books, April 2000.

Larsen, L. Bo., Brøndsted, T., H., Dybkjær, Dybkjær, L., Music, B. and C. Povlsen., **Spoken Language Dialog Systems**, Report 1, September 1992.

Pallett, D. S., Fiscus, J. G., Garofolo, J. S., Martin, A. and Przybocki, M., **1998 broadcast news benchmark test results: English and non-English word error rate performance measures.** In DARPA Broadcast NewsWorkshop, Hernon, *VA*, February 1999.

Pinker, S., **The language instinct** (New York: Morrow, 1994)

Schwenk, H. and Gauvain, J., **Improved ROVER using Language Model Information**, In ISCA ITRW Workshop on Automatic Speech Recognition: Challenges for the new Millennium, Paris, pp. 47–52, Sept 2000.

Schwenk, H., Gauvain, J., **Combining multiple speech recognizers using voting and language model information.** Proc 6th ICSLP, p 915–918, Oct 2000.

Stevenson, B. and McQuivey, J., **The aging of the US population and its impact on computer use** [Online] Available http://www.microsoft.com/enable/research/agingpop.aspx, The Market for Accessible Technology—The Wide Range of Abilities and Its Impact on Computer Use**,** Study commissioned by Microsoft, conducted by Forrester research, Inc., 2003.

Sullivan, D., **What are stop words?** [Online] Available http://searchenginewatch.com/showPage.html?page=2156061, SearchEngineWatch.com – the source for search engine marketing, 2006.

Sykes, D. A. and McGregor, J. D., **Practical guide to testing object-oriented software**, Addison-Wesley professional, March 2001.

The American Heritage® Book of English Usage, **A Practical and Authoritative Guide to Contemporary English** [Online] Available http://www.bartelby.net/64/C001/056.html, 1996.

Wikipedia, the free encyclopedia, **Contraction (grammar)** [Online] Available http://en.wikipedia.org/wiki/Contraction_%28grammar%29, June 2006