

TURKISH DICTATION SYSTEM FOR RADIOLOGY AND BROADCAST NEWS
APPLICATIONS

by

Ebru Arisoy

B.S. in E.E., Boğaziçi University, 2002

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Master of Science

Graduate Program in Electrical and Electronics Engineering
Boğaziçi University

2004

TURKISH DICTATION SYSTEM FOR RADIOLOGY AND BROADCAST NEWS
APPLICATIONS

APPROVED BY:

Assoc. Prof. Levent M. Arslan
(Thesis Supervisor)

Assist. Prof. Hakan Erdoğan

Prof. Bülent Sankur

DATE OF APPROVAL: 09.07.2004

ACKNOWLEDGEMENTS

I would like to thank Assoc. Prof. Levent M. Arslan for his kindness and support during this difficult process. I could not complete this thesis without his diligent advisement, creative ideas, and the inspiration he gave. He is not only an excellent advisor and an enthusiastic academician but also an encouraging teacher and a supportive friend.

I would also like to thank Prof. Bülent Sankur for his help and participation in my committee. His motivating contribution behind the anxiety he fostered was very helpful and is extremely valuable to me. I would like to thank Assist. Prof. Hakan Erdoğan for his gentle attitude and participation in my thesis committee.

My valuable thanks go to my family for their limitless tolerance. I also thank my little brother, Doğuş for his help in corpus preparation in exchange of two packs of ice cream. Also special thanks to my dear sister Esra for her contribution in radiological data collection process.

I thank to my friend Suna. She diligently stood my continuous grumbling and provided her unique friendship under every condition. Even though she was not familiar to my research area at the beginning, she put her labor into every line of this thesis. I would also like to thank my friend Suncem for washing all my coffee cups while I was consuming extreme amount of coffee during sleepless nights. Although, she can not have a chance to read this acknowledgement, I would like to thank my little sweet cat Meze Kepece for the enjoyment she gave.

Also, I would like to thank to my friends Esra, Selin, Dilek, Oktay, the day of gold members, the people in BUSİM for their contribution during my data collection. Finally, I would like to thank Hacettepe University Radiology Department for their help in supplying radiological reports.

ABSTRACT

TURKISH DICTATION SYSTEM FOR RADIOLOGY AND BROADCAST NEWS APPLICATIONS

In this thesis, we have designed a Turkish dictation system for Radiology and Broadcast news applications. Turkish is an agglutinative language with free word order. These characteristics of the language result in the vocabulary explosion and the complexity of the N-gram language models in speech recognition. In order to alleviate this problem, we propose a task-specific, radiology, dictation system. Using classical word-based language models, we achieve 87.06 per cent recognition performance with a small vocabulary size in a speaker independent radiology speech recognition system. However, the same system results in 46.29 per cent recognition rate for the broadcast news dictation due to the large number of out-of-vocabulary (OOV) words. Therefore, we parse some of the words to smaller recognition units like stems, endings and morphemes, and introduced these smaller units and the unparsed words to the speech recognizer as lexicon entries. This time, we manage to overcome to the problem of large number of OOV words with a moderate vocabulary size and get better estimates for the N-gram language models. However, best recognition result is in the word-based language model.

ÖZET

RADYOLOJİ VE HABER UYGULAMALARI İÇİN TÜRKÇE DİKTE SİSTEMİ

Bu tezde, radyoloji ve haber uygulamaları için türkçe dikte sistemi tasarlanmıştır. Türkçe sondan eklemeli bir dildir ve serbest kelime dizilimi vardır. Dilin bu özellikleri konuşma tanımada dağarcık patlamasına ve dilin istatistiklerinde karmaşıklığa sebep olmaktadır. Bu sorunların üstesinden gelebilmek için uygulamaya yönelik, radyoloji, dikte sistemi önerilmiştir. Küçük dağarcıklı, konuşmacı bağımsız, radyoloji konuşma tanıma sisteminde kelime tabanlı dil modeli kullanılarak yüzde 87.06'lık konuşma tanıma başarımına ulaşılmıştır. Buna rağmen, aynı sistem haberlerin diktesi için kullanıldığında, dağarcık dışı kelimelerin çokluğundan dolayı yüzde 46.29'luk tanıma başarımı vermiştir. Bu yüzden, birkısım sözcükler, kök, köksonrası ve morfemler gibi daha küçük tanıma birimlerine bölünmüş ve bu küçük birimler, bölünmemiş sözcüklerle birlikte, sözlük elemanları olarak konuşma tanıyıcıya tanıtılmıştır. Bu durumda, orta boyutlu bir dağarcıkla, dağarcık dışı kelime çokluğu sorunu halledilebilmiş, ve dilin istatistiksel modelleri için daha iyi kestirimler elde edilmiştir. Buna rağmen, en iyi tanıma başarımı kelime tabanlı dil modeliyledir.

TABLE OF CONTENTS

| | |
|---|------|
| ACKNOWLEDGEMENTS | iii |
| ABSTRACT | iv |
| ÖZET | v |
| LIST OF FIGURES | viii |
| LIST OF TABLES | ix |
| LIST OF SYMBOLS/ABBREVIATIONS | xii |
| 1. INTRODUCTION | 1 |
| 1.1. Large Vocabulary Continuous Speech Recognition | 1 |
| 1.2. Language Modeling | 2 |
| 1.3. Modeling Turkish | 2 |
| 2. SPEECH RECOGNITION | 4 |
| 2.1. Acoustic Modeling | 5 |
| 2.2. Statistical Language Modeling | 6 |
| 3. TURKISH RADIOLOGICAL DICTATION SYSTEM | 9 |
| 3.1. Recognizer Overview | 10 |
| 3.1.1. Acoustical Training | 10 |
| 3.1.2. Language Model Training | 11 |
| 3.1.3. Graphical User Interface | 12 |
| 3.2. Statistics of the Radiology Corpus | 14 |
| 3.2.1. Training and Test Data | 14 |
| 3.2.2. Statistics of the Training Corpus | 14 |
| 3.2.3. Statistics with Respect to Test Set | 15 |
| 3.3. Recognition Experiments | 16 |
| 4. TURKISH BROADCAST NEWS DICTATION SYSTEM | 17 |
| 4.1. Turkish Morphology and Morphological Parser | 18 |
| 4.1.1. Turkish Morphology | 18 |
| 4.1.2. The Morphological Parser | 20 |
| 4.2. Proposed Language Models | 22 |
| 4.2.1. Word-based Model | 22 |

| | | |
|----------|---|----|
| 4.2.2. | Combined Model | 23 |
| 4.3. | Statistics of the Corpus | 27 |
| 4.3.1. | Training Corpus | 27 |
| 4.3.2. | Test Data | 28 |
| 4.3.3. | Number of Distinct Tokens | 28 |
| 4.3.3.1. | Word-based Model | 29 |
| 4.3.3.2. | Combined Model with 2.5K Words | 29 |
| 4.3.3.3. | Combined Model with 5.0K Words | 30 |
| 4.3.3.4. | Comparison of Models | 31 |
| 4.3.4. | Coverage | 31 |
| 4.3.4.1. | Coverage with the Word-based Model: | 31 |
| 4.3.4.2. | Coverage with the Combined Model with 2.5K Words | 33 |
| 4.3.4.3. | Coverage with the Combined Model with 5.0K Words | 34 |
| 4.3.4.4. | Comparison of Models | 34 |
| 4.3.5. | Bigram Models | 35 |
| 4.3.5.1. | Bigram Analysis of the Word-based Model | 35 |
| 4.3.5.2. | Bigram Analysis of the Combined Model with 2.5K Words | 36 |
| 4.3.5.3. | Bigram Analysis of the Combined Model with 5.0K Words | 37 |
| 4.3.5.4. | Comparison of Models | 38 |
| 4.3.6. | Statistics with Respect to Test Set | 38 |
| 4.4. | Recognition Experiments | 39 |
| 4.4.1. | The Recognizer | 40 |
| 4.4.2. | Recognition Experiments with Word-based Model | 41 |
| 4.4.3. | Improvements to the Word-based Model | 45 |
| 4.4.4. | Recognition Experiments with Combined Model with 2.5K Words | 46 |
| 4.4.5. | Recognition Experiments with Combined Model with 5.0K Words | 48 |
| 4.4.6. | Comparison of Models | 50 |
| 5. | CONCLUSIONS | 51 |
| | REFERENCES | 53 |

LIST OF FIGURES

| | | |
|-------------|---|----|
| Figure 2.1. | Block diagram of subword unit base continuous speech recognizer [3] | 5 |
| Figure 3.1. | Graphical user interface of the radiological dictation system | 13 |
| Figure 3.2. | Advanced settings of the radiological dictation system | 14 |
| Figure 4.1. | Word-based model | 23 |
| Figure 4.2. | Basic idea behind the combined model | 25 |
| Figure 4.3. | Combined model | 26 |

LIST OF TABLES

| | | |
|------------|--|----|
| Table 3.1. | Statistics of the training corpus | 14 |
| Table 3.2. | Statistics of the test data | 15 |
| Table 3.3. | Recognition experiments with the radiology corpus | 16 |
| Table 4.1. | Percentage of different word orders in Turkish | 20 |
| Table 4.2. | Number of words and percentage of each domain in the training corpus | 27 |
| Table 4.3. | Different domains in the training data groups | 28 |
| Table 4.4. | Number of words and percentage of each domains in the test data | 28 |
| Table 4.5. | Number of tokens (words), number of distinct tokens and number of new distinct tokens in the word-based model | 29 |
| Table 4.6. | Number of words, number of tokens, number of distinct tokens and number of new distinct tokens in the combined model with the most frequent 2500 words | 30 |
| Table 4.7. | Number of words, number of tokens, number of distinct tokens and number of new distinct tokens in the combined model with the most frequent 5000 words | 30 |
| Table 4.8. | Coverage with respect to the word-based model | 32 |
| Table 4.9. | Coverage with respect to the combined model with 2.5K words . . . | 33 |

| | | |
|-------------|---|----|
| Table 4.10. | Coverage with respect to the combined model with 5.0K words . . . | 34 |
| Table 4.11. | Bigram analysis for the word-based model | 36 |
| Table 4.12. | Bigram analysis for the combined model with 2.5K words | 37 |
| Table 4.13. | Bigram analysis for the combined model with 5.0K words | 37 |
| Table 4.14. | Coverage analysis of the test set | 39 |
| Table 4.15. | Bigram analysis of the test set | 39 |
| Table 4.16. | Selection of the p and s parameters for the word-based model (correct/accuracy) | 41 |
| Table 4.17. | Test results for the i.'th and the ii.'th experiments in terms of percent of correct | 43 |
| Table 4.18. | Test results for the i.'th and the ii.'th experiments in terms of percent of accuracy | 43 |
| Table 4.19. | Test results for the iii.'th experiment | 44 |
| Table 4.20. | Test results for all the experiments | 44 |
| Table 4.21. | Test results for all the experiments for modified word-based model | 45 |
| Table 4.22. | Selection of the p and s parameters for the combined model with 2.5K words (correct/accuracy) | 46 |
| Table 4.23. | Test results for the i.'th and the ii.'th experiments in terms of percent of correct | 47 |

| | | |
|-------------|---|----|
| Table 4.24. | Test results for the i.'th and the ii.'th experiments in terms of per cent of accuracy | 47 |
| Table 4.25. | Test results for all the experiments | 48 |
| Table 4.26. | Selection of the p and s parameters for the combined model with 5.0K words (correct/accuracy) | 48 |
| Table 4.27. | Test results for the i.'th and the ii.'th experiments in term of per cent of correct | 49 |
| Table 4.28. | Test results for the i.'th and the ii.'th experiments in terms of per cent of accuracy | 49 |
| Table 4.29. | Test results for all the experiments | 50 |
| Table 4.30. | Comparison of all the proposed models | 50 |
| Table 5.1. | Summary of the proposed models in terms of the defined comparison statistics | 51 |

LIST OF SYMBOLS/ABBREVIATIONS

| | |
|-------------------------|--|
| A | Acoustic signal |
| $A_n \ n = 1, \dots, n$ | Acoustic feature vectors |
| C | Counting function |
| W | Word sequence |
| \hat{W} | Word sequence selected by the recognizer |
| $w_n \ n = 1, \dots, n$ | Words in W |
| X | Random variable |
| χ | Ranges of the random variable X |
| HMM | Hidden Markov Model |
| LVCSR | Large Vocabulary Continuous Speech Recognition |
| OOV | Out of Vocabulary |
| SLM | Statistical Language Modeling |
| WER | Word Error Rate |

1. INTRODUCTION

1.1. Large Vocabulary Continuous Speech Recognition

The aim of speech recognition is to understand the human speech by machines and then machines perform the task based on this understanding. The main parts of speech recognition are the acoustic modeling where models are derived from the speech feature vectors, lexicon which contains the phonemic representations of words, language modeling that characterizes the statistical regularities of the language and the decoding part which decides the best word sequence. Finally, the recognizer transcribes the acoustic signal to symbols.

The important decisions for the speech recognition systems are the size of the vocabulary and the selection of the base recognition units. The vocabulary size depends on the speech recognition application. For large vocabulary continuous speech recognition (LVCSR) system, the vocabulary size increases to thousands of words, however a credit card verification system needs only the 10 digits as the vocabulary entries. If words are recognized in isolation as in the digit recognition case, all the words in the lexicon have equal chance to follow each other but in continuous speech recognition or a dictation task, the regularities of the language get great importance. In that case, the chance of all the words to follow each other is determined by the language itself which is the main idea behind language modeling.

Second decision criterion is the selection of the base recognition units. There is a high tendency to select the words as units. However, the selection decision has to be changed according to the characteristics of the language. For English, it is a good choice, however for agglutinative languages, words as recognition units will be failed due to the productive morphology of the language. The criterion for appropriate base recognition units is that, the units have to be longer enough in terms of acoustic information to make a reliable decision. Also the units will be able to cover the language with the moderate vocabulary size. The selection of the recognition units for LVCSR

applications will be main issue of this thesis.

1.2. Language Modeling

Language Modeling is an essential part of the speech recognition. The aim of language modeling is to capture the regularities of the natural language to improve the recognition performance. It estimates the distribution of the language units using the training data which is the text. Language model helps to the recognizer in the determination of the best words sequence. Due to the characteristics of the language, some words have a high probability to follow each other, and some words have no change to occur in the same context. Therefore, language modeling gives the recognizer an idea about the next recognized word, and decreases the number of the candidates. It is impossible to estimate the distribution all the consecutive word sequences, however this approach can be feasible by estimating the probability distribution for only the N consecutive words, which is called the N -gram language models. For small vocabulary speech recognition applications, estimation of the distribution of words is a possible approach using a moderate size text corpus. However, for large vocabulary applications, it becomes impossible to estimate the distribution of the language accurately. IBM shows that several hundred million words are needed to saturate the N -gram language models where N is equal to two and a few billion words for N is equal to three [1].

1.3. Modeling Turkish

Turkish is a challenging language for LVCSR applications. There are two reasons for this. One of them is the language characteristics of Turkish because Turkish characterizes an agglutinative nature with lots of inflectional and derivational suffixes. The other one is the lack of resources like speech databases, text corpora and pronunciation dictionary for Turkish.

Most of the research in LVCSR is done in English, and for languages like English, many LVCSR engines have been evaluated. However, the morphological structure of Turkish is completely different than English. Continuous speech recognition and a dic-

tation system need a huge vocabulary size. However, the inflectional nature of Turkish increases the vocabulary size drastically if words are selected as base recognition units. The words that are introduced to the recognizer but are not found in the vocabulary are called the “out-of-vocabulary (OOV)” words and due to the agglutinative nature of Turkish the OOV words are very large. As a consequence applying the same methods to Turkish will give poor recognition results because of the large number of OOV words. Therefore, new methods for languages like Turkish have to be investigated.

This thesis is an attempt to solve the problem of vocabulary explosion due to the agglutinative nature of the language. If words are selected as base recognition units, the vocabulary size will be very large to cover most of the words in the language. Also in the vocabulary, there will be lots of words that derive from a single stem. Therefore, we try to solve this problem by using the morphological properties of Turkish. The aim will be to decrease the number of OOV words with a decrease on the vocabulary size for the LVCSR engine of Turkish. To achieve this goal, our base concern will be to use the combinations of previously proposed language modeling units.

2. SPEECH RECOGNITION

Dictation is simply the process of converting acoustic speech signals into written form. The first step in designing a dictation system is building a speech recognizer. Therefore, in this chapter firstly the fundamentals of the speech recognition will be explained. The components of the speech recognizer are shown in the Figure 2.1.

In the first step of the recognition, feature vector of the input speech is derived using the spectral analysis. In subword models part, subword (like phones, syllable) Hidden Markov Models (HMMs) are generated, then using the lexicon, consist of the transcriptions of the words, word models, HMM's are created. Language Models are generated using the training text input and its role in the recognition network is to control the best phoneme or word sequence and improve the recognition performance. Finally decoder finds the optimum word sequence.

If we consider the mathematical formulation of the speech recognition, for a given acoustic observation vector $A = A_1A_2...A_n$, the aim of speech recognition is to find out the corresponding word sequence $\hat{W} = w_1w_2...w_n$, that has the maximum posterior probability $P(W|A)$, expressed as [2]:

$$\hat{W} = \arg \max_w P(W|A) \quad (2.1)$$

If we apply Bayes' rule to (2.1):

$$\hat{W} = \arg \max \frac{P(A|W)P(W)}{P(A)} \quad (2.2)$$

In (2.2), acoustic observation A is fixed, so maximizing this equation is equal to maximizing the following equation:

$$\hat{W} = \arg \max P(A|W)P(W) \quad (2.3)$$

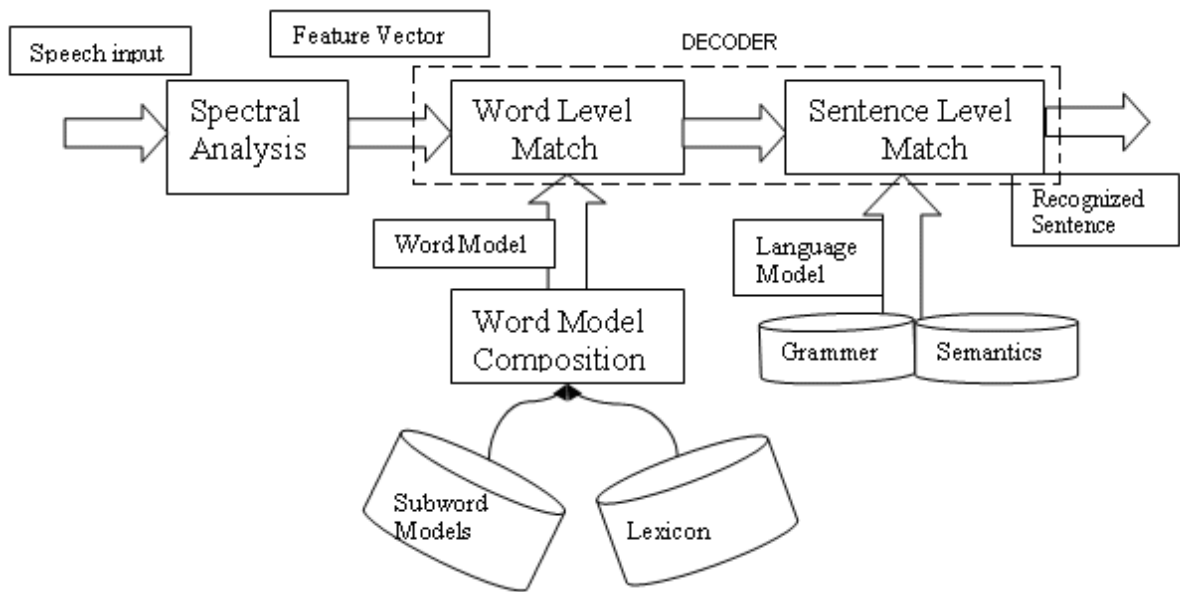


Figure 2.1. Block diagram of subword unit base continuous speech recognizer [3]

where $P(A|W)$ is related to acoustic modeling and $P(W)$ is calculated using language modeling. Building these models accurately is very important for the performance of the speech recognizer.

2.1. Acoustic Modeling

In (2.3), $P(A|W)$ is the acoustic modeling part. It is the probability of the acoustic observation vector A when the speaker intend to utter the word sequence W . In Large Vocabulary Speech Recognition systems, there are lots of words, therefore, instead of modeling each word, they are decomposed into smaller subword units like phones, triphones, syllables and then acoustic models for these subword units are generated. Word models are produced by concatenation of these subword units. Mostly, Hidden Markov Models [4] are used for acoustic modeling.

In continuous speech recognition, for word model generation, instead of the concatenation of individual phone models, concatenations of triphones are used. Triphone model takes into consideration of the neighboring left and right phones. If a phone has different neighboring phones in two different contexts, they have been considered as different triphones. This context dependency can improve the system performance,

if the parameters of the triphones are estimated using a large amount of training data [2].

2.2. Statistical Language Modeling

The aim of Statistical Language Modeling (SLM) is to estimate the distribution of the natural language for the purpose of speech recognition and other language technologies [1]. SLM estimates the necessary linguistic information using training data that is text. So estimation critically depends on the availability of the large amounts of the training data.

If we consider the mathematical formulation of the statistical language modeling, in (2.3), $P(W)$ refers to the language modeling part of the speech recognition. It is the probability of the word string W and formulated as

$$P(W) = P(w_1, w_2, \dots, w_n) \quad (2.4)$$

Using the Chain Rule $P(W)$ can be decomposed as [5]

$$P(W) = \prod_{i=1}^n P(w_i | w_1, \dots, w_{i-1}) \quad (2.5)$$

where $P(w_i | w_1, \dots, w_{i-1})$ is the probability that w_i will be spoken given the words w_1, \dots, w_{i-1} , which were previously spoken. It is called the history.

It is impossible to estimate the probabilities $P(w_i | w_1, \dots, w_{i-1})$ for large vocabulary sizes. Since some histories are unique or repeating only a few times. A practical solution to this problem is to assume that the history equals to the several previous words. If the history equals to $N - 1$ previous words, we have N -gram language model and $P(W)$ is calculated as,

$$P(W) = \prod_{i=1}^n P(w_i | w_{i-N+1}, \dots, w_{i-1}) \quad (2.6)$$

We have the *trigram* model: $P(w_i|w_{i-2}, w_{i-1})$, if w_i depends on the previous two words. If w_i depends on the previous word, then we have the *bigram* model: $P(w_i|w_{i-1})$. Finally if we make the assumption that all the words are uttered independently, then we get the *unigram* model: $P(w_i)$.

The statistics of the N-gram models are calculated using a training text corpus. The domain and the size of the training corpus has importance in estimating Language Model probabilities. The calculation of N-gram probabilities is simply a counting and a normalization process. The occurrences of a particular N-gram is counted from a text corpus and it is normalized with the occurrences of all of the N-grams, sharing the same N-1 previous words. The formulation of the N-gram probabilities is as follows:

$$P(w_i|w_{i-N+1}, \dots, w_{i-1}) = \frac{C(w_{i-N+1}, \dots, w_{i-1}, w_i)}{\sum_w C(w_{i-N+1}, \dots, w_{i-1}, w)} \quad (2.7)$$

In (2.7), $C(w_{i-N+1}, \dots, w_{i-1}, w_i)$ is the count, showing how many times the N-gram $w_{i-N+1}, \dots, w_{i-1}, w_i$ occurs in the training corpus.

One of the problems in N-gram language modeling is the data sparseness. If the training corpus is not large enough, then extremely small probabilities can be assigned to many possible word sequences. In that case, N-gram smoothing should be applied. In N-gram smoothing the extremely small probabilities like zero probability are increased and high probabilities are decreased to make the probability distribution of the model flatter. Therefore, smoothing techniques produce more robust probabilities for unseen data, in spite of the fact that the likelihood for the seen data may be hurt slightly [2].

The quality of a language model is evaluated mostly over the test set, new data. If speech recognition is involved, word error rate (WER) is the quality metric of the language model. However, if speech recognition is not participated, then the quality measure is the test-set perplexity which is calculated from entropy.

Entropy is a measure of information describing the uncertainty about an event.

If X is a random variable which ranges over χ and with particular probability function $p(x)$, the entropy of this random variable is defined as

$$H(X) = - \sum_{x \in \chi} p(x) \log p(x) \quad (2.8)$$

If the probability distribution of p does not known actually, it is the case in language modeling, and then cross entropy becomes useful. It allows us to use some p_m , which is a model of p , [1].

$$\text{cross-entropy}(p, p_m) = - \sum_D p(D) \log p_m(D) \quad (2.9)$$

where $D = (D_1, D_2, \dots, D_n)$ is the new data sample. Finally, average number of choices that a random variable has to make, it is the perplexity, is calculated as,

$$\text{perplexity}(p, p_m) = 2^{\text{cross-entropy}(p, p_m)} \quad (2.10)$$

From the speech recognition framework, perplexity is the average branching factor. If the perplexity is high, this means that after the recognizer recognizes a word, the recognizer will consider a large number of choices as the next word, because lots of words can follow the previously recognized word. Therefore, lower perplexity means the better language model. However, lower perplexity does not result in lower word error rates every time.

3. TURKISH RADIOLOGICAL DICTATION SYSTEM

Dictation is one of the most challenging areas in automatic speech recognition. There is a large demand for speech-to-text systems because speaking is faster than typing in most of the languages. However, today most dictation systems do not perform at desired recognition rates, since the vocabulary size can be huge for any given language. In order to alleviate this problem, task-specific dictation systems are proposed in many areas. One common example is dictation for radiologists who are often eyes and hands-busy at work. In Turkey, in most of the hospitals, radiologist perform their task by recording the diagnosis about the X-ray photograph or the MRI of the patient and then a secretary converts these recordings into written form. Therefore using a dictation system can make the life easier from the point of the radiologist.

The aim of this chapter is to build a state-of-art LVCSR system for radiological dictation. The reasons that radiology area is selected for this dictation system are as follows:

- i. Using keyboards for input entry is not appropriate for hands-busy and eyes-busy applications, and radiologist is a specific example of this type.
- ii. Huge vocabulary sizes degrades the recognition performance, however, in radiology area vocabulary size is small as radiology science has its own specified vocabulary.
- iii. Although, Turkish is a difficult language for speech recognition applications, because of its agglutinative nature and free word order, the systematic arrangement of words in sentence formation, make the radiology area suitable for the dictation applications.

Therefore, in Turkish Radiological dictation system, the vocabulary size can be reduced to only several thousand words, and the perplexity, average branching factor from a word, can be very small.

There is a previous research effort [6] on the design of a dictation system for Radiology area. This research considers only the acoustical modeling and higher recognition performance is achieved on their test set. However, our main goal in designing the radiological dictation system is to see the performance of a LVCSR system over a test set with a high coverage, small perplexity and using words as base recognition units. This is the counterexample of general Turkish. It has been shown in [7] that because of the agglutinative characteristics of Turkish, LVCSR efforts on Turkish Language shows higher perplexities and smaller coverage over the test set if words are selected as base recognition units.

In this chapter, firstly an overview of our recognizer is given. Then, the statistics of the radiology text corpus and the recognition results will be explained.

3.1. Recognizer Overview

In this thesis, Hidden Markov Model Toolkit (HTK) [8] is used for the design of the speech recognizer. HTK is a toolkit for building HMM's. Although, it is designed for general purpose, its primary usage is in particular in speech recognition.

3.1.1. Acoustical Training

The first step in recognizer development is data preparation. Large amounts of training data is needed for better models. We need some amounts of labeled data for initial model generation, and then some unlabeled data for building better models.

We used the labeled recordings of 10 people each of them uttering 149 different phonetically balanced words and sentences for the initial estimates of our monophone models. We generate 29 monophone models with three states and six mixtures. Then we generate our own training database recordings for radiological dictation system. We select 95 sentences covering the most frequent triphones from our radiology reports. Speech data from these sentences are recorded from 16 different people. These unlabeled data are labeled using the initial monophone estimates by force alignment, and

these recordings are used for the final estimates of our monophone models. Also a three state silence model, for modeling the beginning and the ending of the utterances, and a one state short pause model, for representing the pauses between words, are generated.

The next step is making context-dependent triphones from monophone models and re-estimating them. This step is also generated using the HMM training tools of HTK. By this way we generate 1680 triphone models and we decrease this number to 1650 by applying data driven clustering. Also, triphones that will be used for word generation in radiological report entry, but is not available in the acoustical training data are mapped to these 1650 physical models using the acoustical similarity criterion of phonemes.

Finally, we have 1650 physical models that will be used in radiological reporting via speech.

3.1.2. Language Model Training

The language modeling library of HTK can support general N-grams. However, constructing and using N-grams are limited to bigram. Therefore, we will use bigram in our recognition experiments. Bigram probabilities with back-off smoothing are calculated using HTK. The back-off bigrams are given by [8]

$$p(i, j) = \begin{cases} (N(i, j) - D)/N(i) & \text{if } N(i, j) > t \\ b(i)p(j) & \text{otherwise} \end{cases} \quad (3.1)$$

In (3.1), $N(i, j)$ is the number of times word j follows word i , $N(i)$ is the total number of word i in the training text, D is the discount constant, t is the bigram count threshold, and $b(i)$ is the back-off weight which is calculated as:

$$b(i) = \frac{1 - \sum_{j \in B} p(i, j)}{1 - \sum_{j \in B} p(j)} \quad (3.2)$$

where, B is the set of all words for which $p(i, j)$ has a bigram.

Then we generated our recognition network with back-off bigram language model probabilities. Also we generated sublattice networks for the month, date and digits which will be appended to the radiology words network. We assign small back-off bigram probabilities to these networks.

One important point in dictation is to make the transcript document as close as possible to the original one. Therefore the punctuation marks are very important. However, while recording the radiological reports some of the doctors preferred to utter all of the punctuation marks and some doctors do not utter all of them. Also the collected report from the hospital is very limited. Therefore, we train the language models using the texts of the same reports with all the punctuation marks and without the punctuation marks.

3.1.3. Graphical User Interface

We designed a graphical user interface for our radiological dictation system. The screen shot of our system is shown in Figure 3.1. The parts of our interface are as follows:

- *Patient Number, Name-Surname*: This part is an entry from the user, the patient number and the name-surname will be shown on the left top of the output report.
- *Record button*: Record button is used for recording voices for later use in off-line recognition. It records at 16 kHz, 16 bits wav format.
- *Run Recognizer Live*: This part is designed for recognition from direct audio input. However, this part is not working as real time effectively due to the recognition properties of HTK.
- *Recognize from a file*: This part is designed for off-line recognition. It gives better recognition results than online recognition.
- *Start-Stop button*: This buttons are used to start and stop the recognition.
- *Advanced button*: This button is a part to arrange the advance settings of the recognizer. It is used to arrange the grammar scale factor, word insertion penalty and pruning threshold parameters. This part is shown in Figure 3.2.

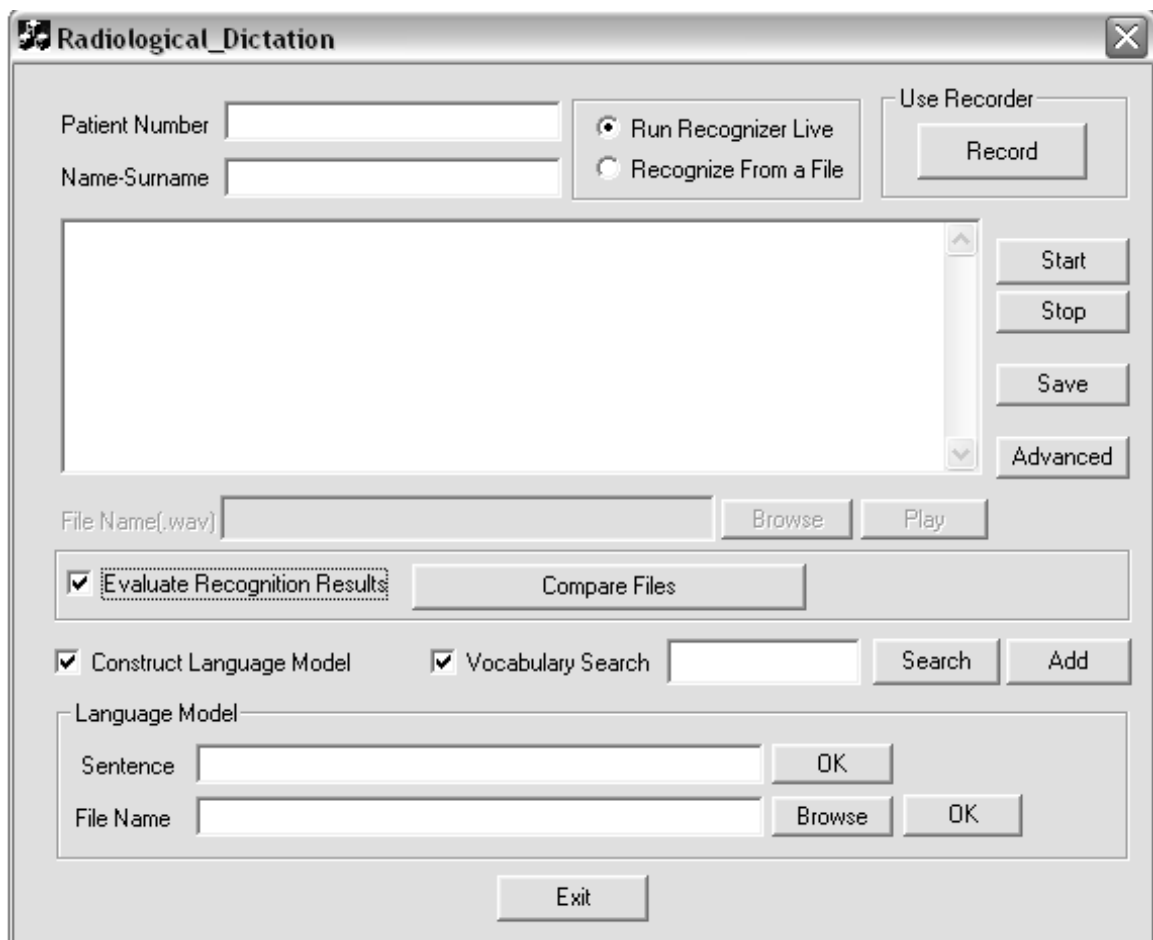


Figure 3.1. Graphical user interface of the radiological dictation system

- *File name*: This part is used to select the wav file for recognition and to listen.
- *Evaluate Recognition Results*: This part is designed for the evaluation of the recognition results. Original and the recognized files are compared with each other and the recognition statistics are obtained.
- *Construct Language Model*: In this part a sentence or a .txt file can be directly appended to the available language model.
- *Vocabulary Search*: An unrecognized word can be checked if it is in the vocabulary or not, using this button. Also this new word can be added directly to the vocabulary.

Our final output is a radiology report saved as a text document. A radiologist can listen to the recorded wav file with the play button and can edit the recognized output file at that time.

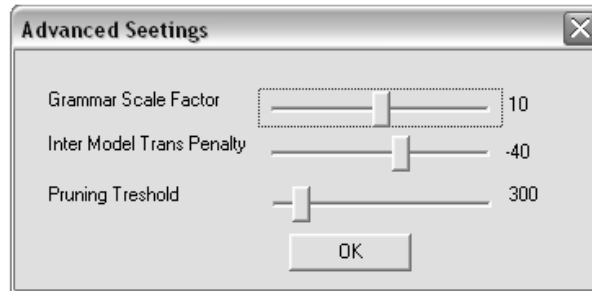


Figure 3.2. Advanced settings of the radiological dictation system

3.2. Statistics of the Radiology Corpus

3.2.1. Training and Test Data

The radiological reports needed for the training of the speech recognition system are collected from Hacettepe University Radiology Department. All of them are ultrasonography reports belonging to 28 different areas. We have collected 507 radiological ultrasonography reports. There are 437 reports belonging to 28 different radiological domains in the training corpus and there are 60 reports belonging to 17 different radiological domains in the test data.

3.2.2. Statistics of the Training Corpus

In this part Training Corpus is evaluated according to the number of distinct tokens. The detailed description of the number of distinct tokens will be explained later in the analysis of the Broadcast news corpus. In radiology LVCSR system, the tokens will be the words in the training corpus and the distinct tokens will be the number of the distinct words.

Table 3.1. Statistics of the training corpus

| | Number of Tokens (words) | Number of Distinct Tokens (distinct words) |
|-----------------|-----------------------------|---|
| Training Corpus | 91469 | 1562 |

The analysis of the radiology corpus is given in Table 3.1. It is clear from the table that, the number of distinct words to cover all the words (91469 words) in the training corpus is very small. Therefore all of these distinct words are added to the vocabulary of the radiological dictation system.

3.2.3. Statistics with Respect to Test Set

In this part coverage and perplexity analysis over the test set will be given with the available training data. By coverage over the test set, we mean the per cent of the words which are both in the vocabulary and the test data. The recognizer has no chance to recognize the words that are not in the vocabulary. Therefore, high coverage over the test set means a high theoretical recognition performance. By perplexity analysis over the test set, we mean the bigram perplexity of the test set calculated over the training text corpus. Small perplexity means a better language model because from the point of the recognizer the number of choices for a new word decreases. However small perplexity does not grantee higher recognition performance.

As mentioned before, from totally collected 507 radiology reports 60 of them are selected for test and 437 of them for training. The analysis of the test data in term of coverage and perplexity is given in Table 3.2.

Table 3.2. Statistics of the test data

| | Coverage (%) | Perplexity |
|-----------------|--------------|------------|
| Training Corpus | 95.54 | 13.62 |

As shown from the table, the coverage is very high and the perplexity is very small especially compared to the general Turkish. This is the indication of the specific and limited vocabulary of the radiology area.

3.3. Recognition Experiments

We perform recognition experiments with the recordings of the test data. The test data are recorded with the utterances of six female and four male speakers, only two of them are doctors. The pronunciation of radiological terminology is not easy for an unfamiliar speaker. Therefore, the pronunciation variability between speakers are very high. Also the recordings are taken in two different ways, as reading reports slowly (Recordings-1) and very fast (Recordings-2). The recognition results are explained in Table 3.3.

Table 3.3. Recognition experiments with the radiology corpus

| | Correct (%) | Accuracy (%) |
|--------------|-------------|--------------|
| Recordings-1 | 87.06 | 84.35 |
| Recordings-2 | 82.47 | 80.79 |

In Recordings-1, where reports are uttered in a slow manner, the recognition performance is better. However, in most of the radiological reports, same words are written differently, considered as different tokens. There are lots of these kinds of words and no preprocessing is applied to write all of them in the same manner. In the evaluation of the results, this will cause substitution errors. Therefore, real recognition performance of the dictation system is better than the given one.

4. TURKISH BROADCAST NEWS DICTATION SYSTEM

As we mentioned before the aim of this thesis is to make a state-of-art LVCSR system. In chapter 3, it has been shown that, in a specific area like radiology, a dictation system can perform well if words are selected as base recognition units. The reason is that, the OOV rate and perplexity is very small compared to the general Turkish because in radiology, there is a limited vocabulary with regular word formations. Also from the point of the recognizer, large recognition units, words, contain enough acoustic information to make a reliable decision. Therefore small error rates are achieved.

However, if we consider the general Turkish, the selection of base units will be a crucial problem. If words are selected as base units, the OOV rate will be higher because of the agglutinative nature of Turkish. By suffixation, millions of word forms can be derived from a single Turkish root. Therefore the selection of words as base units will increase the vocabulary size drastically, although words contain more acoustic information compared to the smaller recognition units like phonemes, syllables, morphemes, etc...

In this chapter, we will try to solve this trade-off using the morphological properties of Turkish. Especially, we will concentrate on the transcription of broadcast news. In [7], different base recognition units like words, stems and endings, roots and morphemes are compared with each other. It has been found that, the best solution is to parse the words as stems and endings, and then to select these parts as base recognition units. In our thesis, we proposed a new model with the combination of all the proposed models in [7]. In the next parts, firstly, we will give a brief information about the morphology of Turkish, and the morphological parser that we have been used. Then our proposed model will be compared with the word based model in terms number of tokens, coverage, perplexity and recognition performance.

4.1. Turkish Morphology and Morphological Parser

4.1.1. Turkish Morphology

Morphology is the study of the internal structure of words and the rules by which words are formed [9]. Turkish is an agglutinative language which means from the same root, very high number of words can be formed by suffixation [10].

The suffixes of Turkish are categorized as derivational or inflectional in terms of their function. Derivation is used to produce new lexical items, and it may change the grammatical category. Some examples are [11]:

büz+gü (noun derived from verb stem)

kat+la (verb derived from noun stem)

tuz+luk (noun derived from noun stem)

kan+dır (verb derived from verb stem)

However nominal inflection only marks the grammatical notions like number, person, gender, and verbal inflection marks tense, aspect, modality and person. The morphotactics for verbal inflection is more complex than the nominal inflection. The examples of nominal and verbal inflection is given in the below cases.

nominal inflection:

ev+im+de+ki+ler+den (one of those that were in my house)

verbal inflection:

yap+tır+ma+yabil+iyor+du+k (It was possible that we did not make someone do it)

A popular example of word formation showing the complex morphotactics of Turkish is [12]:

“OSMANLILAŞTIRAMAYABİLECEKLERİMİZDENMİŞSİNİZCESİNE”

which can be decomposed into morphemes as:

OSMAN+LI+LAŞ+TİR+AMA+YABİL+ECEK+LER+İMİZ+DEN+MIŞ+SİNİZ
+CESİNE

Its meaning is;

“as if you were of those whom we might consider not converting into an Ottoman”

Although there is not a one-to-one correspondence between Turkish morphemes and English words, we can clearly say that a single Turkish word can correspond to a group of English words. This example is the illustration of the drawback that we have to encounter if we apply the same methods used in English speech recognition engines directly to Turkish.

During the suffixation process, the first vowel of the morpheme must be compatible with the last vowel of the stem which is the rule of vowel harmony. According to vowel harmony, stem ending with back/front vowel takes a suffix starting with back/front vowel. There are some exceptions, especially for the foreign words which are enrolled to Turkish lexicon. Morphological parser overcomes this problem by adjusting the last vowel of these stems according to the suffix they take.

Other characteristic of Turkish is the free word order. This is a challenging nature from the perspective of the speech recognizer, because it increases the average branching factor from a word which leads to an increase in the perplexity of the language. Although Turkish characterizes a *subject-object-verb* (SOV) type language, the order of constituents can be changed without effecting the grammar of the sentence. The effect is only to emphasize the meaning. The word which will be emphasized in the sentence is generally placed just before the verb, some examples are [13]:

Ben çocuğa **kitabı** verdim (I gave the book to the children)

Çocuğa kitabı **ben** verdim (It was me who gave the child the book)

Ben kitabı **çocuğa** verdim (It was the child to whom I gave the book)

Table 4.1 [14] shows the commonness of other constitutes order. As shown from the table, there is a high tendency for SOV type in both the adult and the children speech. However, none of them prefer to utter the sentences in VOS type.

Table 4.1. Percentage of different word orders in Turkish

| Sentence Type | Children Speech (%) | Adult Speech (%) |
|---------------|---------------------|------------------|
| SOV | 46 | 48 |
| OSV | 7 | 8 |
| SVO | 17 | 25 |
| OVS | 20 | 13 |
| VSO | 10 | 6 |
| VOS | 0 | 0 |

4.1.2. The Morphological Parser

The morphological parser used in this thesis is firstly developed at Boğaziçi University Computer Engineering Department by Çetinoğlu [15] and then modified by Dutagacı [7]. In the parser, the stems and the suffixes with their properties are defined; also the transitions between the morphemes are defined by the grammatical suffixation rules.

There are 29540 nominal and verbal stems defined in the modified parser. In addition, 5963 new stems are added to the parser from the broadcast news corpus. Also there are lots of foreign words which take Turkish suffixes in the corpus. Therefore 462 foreign words are added to the parser as the nominal stems. The reason is to decrease the number of indecomposable words, especially for the most frequent words.

The properties of the modified parser can be summarized as [7]:

- Each stem is added to the parser with a line describing the below properties
Character sequence of the stem

Last letter of the stem

Last vowel of the stem

Category of the stem (verbal or nominal)

Type describing whether the surface form reflects an inflection or not

- If after suffixation, there is a change in the stem like deletion or modification of the last letter and insertion of a new letter, then the surface realization of the stem is added as a separate item. Some examples are “kitap” and “kitab+ı”, “fikir” and “fkr+i”, “his” and “hiss+i”. The stems “kitap” and “kitab”, “fikir” and “fkr”, “his” and “hiss” are added to the parser as separate items.
- Because of the vowel harmony, last vowel of the stem determines the surface form of the morphemes added to the stem. However, there are some exceptions. For example, if the word “ampul” takes a plural suffix it becomes “ampuller” instead of “ampullar”. By examining the indecomposable words, we add the last vowel of these stems according to the suffixes they take.

One of the problems that we have to encounter is the morphological disambiguation which is the problem of finding the suitable morphological parses given a sequence of words because a word can be decomposed into its morphemes in different ways. In [10], this problem is tried to be handled using the trigram model of inflectional groups derived from the words and the accuracy of ambiguity is increased to 95.07 per cent.

Extensive usage of suffixes causes ambiguities however; these ambiguities can sometimes be resolved at phrase level. An example is the word “çocukları”, which can be parsed as [13]:

child+PLU+3SG-POSS (his children)

child+3PL-POSS (their child)

child+PLU+3PL-POSS (their children)

child+PLU+ACC (children-accusative)

Also a word that has different parts-of-speech causes morphological disambiguation. For example the word “giderim”, which can be parsed as [13]:

N(gider)+ISG-POSS (my expense)

V(git)+AOR+ISG (I go)

During the parsing process, we do not deal with the morphological ambiguity problem directly. In the parser, we sort all the stems according to their character length. Therefore, selecting the first option as the parser output gives us the chance of selecting the one with the longer stem. The reason of this is that stems will be the recognition units and from the recognizer sight, longer units are better because they contain more acoustic information.

4.2. Proposed Language Models

In this part, we will be concentrated on the selection of base recognition units for speech recognition. The mostly used model is taking words as language modeling units especially for English recognition engines. However, if this model is used in modeling agglutinative languages like Turkish, Czech, Hungarian, Finnish and Korean the OOV rate will be very high [16][17], because it is impossible that the lexicon will contain all the words. Our proposed model will be the combination of all the models given in [7], word-based model, morpheme-based model and stem-ending-based model. Word-based model and the combined model will be applied on the data collected from the broadcast news.

4.2.1. Word-based Model

In word-based model, words are selected as lexicon entries for speech recognition and language modeling probabilities are extracted from the training corpus using the words as units. The system for this model is illustrated in Figure 4.1. Here, Z is the model used to represent the short pauses between the consecutive words. The transition probabilities between the Z model and the word model, also between word models are calculated using bigram language models.

This is the model used in the recognition engine of the radiological dictation

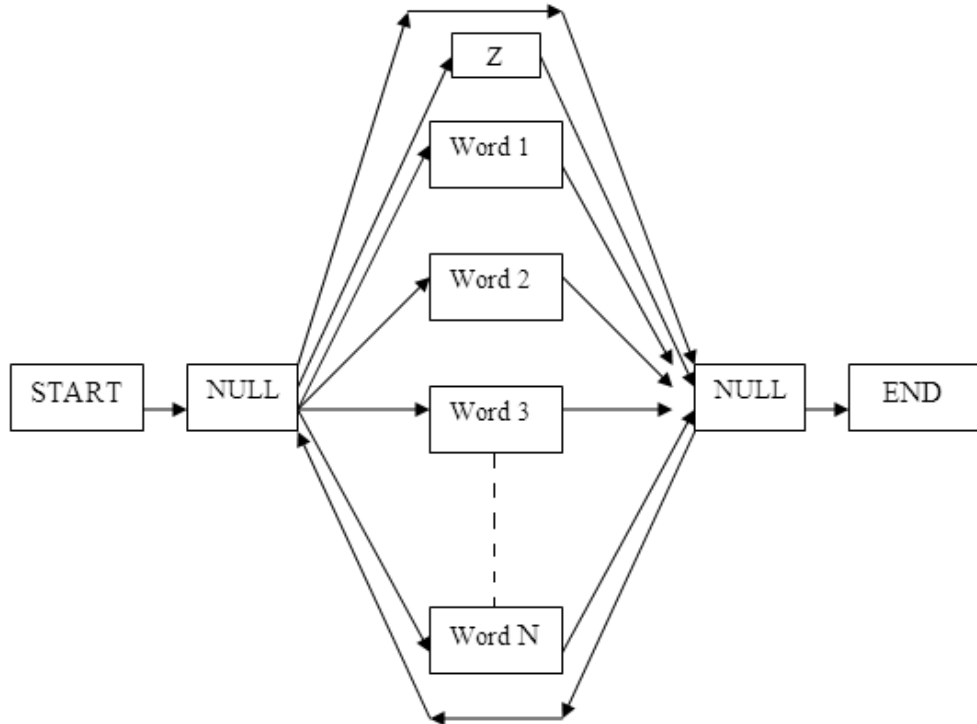


Figure 4.1. Word-based model

system. Using bigram language models, 87.06 per cent recognition rate is achieved with 4.46 per cent OOV rate over the test set. Although words are selected as recognition units, the reason of these results is the specific and limited vocabulary of the radiological terminology and the regularity in sentence formation. However, we expect smaller recognition rate with higher OOV rate, if we used this model in broadcast news dictation. The detailed results will be given in Sections 4.3 and 4.4.

4.2.2. Combined Model

Combined Model is the combination of all the models proposed in [7]. The previously proposed models can be summarized as:

- **Word-based Model:**

This is the same model explained in Section 4.2.1. Words are the lexicon entries and the base recognition units for this model. Although, words are larger recognition units and this is the desired situation from the point of the recognizer, the

agglutinative nature of Turkish, makes it impossible to add all the words to the vocabulary. Therefore with a specified vocabulary size the coverage is small. Also with the addition of new data the vocabulary growth becomes very high because it is possible to derive millions of new words from a single stem by suffixation.

- **Morpheme-based Word Model:**

The model where, all the words are decomposed to their stems and morphemes and then these are parts are taken as lexicon entries. In this model stems and morphemes are utilized as base recognition units. Due to the agglutinative characteristics of Turkish, a stem can be followed with lots of suffixes. By different combinations of stems and morphemes which are determined by the morphotactics of Turkish, lots of Turkish words can be derived. Therefore, the aim of this model is to reduce the vocabulary size and increase the coverage with the specified vocabulary by using stems and morphemes as vocabulary entries.

One of the drawbacks of the morpheme-based word model is that most of the morphemes are smaller recognition units compared to the words. Therefore, these units are lack of enough acoustic information for reliable speech recognition. For this reason, during the recognition experiments it has been found that, although using smaller vocabulary sizes high coverage is achieved, the recognition performance is poorer than the classical word-based model.

- **Stem-ending-based Word Model:**

The model where, all the words are decomposed to their stems and endings and then these are parts are taken as lexicon entries. The concatenation of morphemes which follows a stem is named as the ending. In this model stems and endings are utilized as base recognition units. Here, the proposed idea behind this model is to get rid of the problem of smaller recognition units by concatenating the morphemes. This idea for agglutinative languages is firstly proposed in [18], and this approach is applied to Turkish in [19]. If stems and endings are used as the vocabulary entries, to achieve the same coverage, vocabulary size increases compared to the previous model. However, recognition performance is better than the morpheme based model because of the larger recognition units. Therefore, this model becomes a solution for the the trade off between the small coverage in word-based model and lack of acoustic information in morpheme based model.

The combination of all these models will be our proposed combined model. The idea of the combined model is illustrated in Figure 4.2.

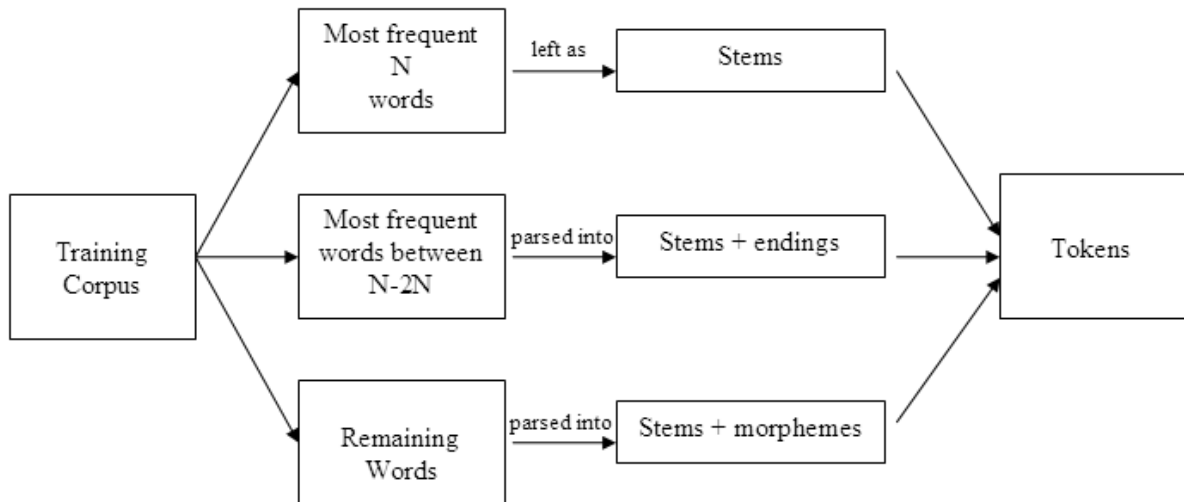


Figure 4.2. Basic idea behind the combined model

As shown from the figure; the tokens, the base language modeling units, are both the words which are left as stems, stems, endings and the morphemes available in the training corpus.

For the parsing process, the modified morphological parser is used. Firstly, all the words in the training corpus is parsed according to the morphotactics of Turkish. Then a post-processing is applied for the three different word groups, most frequent N words, most frequent N-2N words and the remaining words. The idea of the combined model is as follows:

- All the words in the training corpus are sorted according to the frequency of occurrences.
- Most frequent N words are left as stems.

If a word in this group takes no suffixation, then no parsing is applied. However, if it takes suffixes then it is parsed as “word+ending”. For example, if the word “çocukların” is in this group, instead of parsing it as “çocuk+lar+ın”, the whole word is left as stem. If the word “çocuklarından” is in the corpus, it is parsed as “çocukların+dan” where “çocukların” is the word and the “dan” is the ending

part, although the stem of this word is “çocuk”.

- Most frequent N to $2N$ words are parsed into stems and endings.

If a word is in this group, then a post processing is applied and all the morphemes of the word is concatenated to generate the ending part.

- Remaining words are parsed into stems and morphemes.

No post processing is applied to the output of the morphological parser.

- All of the words which are left as stems, all the parsed stems, endings and morphemes are the tokens of the model which will be the lexicon entries.

The system for this model is shown in Figure 4.3.

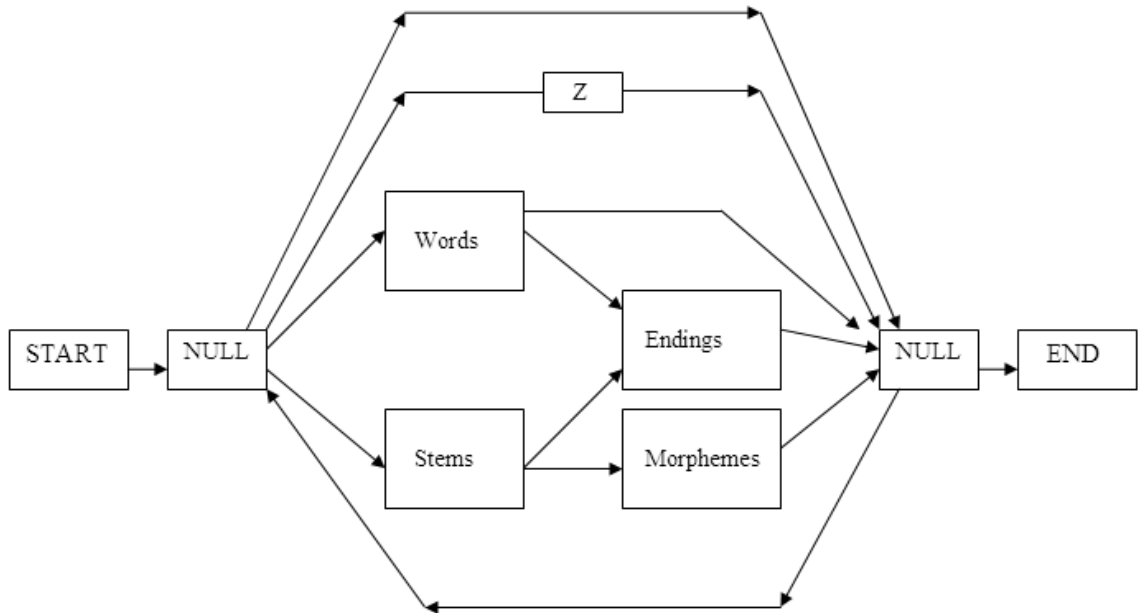


Figure 4.3. Combined model

In this figure, Z is the model used to represent short pauses between the consecutive words and $NULL$ nodes are used to decrease the number of arcs in the representation. As shown from the figure, there will be transitions from words to endings or to other words or stems. The reason is that words are left as stems so they can take suffixes or can be single stems. Also there are transitions between the stems to endings or morphemes. The morphemes part has its own lattice and the lattice is generated using the morphotactics of the Turkish with the available morphemes in the training corpus. Also the transition probabilities between words to endings, stems to endings

or morphemes, also within the NULL nodes are calculated from the training corpus.

In our proposed model, firstly we select N as 2500. The most frequent 2500 words are left as stems; the most frequent words between 2500 and 5000 are decomposed into stem and endings, finally the remaining words are decomposed into their stem and morphemes. This model is named as the combined model with 2.5K words. Then we increase the number of the most frequent words from 2500 to 5000, and apply the same procedure. The aim is to see the effect of the number lexicon entries left as words. This model is named as the combined model with 5.0K words.

4.3. Statistics of the Corpus

4.3.1. Training Corpus

We have collected our text material to construct statistical language models. Our text material is the articles of Milliyet newspaper belonging to different domains collected in a one month period. The domains, the number of words in each domain and the percentage of each domain in the text corpus are given in Table 4.2.

Table 4.2. Number of words and percentage of each domain in the training corpus

| Domain | Number of words | Percentage of the training corpus (%) |
|-------------------|-----------------|---------------------------------------|
| World News | 33534 | 9.43 |
| Economics | 86123 | 24.23 |
| Contemporary News | 65380 | 18.39 |
| Politics | 104467 | 29.39 |
| Daily Life | 65993 | 18.56 |
| Total | 355497 | 100 |

Also, the training data is grouped as: Train-1, Train-2,...,Train-5, using the domain groups given in Table 4.3. The reason is to enlarge the training corpus by adding news from different domains and to see the effect of this on the statistics of the training

corpus and test data.

Table 4.3. Different domains in the training data groups

| | |
|---------|--|
| Train-1 | World News |
| Train-2 | World News, Economics |
| Train-3 | World News, Economics, Contemporary News |
| Train-4 | World News, Economics, Contemporary News, Politics |
| Train-5 | World News, Economics, Contemporary News, Politics, Daily Life |

4.3.2. Test Data

Our test data is still the articles of the Milliyet newspaper collected in one day from five different domains. The collection days of the training corpus and the test data are not the same. The domains, the number of words in each domain and the percentage of each domain in the test data are given in Table 4.4.

Table 4.4. Number of words and percentage of each domains in the test data

| Domain | Number of words | Percentage of the training corpus (%) |
|-------------------|-----------------|---------------------------------------|
| World News | 683 | 9.73 |
| Economics | 1802 | 25.68 |
| Contemporary News | 1886 | 26.88 |
| Politics | 1253 | 17.86 |
| Daily Life | 1392 | 19.84 |
| Total | 7016 | 100 |

4.3.3. Number of Distinct Tokens

Number of distinct tokens is an important concept in the determination of the vocabulary size. It gives the minimum vocabulary size needed to cover 100 per cent of the training data. The selection of the tokens differs from model to model which

will be used during the construction of the language model. In word-based model, the tokens will be the words in the training text and in the combined model the tokens will be the words, stems, endings and morphemes which are available in the training corpus.

4.3.3.1. Word-based Model. Table 4.5 gives the statistics of the training corpus in terms of the number of distinct tokens using word-based model. The second column of the table gives the number of tokens, the third column gives the number of distinct tokens and the fourth column gives the number of new distinct tokens after the addition of the text data in a new domain.

Table 4.5. Number of tokens (words), number of distinct tokens and number of new distinct tokens in the word-based model

| | Number of Tokens (words) | Number of Distinct Tokens (distinct words) | Number of New Distinct Tokens (distinct words) |
|---------|--------------------------------|--|--|
| Train-1 | 33534 | 10258 | 10258 |
| Train-2 | 119657 | 23275 | 13017 |
| Train-3 | 185037 | 35399 | 12124 |
| Train-4 | 289504 | 46996 | 11597 |
| Train-5 | 355497 | 55931 | 8935 |

As shown in Table 4.5, the number of tokens and the number of distinct tokens are very high for this model. Addition of each domain introduces approximately 10 thousand new words to the corpus.

4.3.3.2. Combined Model with 2.5K Words. Table 4.6 gives the statistics of the training corpus in terms of the number of distinct tokens using the combined model with the most frequent 2500 words. In this model, the tokens are both the most frequent 2500 words, stems, endings and morphemes. Different than the previous table, a new column showing the number of words in that training set is added to the table.

Table 4.6. Number of words, number of tokens, number of distinct tokens and number of new distinct tokens in the combined model with the most frequent 2500 words

| | Number of Words | Number of Tokens | Number of Distinct Tokens | Number of New Distinct Tokens |
|---------|-----------------|------------------|---------------------------|-------------------------------|
| Train-1 | 33534 | 47676 | 5584 | 5584 |
| Train-2 | 119657 | 165225 | 9538 | 3954 |
| Train-3 | 185037 | 258952 | 13378 | 3840 |
| Train-4 | 289504 | 406790 | 15762 | 2384 |
| Train-5 | 355497 | 503863 | 18228 | 2466 |

In this model, the number of distinct tokens is comparable smaller than the word-based model. Also, the addition of each new domain to the corpus introduces approximately three thousand new tokens.

4.3.3.3. Combined Model with 5.0K Words. The only difference between this model and the model in part 4.3.3.2, is the number of most frequent words. In this model tokens are both the most frequent 5000 words, stems, endings and morphemes. The statistics of this model is given in Table 4.7. The addition of each new domain introduces approximately four thousand new tokens.

Table 4.7. Number of words, number of tokens, number of distinct tokens and number of new distinct tokens in the combined model with the most frequent 5000 words

| | Number of Words | Number of Tokens | Number of Distinct Tokens | Number of New Distinct Tokens |
|---------|-----------------|------------------|---------------------------|-------------------------------|
| Train-1 | 33534 | 43638 | 6738 | 6738 |
| Train-2 | 119657 | 151936 | 11366 | 4628 |
| Train-3 | 185037 | 238625 | 15401 | 4035 |
| Train-4 | 289504 | 374664 | 17874 | 2473 |
| Train-5 | 355497 | 464214 | 20358 | 2484 |

4.3.3.4. Comparison of Models. If we compare these three models in terms of number of distinct tokens, the combined model with the 2.5K most frequent words has the minimum number of distinct tokens. The reason is that, in this model most of the tokens are left as stems, morphemes and endings instead of words. Because of the morphological productivity of Turkish, it is possible to generate lots of new words by concatenating stems with endings and morphemes. Therefore, this model shows that, most of the words in the training corpus have the same stems, endings or morphemes. Although the combined model with the most frequent 5.0K words has words, stems, endings and morphemes as tokens, the number of distinct tokens is higher than the first model as more words are left as tokens. The word-based model has the highest number of distinct tokens; also addition of new data from different areas increases the number of distinct tokens drastically. The reason is that, if only a morpheme differs between two words, these two words are considered as different tokens.

4.3.4. Coverage

Coverage is an important metric for the recognition performance of a recognizer. The words that are in the test set but not available in the lexicon are called OOV words. If a word is an OOV word, then the recognizer has no chance to recognize it correctly. Therefore OOV words are the main source of recognition errors and the coverage gives us a rough idea about the maximum theoretical performance of the recognizer. Also, it is impossible to add all the words in the training set to obtain 100 per cent coverage over the training set because the vocabulary size will be very large, and this will overload the system. So the vocabulary size decision is very important to balance the trade-off between the OOV rate and the overloading system.

4.3.4.1. Coverage with the Word-based Model:. Table 4.8 shows the coverage of the word based model over the raining and test texts with the specified vocabulary size given in the first column which are the most frequent words in the training corpus.

It is interesting to note that, by using the most frequent 10 words in the corpus, it

Table 4.8. Coverage with respect to the word-based model

| Coverage Analysis (%) | | |
|-------------------------------|--------------|----------|
| Vocabulary Size (in words) | Training Set | Test Set |
| 10 | 15.39 | 14.96 |
| 20 | 19.08 | 18.83 |
| 50 | 24.59 | 24.19 |
| 100 | 29.82 | 28.68 |
| 500 | 44.59 | 43.28 |
| 1000 | 52.22 | 50.28 |
| 2000 | 60.66 | 58.57 |
| 3000 | 65.77 | 63.05 |
| 5000 | 72.37 | 69.51 |
| 10000 | 81.10 | 77.16 |
| 20000 | 89.07 | 83.57 |
| 30000 | 93.25 | 86.73 |
| 40000 | 95.85 | 88.00 |
| 50000 | 98.46 | 89.63 |
| 56931 | 100 | 90.49 |

is possible to cover approximately 15 per cent of the test data. Although the vocabulary size is very large, all the words in the training corpus (55931 words) can cover only the 90 per cent of the test set. The main reason for this is the morphological productivity of Turkish. 53.4 per cent of the words in the training corpus is used only ones and 15.6 per cent of the words are used only twice. However, if we consider the stems, endings or the morphemes of the words which occur only once or twice, we expect to see that they share the same stems, endings and morphemes with the most frequent words and this is the motivation for our proposed combined model.

4.3.4.2. Coverage with the Combined Model with 2.5K Words. In this combined model, instead of words, coverage is determined with the tokens, the most frequent 2500 words, stems, endings and morphemes. The coverage of these tokens over the training and test texts with the specified vocabulary size is shown in Figure 4.9.

Table 4.9. Coverage with respect to the combined model with 2.5K words

| Coverage Analysis (%) | | |
|-----------------------|--------------|----------|
| Vocabulary Size | Training Set | Test Set |
| 10 | 14.50 | 15.03 |
| 20 | 19.67 | 19.70 |
| 50 | 27.87 | 28.18 |
| 100 | 35.33 | 35.66 |
| 500 | 56.38 | 55.50 |
| 1000 | 67.48 | 66.60 |
| 2000 | 79.38 | 78.09 |
| 3000 | 86.22 | 84.94 |
| 5000 | 93.29 | 92.04 |
| 10000 | 98.12 | 96.51 |
| 18228 | 100 | 98.10 |

Although in word-based model, only 81 per cent coverage is attained over the training set with the most frequent 10000 distinct tokens, this coverage increase to 98 per cent for this model. Therefore, this table is the illustration of the productive inflectional and derivational suffixations of the Turkish morphology. As an example, we can give the word “politikalaradaki” (on the politics). It occurs only once in the training corpus and lexicon size of more than 20K words are needed to cover this word. However, this word can be covered by using a lexicon size of approximately 3K words if the combined model is used, because the words “politika (stem)”, and “lardaki (ending)” are one of the most frequent tokens in this models.

4.3.4.3. Coverage with the Combined Model with 5.0K Words. Table 4.10 illustrates the coverage of the combined model with the most frequent 5.0K words over the training and test texts with the specified vocabulary size (distinct tokens) given in the first column.

Table 4.10. Coverage with respect to the combined model with 5.0K words

| Coverage Analysis (%) | | |
|-----------------------|--------------|----------|
| Vocabulary Size | Training Set | Test Set |
| 10 | 13.36 | 12.66 |
| 20 | 17.43 | 17.77 |
| 50 | 24.56 | 24.76 |
| 100 | 31.34 | 31.73 |
| 500 | 50.89 | 50.41 |
| 1000 | 61.16 | 60.32 |
| 2000 | 72.50 | 71.30 |
| 3000 | 79.30 | 78.65 |
| 5000 | 87.92 | 86.65 |
| 10000 | 96.77 | 95.07 |
| 20000 | 99.93 | 97.90 |
| 20358 | 100 | 97.96 |

If we compare this table with the previous one, combined model with the most frequent 2.5K words, it is clear that, with the same vocabulary size higher OOV rates are achieved. The reason is that more words are left as tokens in this model.

4.3.4.4. Comparison of Models. If we compare these three models with respect to coverage, similar results are obtained as in the comparison in terms of number of distinct tokens. The smallest number of OOV rate is achieved with the combined model with 2.5K and the highest OOV rate is achieved with the word-based model if we use the same vocabulary size. With the most frequent 10000 tokens, although

the coverage over the training set is 98.12 per cent with the combined model, coverage decreases to 81.1 per cent for the word-based model. Also, if we left more words as tokens, the coverage decreases to 96.77 per cent. This results show that, using smaller tokens as recognition units, higher coverage is achieved with the same vocabulary size.

In all the languages, the smallest units are the phonemes. Therefore using only phonemes as vocabulary entities will give us 100 per cent coverage independent of the corpus size. However, from the speech recognition point of view, smaller units are lack of acoustic information compared to the larger units and this will cause poorer recognition performance.

4.3.5. Bigram Models

In this thesis, we use bigram models with back-off smoothing. These models are constructed using the Language Modeling tools of the HTK toolkit [8]. Although in models with smaller recognition units, using higher order N-grams are more preferable, the reason of using bigram models is the limited computational aspects of HTK in constructing language models in recognition mode.

In this part, a detailed bigram analysis of the training corpus with respect to the self and the test set will be made, and also the bigram hit, the percentage of the bigrams both occurring in the training and the test set, will be given.

4.3.5.1. Bigram Analysis of the Word-based Model. The main comparison metric in bigram analysis is the perplexity, average branching factor. Although Turkish is a SOV type language, a sentence can be uttered in six different ways. Therefore, due to the free word order of Turkish, the branching factor from a word can be very high. Table 4.11 shows the bigram analysis of the word-based model with respect to the self and the test data, also the bigram hit percentage of the test data.

The second column of the table shows the self perplexity analysis of the training

Table 4.11. Bigram analysis for the word-based model

| | Bigram perplexity with respect to self | Bigram perplexity with respect to test | Bigram hits in test data (%) |
|---------|---|---|---------------------------------|
| Train-1 | 753.95 | 659.26 | 30.4 |
| Train-2 | 711.30 | 959.06 | 34.9 |
| Train-3 | 936.18 | 1105.38 | 36.9 |
| Train-4 | 957.81 | 1217.26 | 38.9 |
| Train-5 | 1063.53 | 1278.17 | 39.7 |

corpus. As expected, adding news from different domains increases the perplexity, because new bigrams and words added to the set. This means the number of alternatives that can follow a word increase with the addition of the new data. At one point, the perplexity has to be saturated for a better bigram model; however our training corpus is very small for this saturation point. Also this small corpus size is the reason of the little bit perplexity decrease from Train-1 to Train-2. If we look at the perplexity analysis with respect to the test set, the perplexity increases with the increasing corpus size. However, if our corpus is large enough, we can expect to see a decrease in the test perplexity, because with larger amounts of data better estimates for the unseen data is achieved. Also the last column of the table shows the level of the uncertainty in our model, as the percentage of the bigram hits is very small. So we can say that, all of the analysis shows that our corpus is not successful enough in modeling the bigrams.

4.3.5.2. Bigram Analysis of the Combined Model with 2.5K Words. Table 4.12 shows the bigram analysis of the combined model with 2.5K with respect to the self and the test data, also the bigram hit percentage of the test data.

From the table, it is interesting to note that, addition of new data from different domains decreases the training set perplexity. Also the self perplexity values are closer to each other. However, the situation is reversed for the word-based case. Although addition of new data introduces new distinct tokens and bigrams to the training corpus,

Table 4.12. Bigram analysis for the combined model with 2.5K words

| | Bigram perplexity with respect to self | Bigram perplexity with respect to test | Bigram hits in test data (%) |
|---------|---|---|---------------------------------|
| Train-1 | 208.11 | 537.30 | 33.6 |
| Train-2 | 171.88 | 473.73 | 44.9 |
| Train-3 | 201.14 | 384.82 | 51.6 |
| Train-4 | 192.23 | 359.75 | 56.0 |
| Train-5 | 197.58 | 334.45 | 58.7 |

the uncertainty of our model decreases. As lots of tokens which lead to the better estimates of the previously seen bigrams added to the corpus. Therefore, this language model works better in modeling the unseen data. This situation is shown in the third column of the table, the test set perplexities decrease with the increasing corpus size. Also, the percentage of hit values is comparable higher than the word-based model.

4.3.5.3. Bigram Analysis of the Combined Model with 5.0K Words. Table 4.13 shows the bigram analysis of the combined model with 5.0K with respect to the self and the test data, also the bigram hit percentage of the test data.

Table 4.13. Bigram analysis for the combined model with 5.0K words

| | Bigram perplexity with respect to self | Bigram perplexity with respect to test | Bigram hits in test data (%) |
|---------|---|---|---------------------------------|
| Train-1 | 322.27 | 784.14 | 28.1 |
| Train-2 | 264.64 | 775.97 | 37.4 |
| Train-3 | 305.72 | 636.86 | 43.5 |
| Train-4 | 285.84 | 579.04 | 48.1 |
| Train-5 | 291.67 | 528.91 | 50.9 |

As shown in the table, perplexity with respect to self and test set decrease with

increasing corpus size. However, the perplexity values are higher than the previous model. Again, this is because of the number of words left as tokens.

4.3.5.4. Comparison of Models. If we compare these three models in terms of bigram perplexity, we can say that combined models are better in language modeling with the available corpus. In word-based model, the addition of new data always increases the perplexity, which means increases the uncertainty of the model. However, in combined models the newly added data improves the certainty of the language model by leading to the better estimates over the previous bigram probabilities. Also, the perplexity for the combined model with 2.5K is lower than the combined model with 5.0K. It can be explained with the number of tokens left as words, stems, ending and morphemes. In the first one, most of the tokens are smaller units compared to the latter one, therefore better in modeling the unseen data.

4.3.6. Statistics with Respect to Test Set

As mentioned before our test set is the one of the daily news of the Milliyet newspaper collected from five different domains, and our training set is the news collected from the same domains in a one month period. In this part, using the statistics of the training corpus, the statistics of the test set will be evaluated in terms of coverage and the bigram models.

If we consider coverage, the vocabulary size will be very huge if we generate the lexicon with all the units in the corpus. Therefore we decided to the optimum vocabulary size as 10K recognition units. This will be the vocabulary size that will be used in the recognition experiments. The reason of this decision is that with 10K words more than 80 per cent self coverage is achieved over the training set in word-based model. However, the coverage is slightly higher for the other models. The coverage statistics of the test set with the proposed models are given in Table 4.15.

As shown from the table, maximum coverage is achieved using the combined

Table 4.14. Coverage analysis of the test set

| Coverage Analysis (%) | | |
|-----------------------|-----------------------|-----------------------|
| Word-based Model | Combined Model (2.5K) | Combined Model (5.0K) |
| 77.16 | 96.51 | 95.07 |

model with the most frequent 2.5K words. The maximum OOV rate is for the word-based model.

Table 4.15. Bigram analysis of the test set

| Bigram Analysis (Perplexity) | | |
|------------------------------|-----------------------|-----------------------|
| Word-based Model | Combined Model (2.5K) | Combined Model (5.0K) |
| 476.68 | 294.36 | 433.78 |

The bigram analysis of the test set is given in Table 4.15. As seen from the table, the smallest perplexity is for the combined model with 2.5K most frequent words. Also the perplexity values for the other combined model and the word-based models are similar to each other.

4.4. Recognition Experiments

We perform recognition experiments on the recording of the test data with only one female speaker. All the test data is recorded at 16 KHz, and 16 bit wav format. As the recognition performance of the models will be tested for LVCSR task, the recordings are not only one utterance. Each recording contains approximately 10 sentences, and each sentence is uttered in a manner like the continuous speech so there are not long silence intervals between each words.

4.4.1. The Recognizer

The recognizer used in these experiments is designed using the training and recognition tools of the HTK toolkit. The properties of the recognizer are as follows:

- It is trained using the labeled recordings of 10 people each of them uttering phonetically balanced 149 sentences and words. Also as unlabeled data, totally 8923 utterances of 195 different speakers are used. The important point is that in each of the models, the context-dependent triphones are trained according to the language model. The reason for this is to better model the morphemes, endings and stems available in the acoustical training data.
- Each of the recordings are at 16 KHz and 16 bit wav format.
- Context dependent triphones are used, and data-driven clustering is applied to decrease the number of least frequent triphones in the data.
- Bigram language models are constructed for each model using the training text that is arranged according to model. Bigram probabilities with back-off smoothing are used.

The performance of the model is evaluated using Per cent Correct and Per cent Accuracy. Firstly optimal string match is made between the recognized utterance and the original text. Then the substitution errors (S), deletion errors (D) and insertion errors (I) can be calculated. If N is the total number of labels, words, stems, endings morphemes, the percentage correct and accuracy are calculated as,

$$\text{Per cent Correct} = \frac{N - D - S}{N} \times 100\% \quad (4.1)$$

$$\text{Per cent Accuracy} = \frac{N - D - S - I}{N} \times 100\% \quad (4.2)$$

Also the selection of the recognition parameters is an important point because they have a great influence on the recognition performance. These parameters are the

pruning threshold (t), word insertion penalty (p) and language model scaling (s). The pruning threshold is necessary to eliminate the tokens that have no chance to succeed. This parameter makes a significant reduction in the computation. Our observations on previous recognition experiments showed that selecting the pruning threshold as 300 is a compromising solution. The word insertion penalty and the language model scaling factor determine the number of the insertion and the deletion errors. If x is the language model probability, it will be converted to $s \times x + p$, before being added to each word-end node [8]. The optimum parameters for p and s are decided after some recognition experiments over one of the test recordings for all of the proposed models.

4.4.2. Recognition Experiments with Word-based Model

In word based model, there are 55931 distinct words in the training corpus. Therefore, if we add all the words to the lexicon, we will have a huge recognition network. Although, the self-coverage will be 100 per cent, this large number of words in the lexicon will overload the recognizer. Also, lots of acoustically similar words will be in the vocabulary, and this will cause an increase in the error rate. Therefore, we decided to add only the most frequent 10000 words to the lexicon. With this vocabulary size, 81 per cent coverage is attained over the training set.

Table 4.16. Selection of the p and s parameters for the word-based model
(correct/accuracy)

| p/s | 0 | 5 | 10 | 15 | 20 |
|-----|---------------|-------------|-------------|--------------------|-------------|
| -10 | 33.00/-7.58 | 45.81/28.08 | 48.77/41.48 | 47.29/44.33 | 45.81/44.33 |
| -5 | 29.06/-28.57 | 44.33/22.17 | 49.75/40.89 | 46.80/43.35 | 46.31/44.83 |
| 0 | 24.14/-54.14 | 43.84/16.75 | 50.25/37.93 | 49.75/45.81 | 46.80/45.32 |
| 5 | 15.27/-122.17 | 42.36/10.84 | 51.23/34.98 | 51.23/47.29 | 47.29/44.83 |
| 10 | 11.33/-217.24 | 38.42/-8.37 | 50.74/31.03 | 52.22/44.83 | 48.77/45.32 |

Firstly, we perform the recognition experiment on the one of the test recordings to select the optimal parameters. The recognition results are given in in Table 4.16. As seen from the table the best results are achieved for the parameters $p = 10$ and

$s = 15$ for this model. Then we perform the experiment on the test recordings for the below cases:

- i. With the lexicon having the most frequent 10000 words in the training corpus. This is the same experiment with the parameter selection case. There will be some OOV words in the lexicon.
- ii. Repeat the experiment in *ii.* with speaker adaptation. Here our aim is to eliminate the effect of speaker variation and also the microphone.
- iii. With lexicon having the most frequent 10000 words in the training corpus, and also the OOV words in the test recording. The test data will be added to training corpus in language model construction. Here our aim is to see our theoretical maximum recognition performance.

The results for the i.'th and the ii'th experiment for each test recordings are given in Table 4.17. "NA" gives recognition results without speaker adaptation and "A" gives results with speaker adaptation. We expect better recognition performance for adapted models. However, in some recordings, 10 and 17, recognition performance drops for adapted models. The reason is that in these recordings the silence interval between sentences is detected as the end of the recognition network. So all the utterances are not recognized and this increases the deletion errors during evaluation of the results. Also in recording 31, the accuracy is dropped, given in Tables 4.17 and 4.18.

In addition to that, in some of the recordings after the speaker adaptation, the recognition performance is significantly better than the others. For speaker adaptation 30 minutes of recordings from broadcast news are collected. The collection data is not from the training corpus. Therefore some of the triphones are adapted better than the others. This is the reason of these slight differences between recognition performances in each of the recordings.

Also it is interesting to note that in recording 16, no recognition results are obtained. The reason is that the pruning threshold is not large enough and the optimal path is pruned before the end of the utterance. Increasing the pruning threshold will

Table 4.17. Test results for the i.'th and the ii.'th experiments in terms of per cent of correct

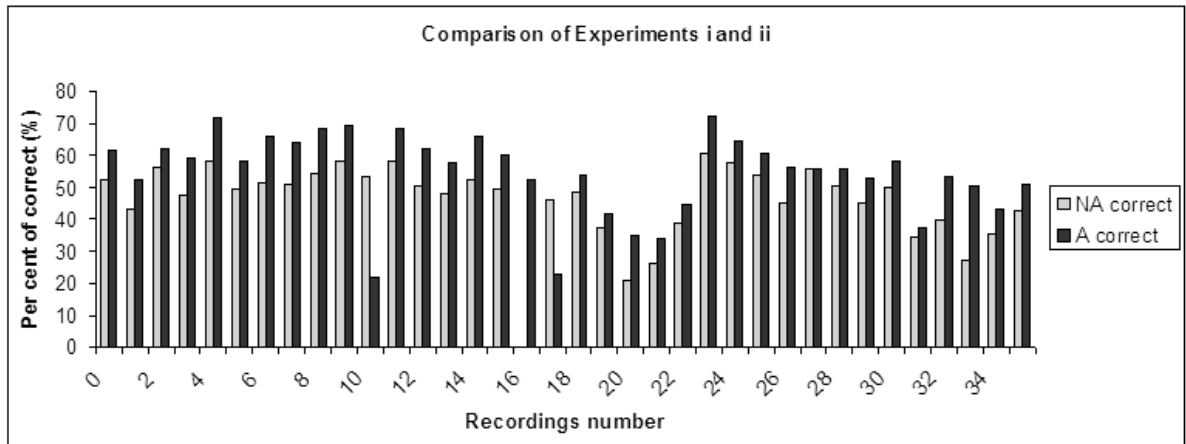
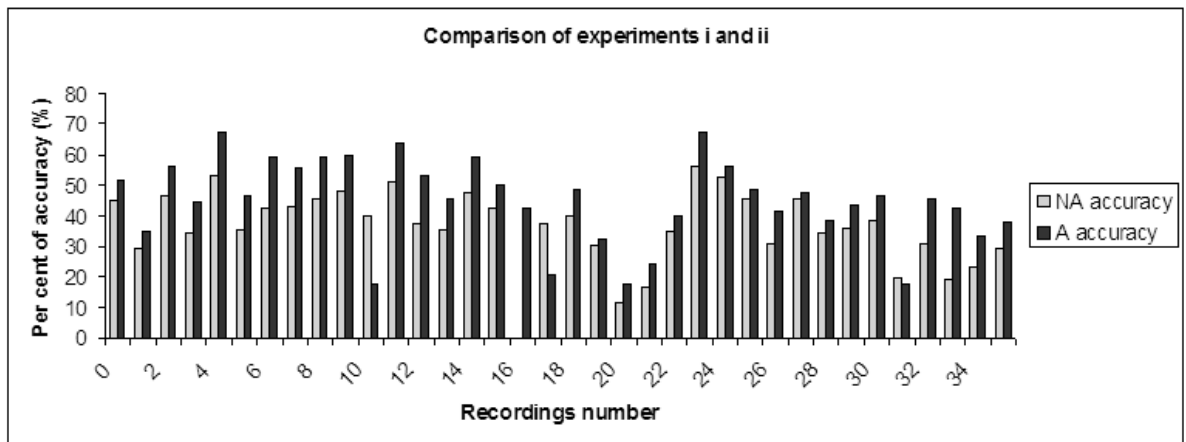


Table 4.18. Test results for the i.'th and the ii.'th experiments in terms of per cent of accuracy

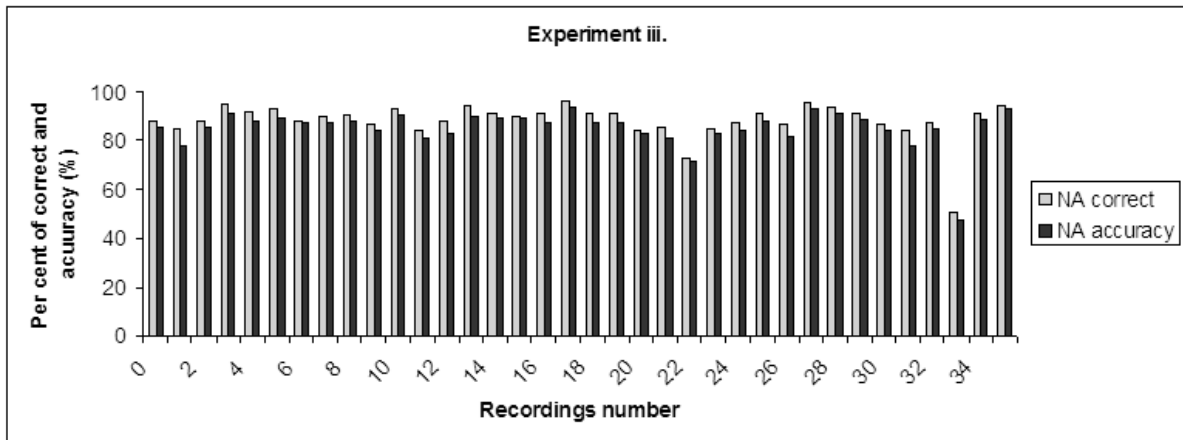


eliminate this problem; however, it will increase the computation time. Therefore, we decided to throw out the recordings that are not recognized.

The results of the iii.'th experiment, no OOV case, is given in Table 4.19. The results are comparable better than the experiments i.'th and the ii.'th, because the test data is involved in the training corpus so no OOV words is available, also language model is now better representing the test set.

The over all results of these three experiments is given for comparison issues. Also

Table 4.19. Test results for the iii.'th experiment



the OOV rates and the perplexities are given. To make better comparison between each experiment, we throw out the recordings 10, 16 and 17 from the test data. Test results are shown in Table 4.20.

Table 4.20. Test results for all the experiments

| Experiment no | OOV rate (%) | Perplexity | Correct (%) | Accuracy (%) |
|---------------|--------------|------------|-------------|--------------|
| i. | 22.60 | 485.49 | 47.56 | 37.96 |
| ii. | 22.60 | 485.49 | 57.17 | 47.39 |
| iii. | 0 | 940.20 | 88.34 | 85.44 |

As seen from Table 4.20, speaker adaptation improves the recognition performance 16.8 per cent. Also with only 77.4 per cent coverage (recognizer has no chance to recognize 22.6 per cent of the data) 57.17 per cent correct recognition is achieved with the adapted models. If we add test data to the training corpus, the perplexity increases due to the addition of the new bigrams. However, there is a significant increase in the recognition performance, because there are no OOV words and language model contains also the test data.

4.4.3. Improvements to the Word-based Model

Our motivation for this model is that, although some words follow each other in most of the context, they are behaved as different tokens. For example, one of the mostly known words is “her iki”. Although, in some texts they are written together, in some of them they are written separately, so considered as different words. If we look at the situation from the recognizer’s point of view, both of the tokens “her” and “iki” are smaller recognition units, therefore they have less acoustical information compared to the word “heriki”. Here, our aim is to combine these small unit word pairs that have higher bigram probabilities to achieve better acoustical information.

We apply two different approaches for this. Firstly, we generate our language model using the most frequent 10000 words using the training corpus. Then we combine word pairs having higher bigram probabilities and whose total length is smaller than 10 characters. By this way, we combine 195 word pairs and added this new pairs to the lexicon. Now, we have a lexicon size of 10195 words. We made the same modifications to the training corpus. Finally, we generate the language model using the modified corpus. This model is named as `wm_2` in the tables. Secondly, we generate all the bigrams available in the training corpus then apply the same criterion to combine the words. By this way we combine 386 word pairs. After this modification to the corpus we select the most frequent 10000 words for our new vocabulary. This model is named as `wm_2` in the tables. The label of our word-based model becomes `wm_1`.

Table 4.21. Test results for all the experiments for modified word-based model

| Models | OOV words | OOV rate (%) | Perplexity | Correct (%) | Accuracy (%) |
|-------------------|-----------|--------------|------------|-------------|--------------|
| <code>wm_1</code> | 1668 | 22.75 | 476.74 | 47.69 | 38.00 |
| <code>wm_2</code> | 1667 | 22.89 | 503.88 | 47.31 | 37.97 |
| <code>wm_3</code> | 1680 | 23.10 | 488.55 | 47.29 | 37.82 |

Table 4.21 shows the comparison of the recognition experiments for the modified word-based models. The recognition results are very similar; however the best results are achieved with the original word-based model. The most evident improvement of

the modified models is the decrease in the insertion errors.

4.4.4. Recognition Experiments with Combined Model with 2.5K Words

In this model, there are 18228 distinct tokens, which are from the most frequent 2500 words, stems and endings of the most frequent 2500 words after the first most frequent 2500 ones and also the stems and morphemes of the remaining words. The most frequent 10000 words are the lexicon entities. With this vocabulary size 96.51 per cent coverage is attained over the test set.

The first step is the selection of the p and s parameters. Again the recognition experiments are performed on the same recording to decide the best parameter pair. The results of the parameter selection experiments are given in Table 4.22. Best results are achieved with the parameter set $p = -10$ and $s = 10$.

Table 4.22. Selection of the p and s parameters for the combined model with 2.5K words (correct/accuracy)

| p/s | 0 | 5 | 10 | 15 | 20 |
|-----|---------------|--------------|--------------------|-------------|-------------|
| -10 | 27.24/-14.18 | 47.76/31.34 | 55.22/45.15 | 48.88/42.16 | 44.4/42.16 |
| -5 | 23.51/-35.82 | 45.52/22.76 | 54.85/42.91 | 50.75/42.91 | 46.27/42.54 |
| 0 | 17.16/-81.72 | 44.78/13.43 | 54.10/36.19 | 50.37/38.81 | 47.01/41.04 |
| 5 | 15.30/-179.10 | 42.54/4.10 | 53.36/31.72 | 53.36/39.18 | 48.13/39.55 |
| 10 | 13.43/-344.40 | 35.45/-19.40 | 52.99/28.36 | 54.85/37.69 | 47.39/35.82 |

Then we perform recognition experiments according to the i.'th and the ii.'th cases. We do need to perform the iii.'th experiment because in this model the OOV rate is very small compared to word-based model. The results of this experiments are shown in Tables 4.23 and 4.24.

In those experiments, in addition to recordings 10 and 17, in recording 15, adaptation decreases the recognition performance. Again the long silence interval is detected

Table 4.23. Test results for the i.'th and the ii.'th experiments in terms of per cent of correct

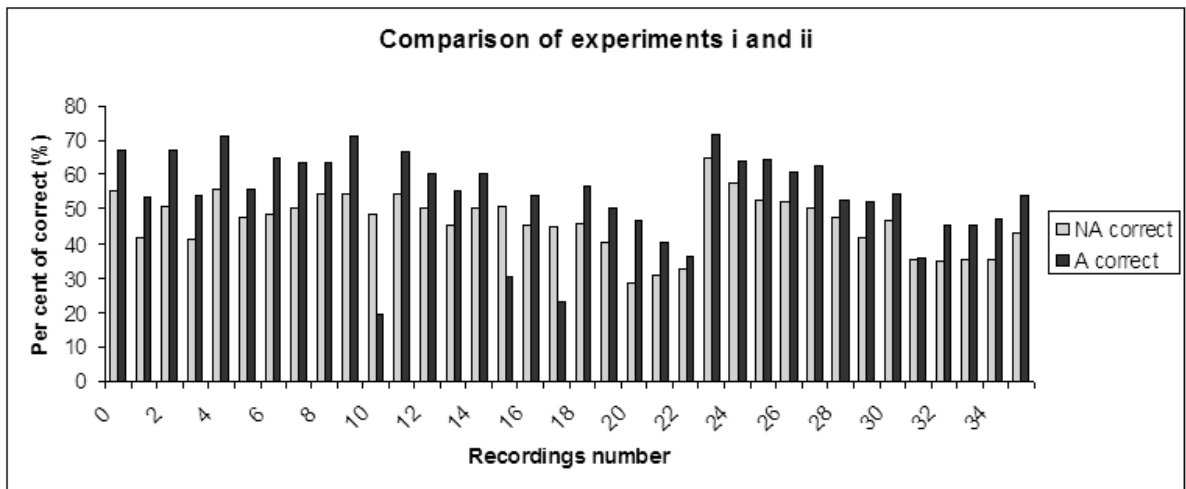
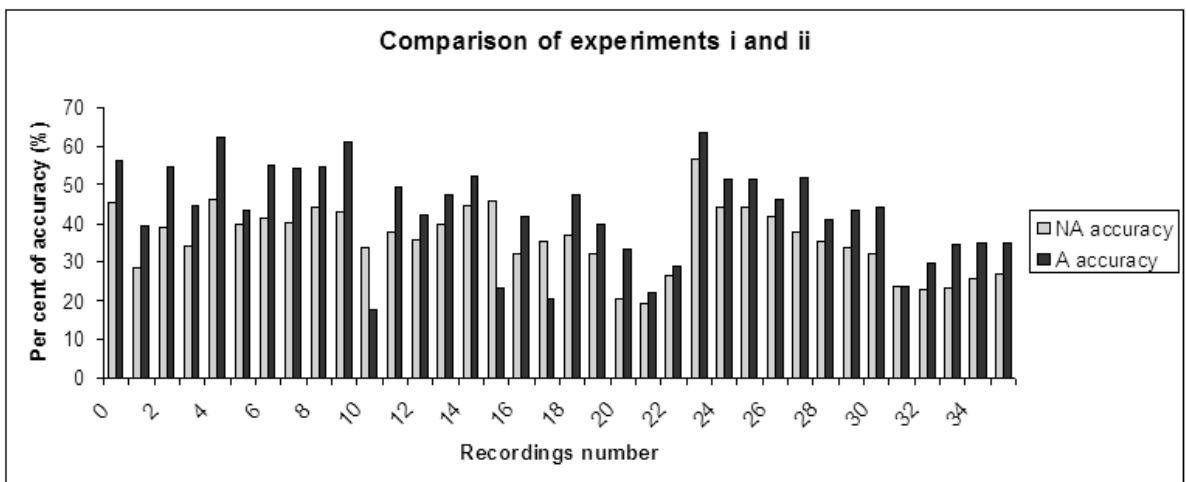


Table 4.24. Test results for the i.'th and the ii.'th experiments in terms of per cent of accuracy



as the end of the speech. Therefore all the utterances are not recognized. Also, recording 16 can be recognized with the same pruning threshold. Acoustic models for each of the language model are trained with the modified triphones arranged according to the model. Therefore, the poorly trained triphones change from model to model. Therefore a perfectly recognized recording in a model may not be recognized with other model. In that case, we throw out the recordings 10, 15 and 17 from the test set and then compare the recognition performances in terms of percentage of correct and accuracy between the adapted and the general acoustic models.

The the overall results of the recognition experiments are illustrated in Table 4.25. In this language model, speaker adaptation increases the recognition performance 18.8 per cent.

Table 4.25. Test results for all the experiments

| Experiment no | OOV rate (%) | Perplexity | Correct (%) | Accuracy (%) |
|---------------|--------------|------------|-------------|--------------|
| i. | 3.51 | 303.32 | 46.14 | 35.59 |
| ii. | 3.51 | 303.32 | 56.83 | 44.98 |

4.4.5. Recognition Experiments with Combined Model with 5.0K Words

In this model, the vocabulary is generated using the most frequent 10000 tokens, most frequent words, stems, endings, morphemes, and recognition experiments with the general and the adapted acoustic models are performed. For this model, with this vocabulary size the coverage over the training set is obtained as 96.77 per cent. Therefore, we do not consider the no OOV case (iii.'th case) during the recognition experiments.

Table 4.26. Selection of the p and s parameters for the combined model with 5.0K words (correct/accuracy)

| p/s | 0 | 5 | 10 | 15 | 20 |
|-----|---------------|--------------|--------------------|-------------|-------------|
| -10 | 26.23/-22.13 | 47.54/30.33 | 53.69/43.03 | 51.64/43.85 | 44.26/40.57 |
| -5 | 19.26/-51.23 | 45.08/19.26 | 53.28/39.75 | 52.87/43.44 | 47.95/41.80 |
| 0 | 14.34/-100.00 | 44.26/11.48 | 54.51/37.30 | 52.05/40.16 | 48.36/42.21 |
| 5 | 12.70/-196.72 | 39.75/-2.05 | 52.87/29.51 | 52.46/37.30 | 51.64/42.21 |
| 10 | 10.25/-380.33 | 37.30/-23.36 | 50.82/21.72 | 52.05/31.15 | 51.64/39.75 |

As seen from Table 4.26, the best percentage of correct results are obtained with the parameters $p = 0$ and $s = 10$, however the accuracy is very small for this parameters. Therefore, we select the parameter pairs $p = -10$ and $s = 10$ for the recognition experiments.

Table 4.27. Test results for the i.'th and the ii.'th experiments in term of per cent of correct

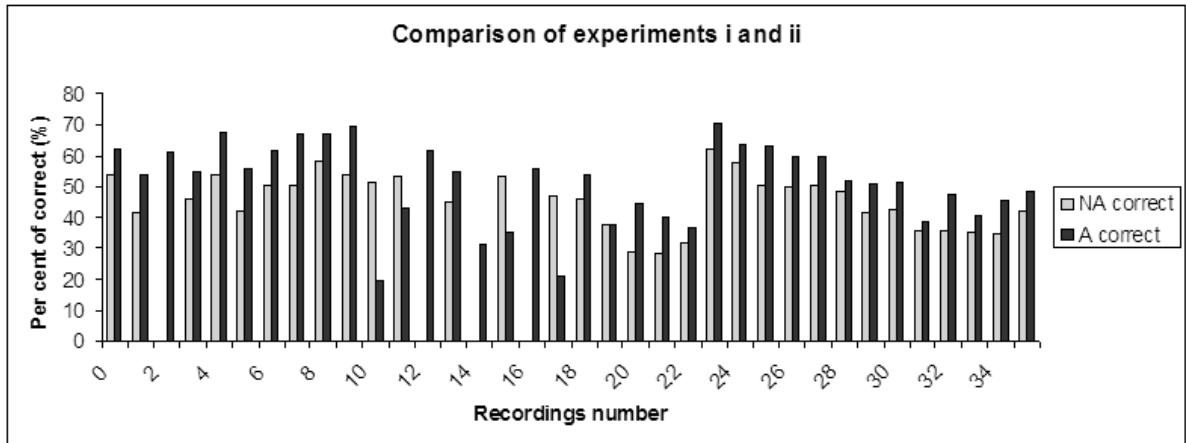
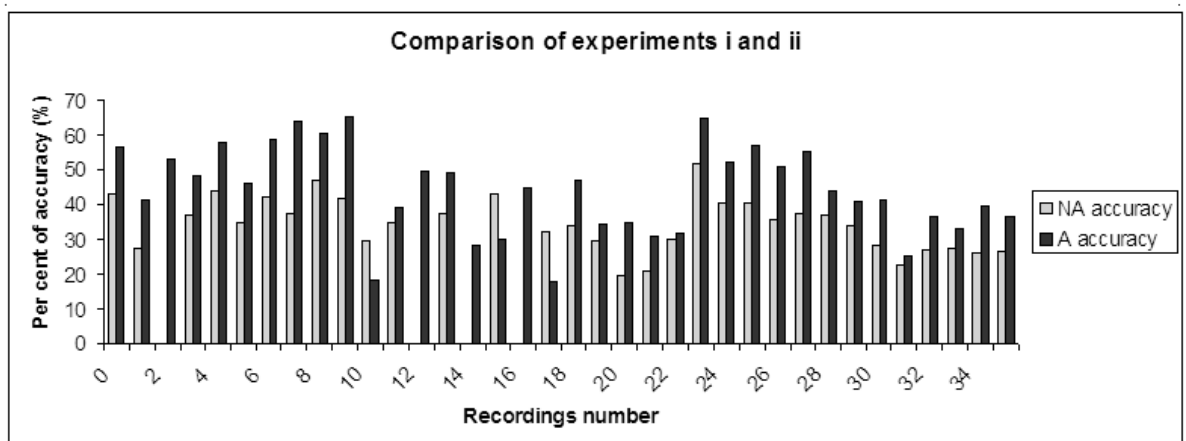


Table 4.28. Test results for the i.'th and the ii.'th experiments in terms of per cent of accuracy



In Table 4.27, recordings 2, 12, 14 and 16 are not recognized with the specified pruning threshold. However, they are recognized perfectly with the adapted models. This is because of the poorly trained triphones. In recordings 10, 11, 15, and 17 adaptation of HMM's degrades the performance, because the long silence interval between sentences is recognized as the end of the utterance. So, we ignore recordings 2, 12, 14, 16 and 10, 11, 15, 17, from the test data. The overall results after this pruning of outlier recordings are given in Table 4.29.

If we evaluate the results of the overall experiments in this model, we see that

Table 4.29. Test results for all the experiments

| Experiment no | OOV rate (%) | Perplexity | Correct (%) | Accuracy (%) |
|---------------|--------------|------------|-------------|--------------|
| i. | 5.08 | 422.36 | 44.76 | 34.33 |
| ii. | 5.08 | 422.36 | 54.37 | 46.84 |

there is a performance increase of 17.67 per cent for the speaker adapted models.

4.4.6. Comparison of Models

As mentioned before, acoustic models for each language model is trained according to the model. Therefore, the completely recognized test data differs from model to model. To make a more fair comparison between the models, we throw out the outlier recordings in all of the models. These recordings are, 2, 10, 11, 12, 14, 15, 16 and 17. The comparison of all the models in terms of OOV rate, bigram perplexity, per cent of correct and accuracy are given in Table 4.30.

Table 4.30. Comparison of all the proposed models

| Models | OOV rate (%) | Perplexity | Correct (%) | Accuracy (%) |
|--------------------|--------------|------------|-------------|--------------|
| Word-based (NA) | 23.45 | 462.23 | 46.29 | 36.37 |
| Word-based (A) | 23.45 | 462.23 | 55.80 | 45.49 |
| Combined-2500 (NA) | 3.63 | 297.23 | 45.38 | 35.12 |
| Combined-2500 (A) | 3.63 | 297.23 | 55.93 | 44.38 |
| Combined-5000 (NA) | 5.08 | 422.36 | 44.76 | 34.33 |
| Combined-5000 (A) | 5.08 | 422.36 | 54.37 | 46.84 |

From Table 4.30, we can say that although maximum OOV rate is for the word-based model, higher recognition rates are achieved with this model. In speaker adapted models, combined model with the most frequent 2.5K words shows a slight improvement which can be explained with the adaptation data.

5. CONCLUSIONS

In this thesis, we have search for the appropriate base recognition units for LVCSR. The idea of words as recognition units, works well for radiological dictation system because radiology area has its own specific vocabulary, also the perplexity over the test set is very small. However, in the general Turkish the situation is reversed. Therefore, in previous researches base recognition units like words, stems, endings and morphemes are proposed. In Turkish broadcast dictation system, we try the combination of all the proposed units. The reason for this is that although each recognition unit has its own superiority over the remaining ones in some comparison statistics, each of them has its drawback. Therefore, in our model, we prefer the combination of all the proposed units to overcome these problems. Our combined model is compared with the classical word-based model in terms of the comparison statistics like number of distinct tokens, coverage, bigram perplexity and recognition performance. The results are summarized in Table 5.1.

Table 5.1. Summary of the proposed models in terms of the defined comparison statistics

| Models | Word-based | Combined-2500 | Combined-5000 |
|---|------------|---------------|---------------|
| Number of distinct tokens (in Training Corpus) | 55931 | 18228 | 20358 |
| Coverage over the test set (%) (with 10K tokens) | 77.16 | 96.51 | 95.07 |
| Bigram perplexity (with 10K tokens) | 476.68 | 294.36 | 433.78 |
| Recognition performance (%) (with 10K tokens) | 46.29 | 45.38 | 44.76 |

Our desired language model is to obtain high coverage with smaller perplexity which is the case in the radiological dictation system. As shown from the table, the desired results are obtained with the combined model with 2.5K words. The reason is

that, the number of distinct tokens is smaller compared to the other models, therefore higher coverage is achieved with the same vocabulary size. Although, this model is the best one in terms of the comparison metrics, number of distinct tokens, coverage and perplexity, recognition performance is lower than the word-based model. It is explained with the number of smaller recognition units in the vocabulary. In the combined model, lexicon entries are the words, stems, endings and morphemes. Therefore, the lengths of the units are very different from each other. Words are considered as longer units and morphemes are considered as smaller units. This unbalanced vocabulary entry situation generates a handicap from the point of the recognizer and this cause a slight decrease in the recognition performance of the combined model with 2.5K words although the coverage is comparable higher than the word-based model.

As a consequence, we can say that although coverage is small and the perplexity is high compared to the other models, the best result is with the word-based model. Also our corpus is not large enough for accurate estimates of the bigram model. Therefore larger amounts of data need to be collected.

A further research can be to apply the word based model to more specific news domains like economics, politics and sports. By this way, we limit our vocabulary and with a moderate size of corpus better statistical model estimates and coverage can be achieved. This can be a solution for a better Turkish dictation system. Also, new base units can be proposed. The main drawback of our proposed models is the unbalanced length recognition units. Therefore, concatenation of specified number of syllables can be used to balance the vocabulary entries. In addition, consecutive morphemes with higher bigram probabilities can be concatenated to overcome the problem of smaller recognition units.

REFERENCES

1. Rosenfeld, R., “Two Decades of Statistical Language Modeling: Where Do We Go From Here?”, *Proceedings of the IEEE*, Vol. 88, pp. 1270-1278, August 2000.
2. Huang, X., A. Acero and H.W. Hon, *Spoken Language Processing A Guide to Theory, Algorithm, and System Development*, Upper Saddle River, New Jersey, 2001.
3. Rabiner L. R. and H. B. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, New Jersey, 1993.
4. Rabiner, L. R., “A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition”, *Proceedings of the IEEE*, Vol. 77, pp. 257-286, February 1989.
5. Jelinek, F., *Statistical Methods for Speech Recognition*, MIT Press, Massachusetts 1997.
6. İskurt, A., *Radiological Report Entry via Speech*, M.S. Thesis, Boğaziçi University, 2002.
7. Dutağacı, H., *Statistical Language Models for Large Vocabulary Continuous Speech Recognition*, M.S. Thesis, Boğaziçi University, 2002.
8. Young S., D. Ollason, V. Valtchev and P. Woodland, *The HTK book (for HTK Version 3.2)*, Entropic Cambridge Research Laboratory, March 2002.
9. Fromkin, V., R. Rodman and N. Hyams, *An Introduction to Language*, Thomson Heinle, Massachusetts, 2003.
10. Hakkani-Tür, D., K. Oflazer and G. Tür, *Statistical Morphological Disambiguation for Agglutinative Languages*, Technical Report, Bilkent University, 2000.
11. Erguvanlı-Taylan, E., A. Göksel, M. Nakipoğlu-Demiralp, *Structure of Modern*

- Turkish*, TK 204 Class Notes, Boğaziçi University, 2003.
12. Oflazer, K., “Two-level Description of Turkish Morphology”, *Literary and Linguistic Computing*, Vol. 9, No.2, 1994.
 13. Oflazer, K. and H. C. Bozşahin, “Turkish Natural Language Processing Initiative: An Overview”, *Proceedings of the Third Turkish Symposium on Artificial Intelligence and Artificial Neural Networks*, Ankara, June 1994.
 14. Erguvanlı E. E., *The Function of Word Order in turkish Grammar*, Ph.D. Thesis, University of California, Los Angeles, 1979.
 15. Çetinoğlu, Ö., *A Prolog Based Natural Language Infrastructure for Turkish*, M.S. Thesis, Boğaziçi University, 2001.
 16. Siivola, V., M. Kurimo and K. Lagus, “Large Vocabulary Statistical Language Modeling for Continuous Speech Recognition in Finnish”, *Proceedings of the 7th European Conference on Speech Technology and Communication, EUROSPEECH 2001*, Aalborg, Denmark, 2001.
 17. Kwon, O. W. and J. Park, “Korean Large Vocabulary Continuous Speech Recognition with Morpheme-based Recognition Units”, *Speech Communication*, Vol. 39, pp. 287-300, January 2002.
 18. Kanevsky *et al.*, “Statistical Language Model for Inflected Languages”, US patent No:5,835,888,1998, 1998.
 19. Mengusoglu, E. and O. Deroo, “Turkish LVCSR: Database Preparation and Language Modeling for an agglunitative Language”, *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP-2001 Student Forum*, Salt Lake City, May 2001.