SUB-WORD LANGUAGE MODELING FOR TURKISH SPEECH RECOGNITION

by

OSMAN BÜYÜK

Submitted to the Graduate School of Engineering and Natural Sciences
in partial fulfillment of
the requirements for the degree of
Master of Science

Sabancı University
August 2005

SUB-WORD LANGUAGE MODELING FOR TURKISH SPEECH RECOGNITION

by

OSMAN BÜYÜK

APPROVED BY:

Assist. Prof. Hakan Erdoğan …….
(Thesis Supervisor)

Prof. Kemal Oflazer …….

Assoc. Prof. Levent Arslan …….

DATE OF APPROVAL: …….

To my family

ACKNOWLEDGEMENTS

I would like to thank my supervisor Assist. Prof. Hakan Erdoğan for his guidance and advice throughout my thesis. He helped me in every step of my thesis and encouraged me to overcome the problems. I could not complete this difficult process without his advisement, enthusiasm and friendly support.

I would like to thank Prof. Kemal Oflazer for his help and participation to my thesis jury. His creative ideas have motivated me throughout my work and gave me inside knowledge.

I would like to thank Assoc. Prof. Levent Arslan for spending his valuable time and energy in my committee.

I also would like to thank my family for their endless love and patience. Their limitless tolerance made everything about me possible.

Not but not least I would like to thank my friends in Sabanci University for their support and enjoyment during the tedious laboratory hours.

ABSTRACT

In large vocabulary continuous speech recognition (LVCSR) for agglutinative languages, we encounter problems due to theoretically infinite full-word lexicon size. Sub-word lexicon units may be utilized to dramatically reduce the out-of-vocabulary rate in test data. One can develop language models based on sub-word units to perform LVCSR. However, it has not always been beneficial to use sub-word lexicon units, since shorter units have higher acoustic confusability among them and language model history is effectively shorter as compared to the history in full-word language models. To reduce the aforementioned problems, we propose using the longest possible sub-word units in our lexicon, namely half-words and full-words only. We also incorporate linguistic rules of half word combination into our statistical language model. The language constraints are represented with a rule-based WFSM, which can be combined with an N-gram language model to yield a better and smaller language model. We study the performance of the proposed system for Turkish LVCSR when the language constraint takes the form of enforcing vowel harmony between stems and endings. We also introduce novel error-rate metrics that are more appropriate than word-error-rate for agglutinative languages. Using half-words with a bi-gram model yields a reduction in word-error-rate as compared to a bi-gram full-word model. In addition, combining a tri-gram half-word language model with the vowel-harmony WFSM significantly improves the accuracy further when re-scoring the bi-gram lattices.

ÖZET

Türkçe gibi eklemeli dillerdeki geniş dağarcıklı konuşma tanıma uygulamalarında, kelimeler tanıma sisteminin birimi olarak seçildiğinde sınama için kullanılan kelimeleri kapsama ile ilgili sorunlar çıkmaktadır. Bu sorunu ortadan kaldırabilmek için kelime altı birimlerden yararlanılabilir. Geniş dağarcıklı konuşma tanıma uygulamasını gerçekleştirebilmek için kelime altı birimler kullanılarak bir dil modeli geliştirilebilir. Bununla beraber kelime altı birimlerin kısa olması nedeniyle yeteri kadar akustik bilgi içermemesi, birimler arasındaki akustik karışıklık olasılığını arttırmaktadır. Ayrıca kelime altı birimlerle elde edilen dil modelinde kelime dil modeline göre daha kısa bir geçmiş kullanılmaktadır. Bu sorunlar nedeniyle kelime altı birimlerin kullanımı ile sistemde her zaman beklenen başarım artışı sağlanamayabilmektedir. Bu problemleri ortadan kaldırabilmek için, bu çalışmada tanıma sözlüğünde kullanılabilecek en büyük kelime altı birim olan yarı-kelimelerin yada tam kelimelerin kullamı önerilmiştir. Buna ek olarak istatistiksel dil modeline yarı kelime birleşimlerindeki dilsel kısıtlamalar da dahil edilmiştir. Ağırlıklı sonlu durum makinesi ile ifade edilebilecek dilsel kısıtlamalar, daha küçük ve daha iyi bir dil modeli elde edebilmek için istatiksel modelleriyle birleştirelebilir. Bu çalışmada önerilen sistemin, ağırlıklı sonlu durum makinesi kelimelerin ekleri ve kökleri arasındaki ünlü uyumunu zorladığındaki başarımı ölçülmüştür. Türkçe gibi eklemeli dillerdeki hata oranlarını kelime hata oranına göre daha iyi gösterebilecek ölçü birimleri de teklif edilmiştir. Yarı-kelimelerle elde edilen ikili dil modeli, tam-kelimelerle elde edilen ikili dil modeline göre kelime hata oranları açısından daha iyi sonuçlar vermiştir. Buna ek olarak üçlü-dil modelinin ünlü uyumunu sağlayan ağırlıklı sonlu durum makinesi ile birleştirilmesi sonucunda elde edilen dil modeli, hata oranlarını önemli ölçüde azaltmıştır.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

LVCSR	Large Vocabulary Continuous Speech Recognition
ASR	Automatic speech recognition
HMM	Hidden Markov Model
OOV	Out of Vocabulary
FSM	Finite State Machine
WFSM	Weighted Finite State Machine
AM	Acoustic Model
LM	Language Model
SLM	Statistical Language Model
WER	Word Error Rate
STER	Stem Error Rate
HWER	Half-word Error Rate
W	Sentence
w_i	Word *i*
φ_i	Phoneme *i*
MFCC	Mel-Frequency Cepstral Coefficients
HTK	Hidden Markov Toolkit
pdf	Probability Density Function

1 INTRODUCTION

1.1 Motivation

Modern speech recognition systems use a limited vocabulary whose size is typically less than 100k. Hypothesis units of the recogniser are constructed from the entries of this vocabulary. Therefore, if a unit is not included in this vocabulary, it does not have a chance to be recognized. These units are called out-of-vocabulary (OOV) units. The percentage of OOV units in a typical test set, called OOV rate, plays a crucial role in the performance of modern speech recognisers.

Unconstrained speech recognition applications in agglutinative and highly inflectional languages suffer from OOV rate problem since any vocabulary size becomes inadequate when words are used as the vocabulary. As a result regular methods, which achieve acceptable performances in other languages such as English, fail for these languages.

Turkish is an agglutinative language in which new words can be generated by adding various suffixes to the end of stems. Morphological productivity of Turkish can be shown using the stem *ev*. In this example, morpheme boundaries are indicated with "–" symbol:

ev	house
ev-i	his house
ev-imiz	our house
ev-imiz-de	at our house
ev-imiz-de-ki-ler	the ones at our house
ev-imiz-de-ki-ler-den	from the ones at our house

Various new words can be obtained from a single stem by suffixation as in this example. This morphological productivity causes very large number of words, which can be uttered in Turkish. As a result we may expect a steady vocabulary growth rate without any saturation if words are chosen as vocabulary units of the speech recogniser. This fast and boundless vocabulary growth rate results in high OOV rates for any vocabulary size.

Although number of words in Turkish is theoretically infinite, sub-word units such as morphemes and syllables can be finitely listed. As a result, high out-of-vocabulary problem seen with word units can be efficiently solved with the utilization of sub-word units. In past studies, syllable, morpheme and stem-ending based vocabularies have been proposed for Turkish speech recognition [1, 2, 10, 11, 14]. The recognition performance of automatic speech recognition systems based on these vocabulary units has been compared to word-based systems and each other.

There has been much less effort to implement a speech recogniser for Turkish compared to other languages such as English. This is due to the inherent challenges of Turkish speech recognition task mentioned above. In this thesis, we focus on challenges in Turkish speech recognition. We have utilised sub-word units such as syllables, morphemes or hybrid units to avoid OOV rate problem. We measured the performance of recogniser based on these different units. The performances of experiments are compared to each other in order to suggest the best vocabulary unit for unconstrained Turkish speech recognition.

1.2 Problem Definition

Speech recognition problem consists of two main modelling parts: language and acoustic modelling. The system learns the properties of particular language in language modelling. On the other hand acoustic features are dynamically modelled in acoustic modelling. A speech recogniser utilises these two models together.

Hidden Markov Models are used in modern speech recognition systems for acoustic modelling. Statistical or rule-based language models can be used in order to represent the recognized language. If the speech recognition application permits the use of a limited-vocabulary handcrafted grammar, then high recognition accuracies may be achieved. In such applications, when the speaker utters a sentence, not covered by the grammar, the system cannot recognize the sentence. In a system-initiative spoken dialog system, if the user can be directed with well-prepared questions, a dynamic limited grammar system may achieve great performance. However, for mixed-initiative systems and for other applications such as dictation and broadcast news transcription, a large vocabulary system with a statistical grammar (language model) is required. Such systems are called large vocabulary continuous speech recognition (LVCSR) systems.

N-gram language models for LVCSR achieve acceptable performance for English. In English dictation systems an accuracy of 90-95% may be achieved. On the other hand, in agglutinative and highly inflectional languages like Turkish, Finnish, Czech, etc., LVCSR is problematic when only the words in the language are used as lexicon units. Total number of words is very high and any vocabulary size becomes inadequate [1-15]. A lexicon that contains sub-word units may be utilized as a solution to the coverage problem [1-16]. These sub-word units are typically morphological components of words or syllables in the language.

Despite solving the coverage problem, using sub-word units in LVCSR for agglutinative languages does not always achieve acceptable performance [1, 2, 3, 5,15]. A reason for bad performance is the shortness of sub-word units as compared to full-words. As expected, shorter units become more acoustically confusable with each other. Also, LVCSR decoders have a tendency to insert short units into wherever they can since matching short units to acoustic data is easier. Another reason for the performance drop in sub-word systems is the smaller effective language model history while using N-gram models. In a sub-word system, the language model effectively uses a shorter history as compared to a full-word based system. The shorter the sub-words, the shorter the effective history one uses in an N-gram language model. As a result, choosing the appropriate sub-word unit is crucial for LVCSR systems. The main criteria in this choice should be to cover the words in the language as much as possible while maintaining small acoustic confusability between units not to decrease the recognition rates. In other words, unnecessary and short units should not be included in the lexicon of the speech recogniser. In this paper, we propose some new ideas to construct a lexicon of half-word and full-word units for a LVCSR system in Turkish.

Another potential problem while using sub-word units in speech recognition is that, the recogniser may output a sequence of sub-words that may not form a legitimate word sequence in the language. In this thesis, we propose to use a rule-based weighted finite state machine (WFSM) to accept only allowable sub-word unit sequences. This WFSM can be composed with an N-gram language model to improve the overall language model and reduce its size. With this approach not only language model size can be decreased but also recognition rates can be improved.

1.3 Literature Review

LVCSR for agglutinative and highly inflectional languages has drawn the attention of researchers recently. The challenges of LVCSR for such languages and different techniques to overcome these challenges have been illustrated in these studies. We review past studies on LVCSR for agglutinative and highly inflectional languages in the following paragraphs.

Carki's work [2] is one of the first attempts for LVCSR in Turkish. 16.9% WER is reported as the best recognition performance when OOV problem is ignored by adding all test words to the vocabulary. To overcome the OOV problem, morpheme-based approach by merging syllable units into larger units by defining word-positioned syllable classes studied. With morpheme-based approach significant OOV rate reduction is obtained, but the recognition rate is unsatisfactory. It is mentioned that, 27% relative improvement in OOV rate could be achieved with hypothesis driven lexical adaptation, but no studies were conducted on this method.

In [17], a speech recogniser is designed for Turkish Radiology Dictation Application. In this task-specific system, vocabulary size is limited to a few thousand words. In addition there is a systematic arrangement of words in sentence formation. As a result, OOV rate and perplexity is very low with a moderate vocabulary size especially when compared to general Turkish LVCSR. High recognition accuracies (approximately 82%) are obtained in this limited task. In addition, pronunciation variants are investigated. The pronunciation variants are added to lexicon by learning common mis-recognized words during speech recognition process. 24.74% relative error rate reduction is observed, when proposed data-driven pronunciation variant approach is introduced to the recogniser.

Comparison of morpheme-based, stem-ending-based, syllable-based and word-based language models for LVCSR of Turkish is carried out in Dutagaci's work [14]. For all three models, a significant improvement is obtained in terms of vocabulary size, coverage, perplexity and sensitivity to context, compared to word-based model. While the smallest perplexity, vocabulary size and OOV rate are obtained for syllable experiment, it is also noted that this base unit less is sensitive to context changes. In addition, morpheme-based model yields lower perplexity compared to word-based and stem-ending-based models. Morpheme-based model is also claimed to be less sensitive to context changes as compared to word and stem-ending model. In this paper, all

evaluations are performed on text data and no speech recognition experiment is performed.

In [1], morphological parts of words are used in order to solve the coverage problem seen in Turkish LVCSR. In this work, a combination of word, morpheme, and stem-ending based models is obtained with respect to most frequent words in training corpus. While most frequent N-words are left as stems, most frequent N to 2N words are parsed as stem + endings and remaining words are parsed to their morphological parts. Although coverage and perplexity of combined model is better than word based model, speech recognition performance is not improved with the combined model due to smaller recognition units. In addition this study suffers from insufficient text corpus available for accurate estimates of bi-gram models.

In [18], recent work to develop text and audio corpora and speech recognition tools for Turkish LVCSR is described. The authors try to overcome one of the biggest problems, which is the lack of phonetically balanced rich databases for Turkish speech recognition research. A phone error rate of 29.3% is demonstrated using a back-off tri-gram phoneme language model.

In [10], data-driven and morphology based methods to split words in Turkish into their morpheme-like units are compared in terms of language modelling and speech recognition performance. The data-driven method is based on the proposed method in [19]. This method discovers morpheme-like units in an unsupervised manner. Expensive expert labour needed to create a morphological analyser can be saved by this method. Morphology-based method is based on the two-level morphology of Turkish [20]. The best results are obtained using data-driven method with 54.8% WER despite its higher perplexity, lower LM quality and higher OOV rate as compared to morphology-based method. In addition recognition accuracy is improved in [11], when the counts of units to be split are introduced to the data-driven method. A WER of 46.6% is achieved with this extension. Moreover compounding of lexical units based on data driven methods, provides further recognition gain. WER is decreased from 54.8% to 52% with the compounding procedure in two iterations.

In [19], two methods for unsupervised segmentation of words into morphemes are proposed. These methods are based on Minimum Description Length (MDL) and Maximum Likelihood (ML) optimisation. Both methods achieve superior performances for Finnish compared to a previous method, called Linguistica. However the methods have similar performances to Linguistica in English. In addition MDL method performs

consistently better than ML method. In [21], excessive segmentation problem is overcome using prior information about morph length and morph frequency. This new method outperforms MDL method especially for Finnish data.

Finnish LVCSR based on the sub-word units obtained with MDL method in [19], is studied in [12]. The results based on syllable units are also given. With the usage of sub-word units like syllables and morpheme-like units, OOV rates decrease significantly compared to word-based model. In addition both models outperform word model in terms of recognition accuracies. Compared to the word tri-gram model, the OOV rate is reduced from 20% to 0% and WER from 56% to 32% in morph based model. In order to have a fair comparison between the accuracies of different experiments and different languages, three error rates are computed: token error rate, word error rate and letter error rate. In [13], more elaborate comparison of recognition accuracy is made between three different base units. Compared units are sub-word units obtained with MDL, grammatical morph-based and word-based units. In word-based experiment phonemes are also added to lexicon to overcome OOV rate problem. In grammatical morph-based experiment, segmentation of words into morphological parts is obtained using two-level morphology [22]. In comparison, both morphology-based experiments outperform word-based experiment significantly in terms of recognition accuracies. As a result grammatical and statistical morpheme-like word segmentation methods were found to be good choices to represent a very large vocabulary efficiently with a moderate number of lexicon units. On the other hand in statistical method, no expert labour is needed since the morphs are found in an unsupervised and language independent manner.

In [23], different SLMs are tried for recognition of broadcast news in Finnish language. The SLMs try to model the language with a limited amount of statistical information of the preceding words. The idea is related to statistical modelling of word inflections, word contexts and document contexts. The best results are obtained by the document context SLMs.

In [6], a LVCSR system is built for an inflected language, Korean. The speech recognition task for this language also suffers from unmanageably large dictionaries with extremely high OOV rates when aejeol (word phrase) is used as lexicon entries. Syllable-based system is investigated in this study. A data-driven approach to repeatedly merge syllable units in order to decrease acoustic confusability is also studied. When acoustic information in syllable sequences is increased with merged

syllables, the performance of the recogniser increases from the base result 74.3% to 79.4%. In this paper, 95% syllable lattice accuracy is reported when syllable system is improved with phone set reduction and a new phone context question group. In [7], morpheme-based and syllable-based approaches are investigated for Korean language. Frequent morphemes are merged using rule-based and statistical unit merging methods. Best performance is obtained with statistical merging methods when appropriate linguistic constraints are used. In this research it was also noted that syllable-based experiment does not give comparable performance to morpheme-base experiment although syllable experiment has the advantage that the OOV rate is nearly zero and can be used in unlimited vocabulary speech recognition tasks. In [8], pseudo-morpheme (morpheme-like) units are used as lexicon units of Korean LVCSR. Inter-morpheme acoustic modelling and pronunciation dependent language models are proposed. 80% pseudo-morpheme recognition accuracy is achieved with pseudo-morpheme base units in the best case. This best result is much higher than the baseline pseudo-morpheme accuracy, which is 75%. In [9], further improvements are obtained when text corpus size and acoustic parameters (number of codebooks, senones, mixtures in a codebook) are increased. 84.5% pseudo-morpheme recognition accuracy which is equivalent to word recognition accuracy in English is achieved with these modifications.

In [16], different decomposition methods originally based on morphological decomposition of words in German are used in speech recognition and various properties of morpheme-based units such as coverage, perplexity and speech recognition performance are compared. It is observed that smaller lexicon units lead to reduction in perplexity and OOV rate. In addition vocabulary growth is much slower in morpheme-based experiments. On the other hand, these improvements in language modelling do not improve word accuracies although speed of recogniser is accelerated by approximately one third because of smaller lexicon size.

In [3], partial morphological analysis is used in language modelling of LVCSR in a highly inflectional language, Portuguese. With this approach, perplexity, memory requirements and the OOV rate are reduced. Processing time spent by the recogniser is also reduced with morpheme-based approach. However, the recognition accuracy is degraded slightly compared to word-based model. The author claims that, this degradation is due to the shortness of many morphemes and invalid morpheme sequences as no linguistic constraints are introduced to the recogniser.

In [4], a stem-ending model is proposed for LVCSR of highly inflected language Slovenian and this model is compared to word-based model. This model is preferred to full morpheme-based model since breaking the acoustic information into too small parts can be detrimental. This model gives improved performance (7.5% absolute) for small vocabulary size, e.g. 20k. On the other hand, acoustic confusability problem between small sub-word units becomes more evident when vocabulary size is increased. As a result, word-based model gives better performance with large vocabulary sizes, e.g. 95k and 60k. In [5] similar recognition results are obtained for the same language with morpheme and word based models. Significant improvement in perplexity results is also observed when corpus-based topic-adapted language models are used.

In [24], LVCSR system is developed for dictation and broadcast news domain in Serbo-Croatian language. The positive effects of vocal tract length normalization and MLLR adaptation are observed in this study. A surprising recognition performance of 36% WER is reported.

In [15], tri-gram re-scoring is applied to N-best list for both morpheme-based and word-based experiments in Czech language. Significant improvements are obtained in both experiments after re-scoring. In addition morpheme-based and word-based experiments give comparable result although OOV rate of morpheme-based experiment is significantly small.

In Geutner's work [25], some language constraints are claimed to be incorporated to a statistical language model. In this work, words in German are divided into two categories: function and content words. In addition to a regular N-gram model that uses previous words in the history, another language model that uses only previous "function words" is also trained. Then these two language models are interpolated. In this study, slightly better performance is observed in terms of perplexity. In addition, the recognition accuracy of speech recogniser is increased to 71% from baseline result 70.6%. In fact, the idea to integrate linguistic constraints to a language model by categorizing the type of history words is first introduced by Isotani for Japanese [26]. In this work, a new language model with reduced number of parameters is obtained after linguistic constraints are integrated to the language model. In addition the performance of the recogniser does not degrade much compared to the other conventional bi-gram and tri-gram language models although the new language model is computationally more efficient than the tri-gram model.

In [27], construction procedure of a new language is explained to develop robust human-machine speech recognition applications. The words of this new language are chosen so that acoustic similarities between these words are minimized. New constructed language is compared to Turkish and English using a digit recognition experiment and it is observed that significant accuracy improvement can be achieved when acoustic similarities between lexicon units of a language are minimum.

1.4 Contributions

This thesis contributes to the recent studies conducted in the research area of LVCSR for agglutinative languages. Utilization of sub-word units is proposed in order to solve the coverage problem, which arises when words are used as the vocabulary of the recogniser. Syllable-based, morpheme-based and hybrid-based recognition lexicons are compared with a word-based lexicon in terms of both language modelling and speech recognition performance. One of these lexicons is chosen as the most promising recognition lexicon for Turkish LVCSR. In addition linguistic constraints such as vowel harmony rule in Turkish are studied. It has been observed that recognition accuracy can be improved when linguistic rules are also incorporated into the language model with a WFSM. One of the most important contributions of this thesis to the past research efforts, is to benefit from the linguistic knowledge available in the specified language for speech recognition performance. This WFSM also forces the correct order of sub-word units. When recogniser lexicon is constructed from sub-word units, illegitimate order of sub-words is one of the sources of recognition errors. In addition, new evaluation metrics, which better reflect the performance of recogniser for agglutinative languages, are introduced. These evaluation metrics are stem error rates and half-word error rates.

The contributions of this research can be summarised with the following items:

- A good way to obtain a sub-word lexicon is proposed in order to solve the OOV rate problem
- Linguistic constraints between sub-word units in the particular language is enforced
- Correct sub-word order is enforced
- New evaluation metrics are introduced for LVCSR in agglutinative languages

1.5 Outline of thesis

This thesis is organised in the following way. In chapter 2, an overview of speech recognition is presented. In chapter 3, important properties of Turkish language in terms of speech recognition task are discussed. Our implementation for Turkish LVCSR is explained in chapter 4. Chapter 5 provides the properties of training text and speech corpora, which are very crucial to obtain a robust speech recognition system. This chapter also describes the details of language and acoustic model training. In chapter 6, the experimental set-up and results obtained for different recognition units are provided. Improvements in speech recognition performance with rule-base WFSM are illustrated in chapter 7. The conclusions and future plans are summarized in the final chapter.

2 SPEECH RECOGNITION

2.1 Overview and usage areas of speech recognition

A speech recognition system can be defined as a piece of software or a device, which transcribes human speech to written text. Human-machine interaction would become much easier if such a device is realised for most of the languages in the world. Moreover with a LVCSR system, any spoken utterances in particular language would be recognized accurately. This will let people interact with machines using the natural medium of speech. Nevertheless implementing such a speech recogniser is not an easy task and serious considerations must be taken. In addition, this task shows different characteristics for different languages. Because of that, individual efforts must be expended for each language to get a global speech recogniser. In the following paragraphs, we review daily life applications of speech recognisers. While some of the mentioned speech recognition applications are already implemented accurately for some languages, some of them are not done yet.

Speech recognisers can be used in Interactive Voice Response (IVR) systems. Most companies use call centers in order to meet the needs of their customers. For this purpose, they hire a lot of call center employees, which is very expensive. If a speech recognition system can be implemented to communicate with the users of an IVR system, the need for employees will decrease and the work done with a lot of employees will be done with a speech recognizer. This will significantly reduce the labor and the costs of these systems.

Robust dictation systems would be used in daily life with the availability of accurate speech recognizers. When a LVCSR system built inside of an editor program for Turkish, a document would be typed with human voice instead of keyboard. This will reduce the amount of time spent for typing documents and will make this work much easier. In fact any human speech would be instantaneously transcribed to written text with a recognizer. Such a dictation system can be very useful to transcribe audio of broadcast programs to written text for deaf people or recordings of employees often eyes and hands-busy at work (such as radiologists [17]).

Another important utility of speech recognition systems is to let people command machines with their voices. For example television can be controlled with human voice instead of remote control. With this technology, one can change the channel by saying its name or turn the television off by a predefined command. Any other household appliance can be controlled with human voice as television. Moreover other technological tools like the electronic tools in cars, robots, computers etc… can be governed with voice commands without a need for any other intermediate tools.

When a speech recognition system is available for most of the languages in the world, it can also be used for translation purposes after it is incorporated with a speech synthesizer and a machine translator. If the available speech recognition system is a LVCSR system, any spoken utterances can be translated to other languages without the need of human translators. For example this type of system can be used in international conferences in which instantaneous translation from various languages are needed. People would be able to communicate with each other with the help of a LVCSR system even if they don't speak each other's language. These systems would reduce the need for learning foreign languages and would let people communicate with others who speak a different language if a speech synthesizer is available in the aimed language.

There are much more places where speech recognizers can be very useful. They dramatically reduce the amount of labor, money and time spent by people in a lot of the areas. In order to start using this technology in daily life, more intense efforts are needed especially for languages like Turkish. This research is one of the initial attempts to obtain an accurate and robust speech recognition system in Turkish.

2.2 Theory of Speech Recognition

A speech production and recognition system can be best explained with the source-channel model of speech recognition in Figure 2-1 [28]. This figure also shows the resemblance of speech recognition process to communication theory:

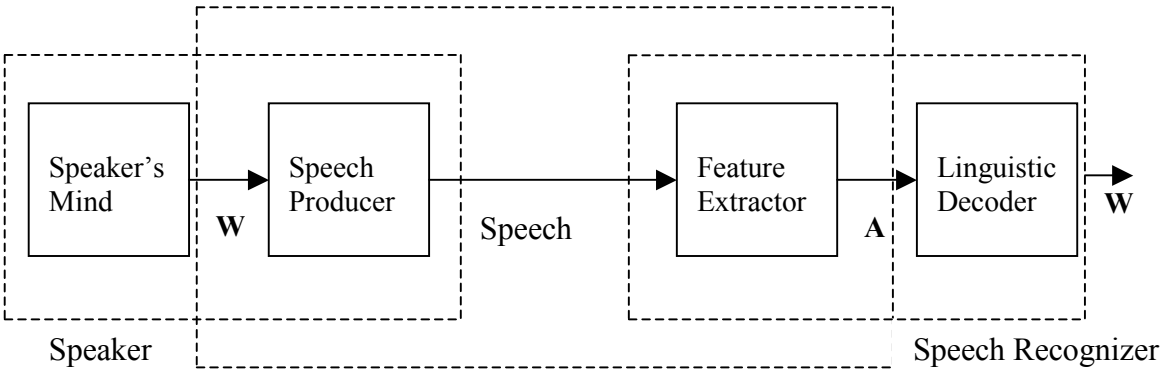

Figure 2-1 Source-channel model of speech recognition

The source of this communication channel is the speaker's mind, which decides on the words that will be pronounced. Then the speech is produced by a complex system involving the vocal tract of the speaker. In the receive side, this speech waveform is transformed to an acoustic feature sequence by the feature extractor. This part and the speech producer constitute the noisy channel of the communication system and it is labelled as the acoustic channel in Figure 2-1. The acoustic feature sequence is mapped to a hypothesized word sequence in the linguistic decoder. As seen in Figure 2-1, a speech recogniser consists of a linguistic decoder and a feature extractor.

Linguistic decoder consists of three important components: acoustic and language models and hypothesis search algorithm (or decoding algorithm). Process in this decoder can be described with the following figure:

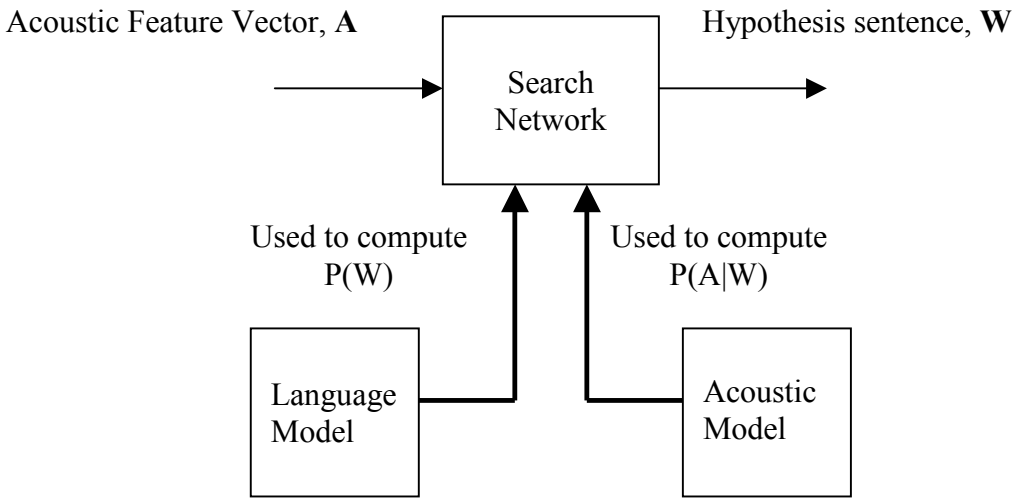

Figure 2-2 Linguistic Decoder

Figure 2-2 will be discussed in the following sections of this thesis. Search network will be shown in section 2.6 with more details. Acoustic and language models will be explained in sections 2.5 and 2.4 respectively.

As seen in Figure 2-2, a speech recogniser converts a speech waveform to a sequence of words. To obtain the optimal transcription of speech waveform, decoder tries to find the best path in the search network using probabilities obtained from acoustic and language models. This search process is called hypothesis search or decoding. In order to run this search, first speech waveforms must be transformed into acoustic feature vectors. This part is called feature extraction or front-end processing and will be described in section 2.3.

If we assume that the hypothesis sentence is constructed from words, the decoding process can be described with the following formula:

$$\hat{W} = w_1 w_2 ... w_n = \arg\max_W P(W \mid A) \tag{2-1}$$

In this equation, $\hat{W}$ represents the hypothesis sentence and w_i represents the words in this sentence. Using Bayes rule this formula can be written as:

$$\hat{W} = \arg\max_W P(W \mid A) = \arg\max_W \frac{P(A \mid W)P(W)}{P(A)} = \arg\max_W P(A \mid W)P)(W) \tag{2-2}$$

In this formula, $P(A \mid W)$ denotes the probability of observing the acoustic sequence when the word sequence is given. This probability is obtained from the statistical acoustic model. On the other hand, $P(W)$ is the a priori probability of the word string, W. This probability is calculated using the language model. Then best matching word sequence for a given speech signal is calculated with the combination of these models inside the search network.

2.3 Feature Extractor

A speech waveform must be transformed to a sequence of parameter vectors before being used in speech recognition process. This part is called feature extraction or front-end processing as shown in Figure 2-1. Short-time spectral analysis performed to extract the features. The speech signal is divided into overlapping frames in time and each frame is handled separately to extract the feature vectors.

There are many different feature extraction methods for speech recognition. Filter-bank analysis is a popular approach since it gives non-linear frequency resolution

as in the case of human ear voice perception. Tri-angular filters, which are equally spaced in mel-scale, are used in HTK (this toolkit will be discussed in Section 2.7) [29]. Mel-scale can be defined as:

$$Mel(f) = 2595 * \log(1 + \frac{f}{700}) \tag{2-3}$$

To implement the filter-bank, magnitude of Fourier Transform of the windowed speech signal is multiplied by the corresponding filter gain and results are accumulated. Thus, each accumulation holds a weighted sum representing the spectral magnitude in that filter-bank channel [29]. The overall paradigm can be summarised with the figure below:

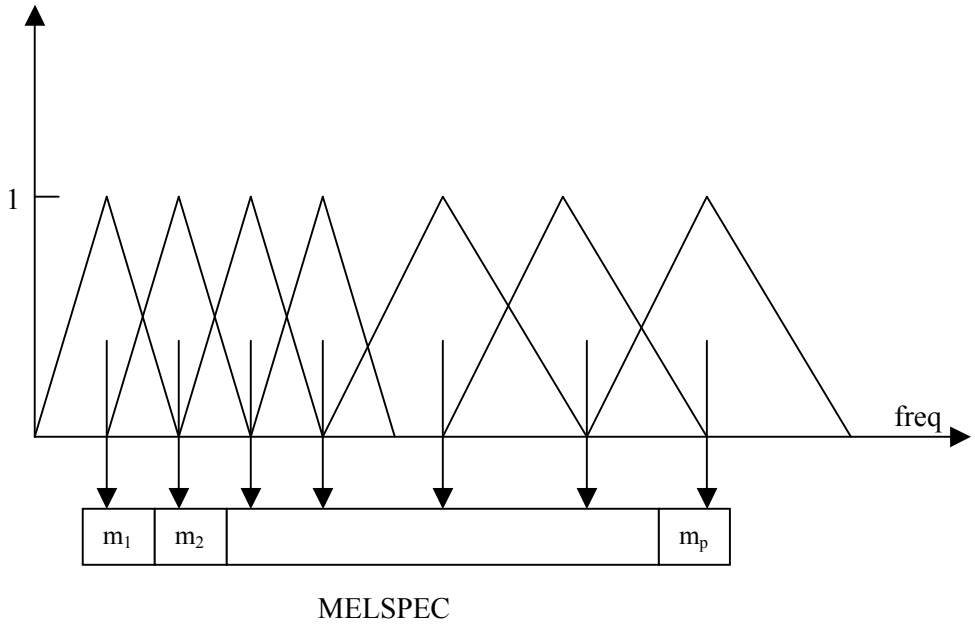

Figure 2-3 Mel-Scale Filter Bank

Most often cepstral features are used which are called Mel-Frequency cepstral coefficients (MFCC) in speech recognition. These parameters are calculated from the log filter-bank amplitudes $\{m_j\}$ using discrete cosine transform (DCT) and keeping the lower order terms:

$$c_i = \sqrt{\frac{2}{N}} \sum_{j=1}^{N} \cos\{m_j (\frac{\pi * i}{N}(j - 0.5))\} \quad \text{i} = 1,...,\text{M} \tag{2-4}$$

Here N is the number of filter-bank channels, and M is the number of DCT coefficients kept. Typically M=12. An energy term is appended to the cepstral

parameters and the resultant parameters called static parameters. In addition, time derivatives are added to static parameters. First order regression coefficients (referred to as delta coefficients), second order regression coefficients (referred to as acceleration coefficients) are appended to parameter vectors to obtain 39 dimensional feature vectors.

In this research, 39 dimensional MFCC features have been used as acoustic features to represent speech waveforms. We also perform cepstral mean normalization and energy normalization on the feature parameters [29].

2.4 Language Model

2.4.1 Grammars

Statistical or handcrafted grammars can be used in order to represent the recognized language. Handcrafted grammars achieve acceptable performance if the set of allowed sentences in the subset of language we want to recognize is relatively small. In a system-initiative spoken dialog system, if the user can be directed with well-prepared questions, a dynamic limited grammar may be successfully used and achieves great performance. Following example is a simple handcrafted grammar, which provides a voice-operated interface for phone dialling:

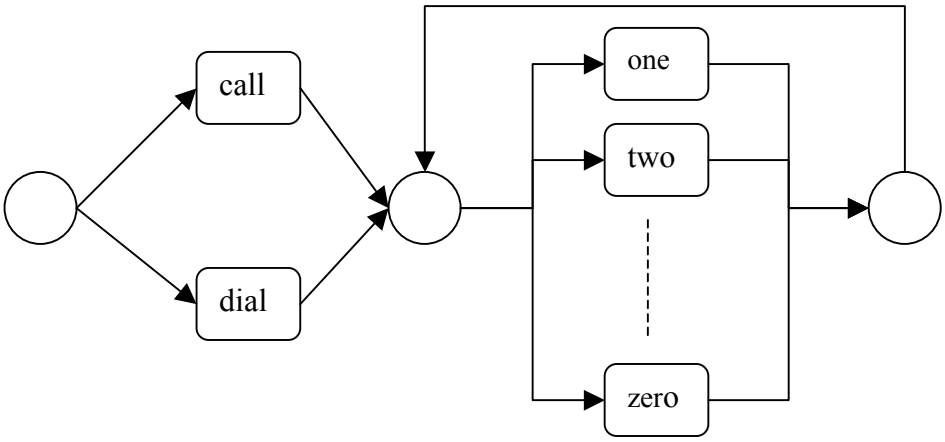

Figure 2-4 Handcrafted grammar for phone dialing

This grammar can be used in a phone dialling recognition application in which user should say the word “dial” or “call” first and then he/she should start saying the phone number. However these types of handcrafted grammars suffer from uncovered parts of language. When a user utters a sentence, which is not covered by the grammar,

the system has a zero probability to recognize it. For example if the user of phone dialling recognition application starts his sentence with the word “phone” instead of “dial” or “call”, recogniser fails in this uncovered sentence. If we want to recognize most part of language as in LVCSR case, these limited, handcrafted grammars cannot be used because of their restriction. In LVCSR applications statistical language models like N-grams are successfully used and they achieve acceptable performance.

N-gram models and most handcrafted grammars are regular grammars. Therefore they can be realised with weighted finite state machines (WFSM) [30]. A finite state machine can be defined as a model of computation, which consists of a set of internal states, an input alphabet, a transition function and an initial state. Transition function maps input symbols and current states to a next state with the arcs of finite state machine. One can add weights to these arcs to handle probabilistic transitions. A finite state machine that contains weights on its arcs is called weighted finite state machine. Computation begins in the initial state with an input string and it changes to new states depending on the transition function.

Language models provide word string probabilities to the recogniser as stated in section 2.2. In this section we will observe that language models could be represented with WFSMs. Acoustic models are typically HMMs. As shown in Figure 2-2, the final part of the recogniser, namely search network, consists of these two modelling parts. The search algorithm, which tries to find the best matching path for the observed acoustic feature vector, conducts its search on a large-HMM whose model parameters are constructed from the probabilities obtained using language and acoustic models. The detailed figures for this large-search HMM will be given in section 2.6, after acoustic modelling is also introduced.

2.4.2 Statistical (N-gram) Language Models

A statistical language model enables to compute the probability of a sentence in a language. If we assume that a sentence is constructed from words, the sentence probability can be written as follows:

$$P(W) = P(w_1, w_2, ..., w_n) = \prod_{i=1}^{n} P(w_i \mid w_1, ..., w_{i-1}) \tag{2-5}$$

In this equation W represents the sentence and each w_i represents a word in the sentence. N-gram language models approximate each term in the right hand side of the equation by using only N-1 words in the history. The equation becomes:

$$P(W) \approx \prod_{i=1}^{n} P(w_i \mid w_{i-N+1},...,w_{i-1}) \tag{2-6}$$

Usually, N-gram probabilities are calculated with a technique known as maximum likelihood estimate from a large text corpus. In order to train N-gram models a counting and normalizing procedure is applied [31]. From the training text corpus counts of a particular N-gram is divided by sum of all N-gram that share the same *N-1* words [31]. This technique can be formulated in the following way:

$$P(w_i \mid w_{i-N+1},...,w_{i-1}) = \frac{C(w_{i-N+1},...,w_{i-1},w_i)}{C(w_{i-N+1},...,w_{i-1})} \tag{2-7}$$

In this equation $C(w_{i-N+1},...,w_{i-1},w_i)$ represents the count in the text corpus for the specified consecutive N words and $C(w_{i-N+1},...,w_{i-1})$ represents the count for specified *N-1* words. As a result N-gram probability is estimated by dividing the observed frequency of a particular sequence by the observed frequency of a prefix. The ratio in this equation is called relative frequency.

As mentioned in 2.4.1, N-gram language models can be realized with WFSM. AT&T's GRMTOOLS is a package for building and manipulating N-gram language models with WFSM [32]. In order to obtain tri-gram language model and rule-based WFSM in chapter 7, we benefit from this toolkit.

2.4.3 Smoothing

We estimate N-gram probabilities from the counts in a training text corpus, these estimates become more robust and reliable if a large training text corpus is available. Although used text corpus is very large, some N-grams may not be seen in the corpus or may be poorly estimated because of small counts. In these cases, smoothing or discounting techniques are used to avoid assigning zero probabilities for unseen N-grams. Essentially in smoothing, a small part of the available probability mass is deducted from the higher N-gram counts and distributed amongst the lower N-gram counts. There are different smoothing techniques and these techniques are used to obtain more reliable language model.

Backed-off language modelling is one of the strategies to overcome the sparseness of N-gram counts. In this smoothing technique, we build N-gram models based on (N-1)-gram models if the count of an N-gram falls below a chosen threshold. For example in bi-gram case, when a bi-gram count falls below the threshold *t*, the bi-gram is

backed-off to uni-gram probability suitably scaled by a back-off weight in order to ensure that all bi-gram probabilities for a given history sum to one [29]. When this technique is used for a bi-gram language model, bi-gram probability estimates can be written as:

$$P(i,j) = \begin{cases} \dfrac{N(i,j) - D}{N(i)} & if \quad N(i,j) > t \\ b(i)\,p(j) & otherwise \end{cases} \tag{2-8}$$

In this equation $N(i, j)$ is the number of times word j follows word i and $N(i)$ is the number of times that word i appears [29]. D is discounting constant and t is the threshold value mentioned above. $p(j)$ is calculated from the uni-gram counts of word j and $b(i)$ is the back-off weight for word i. As we see in this equation, a small part of available probability mass is deducted from frequent bi-grams and distributed among the rare bi-grams.

2.4.4 Evaluation of N-gram LMs

There are some techniques to evaluate the goodness of a statistical language model. One of these evaluation metrics is the N-gram hit percentage. This evaluation metric shows the percentage of N-grams in a test corpus included in trained language model. The bigger this number is the better the language model.

The second evaluation criterion is perplexity. Perplexity of a random variable is calculated from the entropy of the process. Entropy can be defined for a random variable that ranges over X:

$$H(x) = -\sum_{x \in X} p(x) \log_2 p(x) \tag{2-9}$$

Then the value 2^H is called perplexity. Perplexity can be intuitively thought as the weighted average number of choices a random variable has to make.

In order to compare different probabilistic models cross-entropy is used. Cross-entropy is useful when the actual probability distribution p that generated some data is not known and estimated using a model for it, m. The cross-entropy of model m on p for a stationary ergodic process can be written as:

$$H(p,m) = \lim_{n \to \infty} -\frac{1}{n} \log m(w_1 w_2 ... w_n) \tag{2-10}$$

The useful side of cross-entropy calculation is to give an upper bound for entropy of actual random process:

$$H(p) \leq H(p,m) \tag{2-11}$$

From here we can conclude that the better model m is the smaller its cross-entropy and perplexity. Between two models m_1 and m_2 the more accurate model will have smaller cross entropy [31].

In order to calculate the perplexity or cross-entropy of N-gram language models, a test corpus, which is not included in training of the language model, is needed. Then, calculated perplexity or cross-entropy gives the capability of the language model to represent the test corpus. A better language model for the test corpus must have lower perplexity or cross-entropy. From this calculation, we can choose the best SLM to model a test corpus among different SLMs.

2.4.5 Important characteristics of N-gram LM

2.4.5.1 Lack of hard constraints

There are some properties of statistical language models, which are important for a speech recognition system. To obtain a reliable language model, very huge amount of text corpus (million even billion of words) is needed. Despite the usage of huge training data, some infrequent parts of language may be never seen. As a result, a small but nonzero probability is given to illegitimate sequences of words in smoothing. In addition training text corpus should be cleaned, tokenised and checked for typos before being used in N-gram language model training. Since this cleaning procedure is usually not perfect, some noise is always left in the training data and this noise causes imperfect weights in the WFSM. Moreover the noise may introduce nonzero probabilities for illegitimate word sequences.

Linguistic rules in a language may not be learnt with a statistical language model since considering only word order of a given training text might not be enough to capture linguistic constraints typical for a particular language [25]. For example vowel harmony rule between the stem and endings of words is mostly obeyed in Turkish. An N-gram language model can be trained using stems and endings. Although this N-gram language model sees a lot of occurrences of this rule, it does not completely rule out cases where this rule is not obeyed.

On the other hand, enforcement of linguistic rules and legitimate word order may improve the recognition performance of the recognizer. In this research, we will realize hard constraints with WFSMs to enforce correct word order and linguistic rules at the

output of decoder. When these types of hard, rule-based language models are composed with soft, statistical language models, better language models can be obtained which not only calculate the probability of a word sequence statistically but also check whether a sequence obeys the constraints introduced by rule-based WFSM. This language model may improve the performance of the recognizer significantly for languages like Turkish, which are regular in terms of obeying the rules.

2.4.5.2 Brittleness across domains

Another important property or weakness of statistical language models is their brittleness across domains. Current language models are extremely sensitive to changes in the style, topic, or genre of the text on which they are trained. For example in order to model phone conversations it is better to use 2 million words of transcription from telephone dialogues compared to 140 million words of transcript from TV or radio news broadcast. In addition, a language model trained on Dow–Jones newswire text will see its perplexity doubled when applied to the very similar Associated Press newswire text from the same time period [33]. Because of this sensitivity, language models trained with a text corpus, whose domain is not similar to the domain of test data, may not be successful to model test data. Therefore expected high recognition performance in speech recognition may not be achieved although a huge amount of training text corpus is used during the training of SLM.

2.5 Acoustic Model

2.5.1 Hidden Markov Models

Hidden Markov Models are used in modern speech recognition systems for acoustic training. HMM is a natural and highly reliable way of recognizing speech for a variety of applications [34].

In LVCSR application as the vocabulary size is huge, HMMs are used to model smaller acoustic units rather than words. The smallest unit that conveys acoustic information is called phoneme. Each phoneme is modeled with 3 HMM states representing begin, middle and end of phonemes. The model can be depicted as follows:

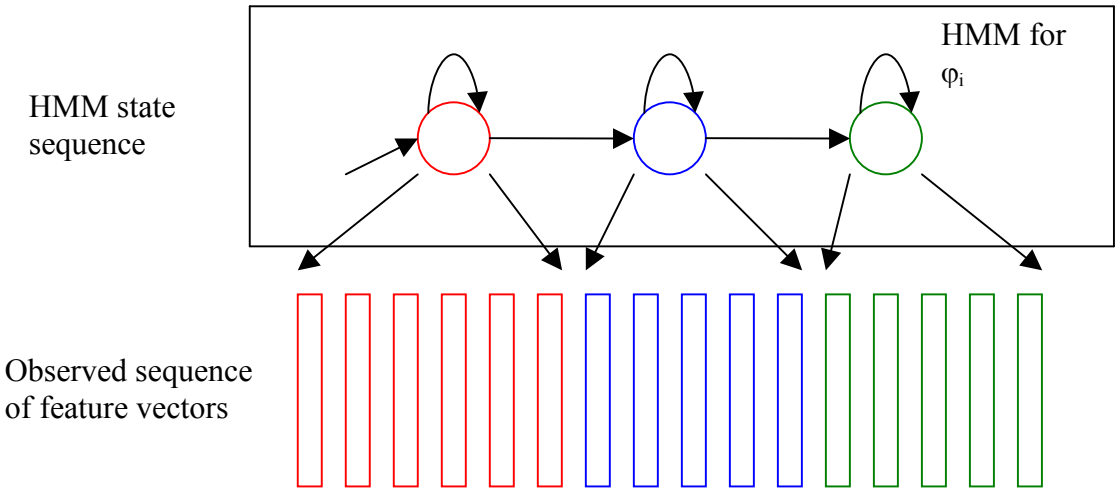

Figure 2-5 HMM for phonemes and aligned feature vectors

In this model the observation sequence is assumed to be produced by underlying hidden states at each instant of time. The model can be thought as a generalized WFSM which changes state once every time unit and at each time *t* that a state *j* is entered an output vector is observed with the continuous probability density $b_j(o)$. Each state of HMM is usually represented with a number of Gaussian mixtures. If state *j* is modeled with *M* Gaussian mixture components, $b_j(o)$ can be written as:

$$b_j(o) = \sum_{k=1}^{M} c_{jk} N(o, \mu_{jk}, U_{jk}) \qquad (2\text{-}12)$$

In Equation 2-12, *o* is the observed sequence of feature vectors, c_{jk} is the mixture coefficient for the k^{th} mixture in state j and *N* is a Gaussian pdf with mean μ_{jk} and covariance matrix U_{jk}. Furthermore, the transition from state *I* to state *j* is also probabilistic and is governed by the discrete probability a_{ij} [34].

Model parameters of an HMM, namely state transition probabilities (a_{ij}), state observation densities ($b_j(o)$) and initial state distributions (π_I) can be calculated using a large amount of speech data by maximizing the probability of acoustic feature vectors by varying the model parameters. In order to estimate these model parameters, first the feature vectors of speech waveform must be obtained as explained in section 2.3 and then these feature vectors must be aligned to states of HMM as shown in Figure 2-5. For the reliable estimate of these parameters, a large amount of speech data is required. When more training speech data is available, more robust acoustic model parameters

can be obtained. Our implementation for estimation or training part will be discussed in section 4.1.

Acoustic models for phonemes or tri-phones are combined to make up HMMs for the words. The word models are then inserted into the search network as indicated in section 2.2. We previously stated that large-HMM search network constitutes probabilities obtained from language and acoustic models. The detailed figures for this large-HMM will be given in section 2.6 for different statistical language models.

2.5.2 HMM embedding and recognition lexicon

In speech recognition applications, each phoneme is usually modelled with 3 HMM states and each state is represented with a number of Gaussian mixtures. In this study, we have used 29 phonemes, which correspond to the number of letters in Turkish. Each phoneme is modelled with 3 states and 12 Gaussian mixtures are used for each state.

The word or sub-word models are obtained with the concatenation of phonetic HMMs. In order to realise this concatenation, speech recogniser needs a recognition lexicon, which includes the pronunciation of words given to the recogniser. The pronunciations for each word are constituted using phonemes for which HMMs were trained.

Context independent phonetic models (mono-phones) have a high degree of variability partly due to differences in context in different realizations of the same phoneme. It is beneficial to use context dependent acoustic models such as tri-phones to reduce variability. A tri-phone indicates a realization of a phoneme when the preceding and the following phonemes are specified as well. As we benefit from the context of phonemes, tri-phone acoustic models perform much better as compared to mono-phone models. In this case, the recognition lexicon should include tri-phone pronunciations of words and phonetic model concatenation is realised according to this pronunciation lexicon.

The idea stated here can simply be summarised with the following figure:

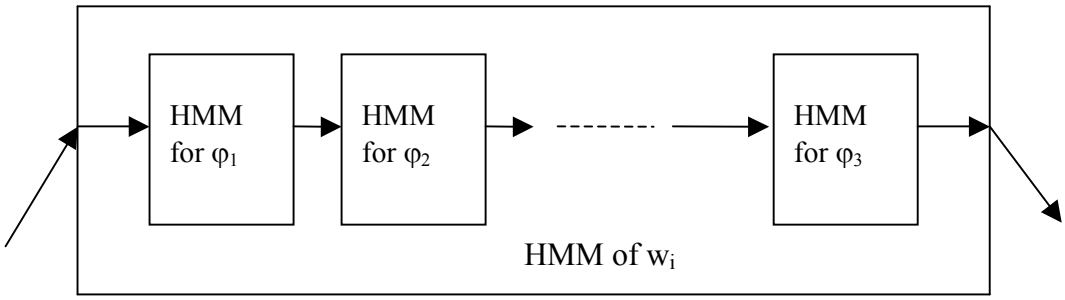

Figure 2-6 Phoneme representation of word HMMs

In this figure phonemes are represented with the symbol φ_i and word is represented with the symbol w_i.

2.6 Search Network

In the decoding part of recognition, a search on the resulting large-HMM to find the best path that best explains the observed acoustic feature vector is performed. The large-HMM is formed using language and acoustic models. Viterbi algorithm is commonly used to find the best path on this large search network [35].

As explained in 2.4 for the simplest language model, in which all histories $w_1 w_2 ... w_{i-1}$ are equivalent, sentence probability can be given as:

$$P(W) = \prod_{i=1}^{n} P(w_i) \tag{2-13}$$

If vocabulary units are represented with w_i, decoder searches the most likely path in the following network. This network can also be viewed as a large-HMM whose model parameters are obtained using language and acoustic model probabilities:

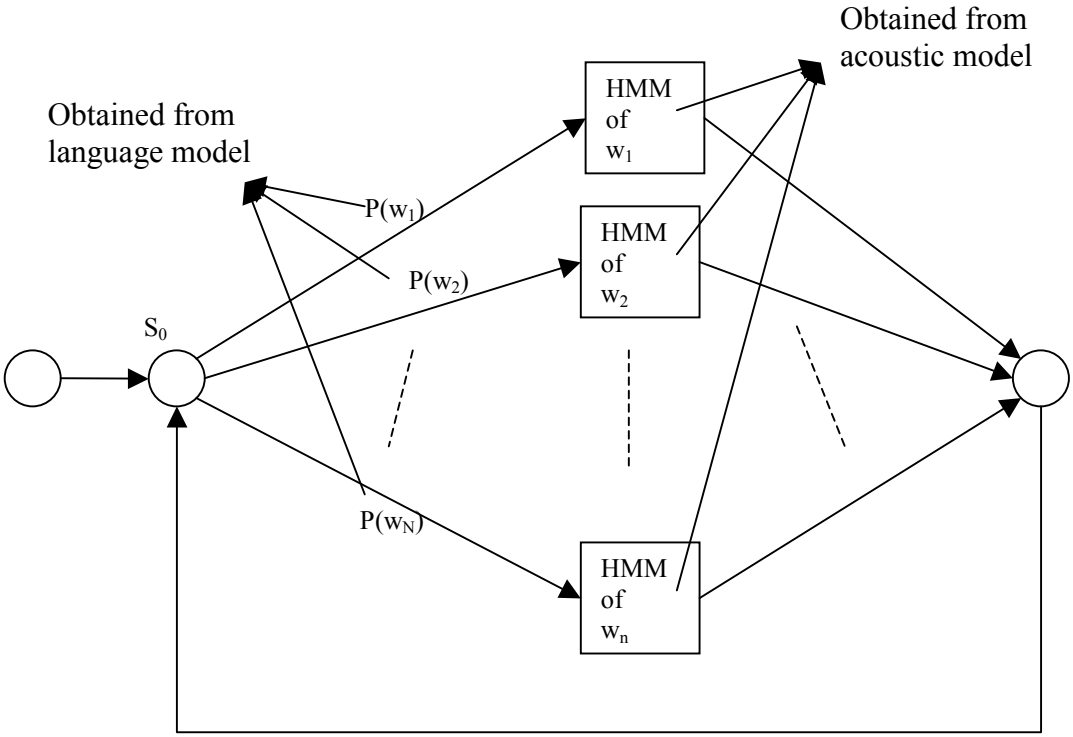

Figure 2-7 Search network with a uni-gram language model

Word HMMs are obtained with the concatenation of smaller phoneme HMMs as shown in Figure 2-6. Each phoneme is represented with 3 HMM states as depicted in Figure 2-5. These embedded boxes constitute the HMM of w_i in Figure 2-7. On the other hand the weights, $P(w_i)$, on the arcs of the search network are obtained from language model as indicated in Figure 2-7. A branch of this large-HMM can be shown as below:

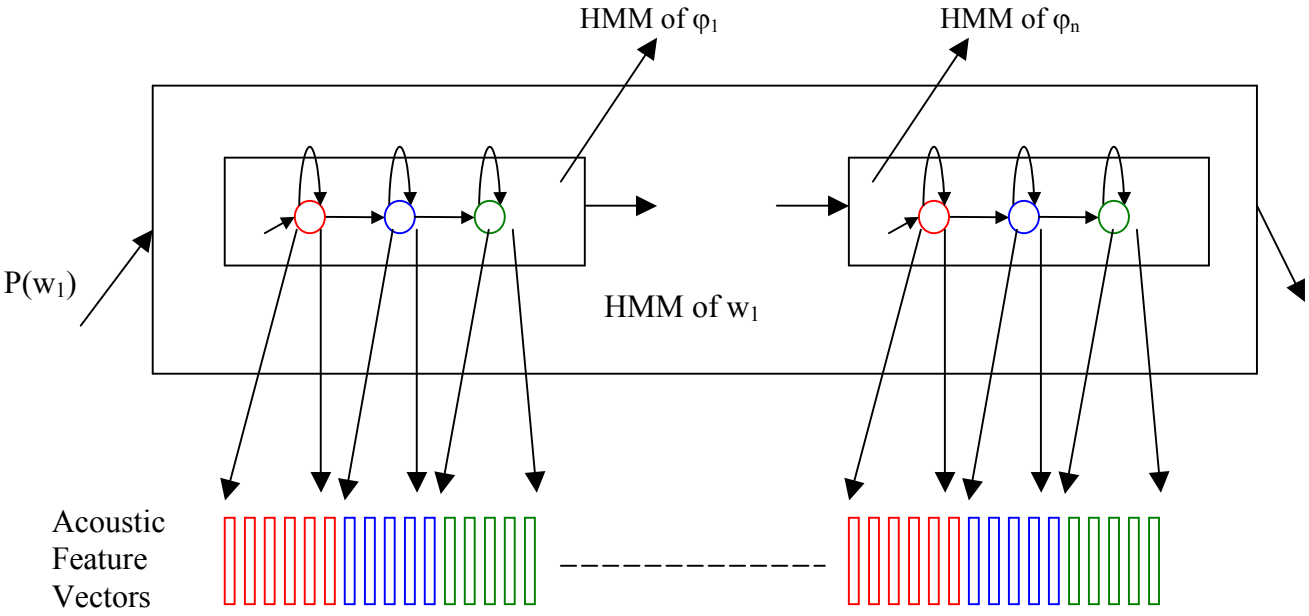

Figure 2-8 A branch of large search network in uni-gram language model

In bi-gram case, probability of a sentence can be written as;

$$P(W) = P(w_1)\prod_{i=2}^{n} P(w_i \mid Pw_{i-1}) \tag{2-14}$$

As a result, decoding algorithm searches the best path in the following more complicated WFSM:

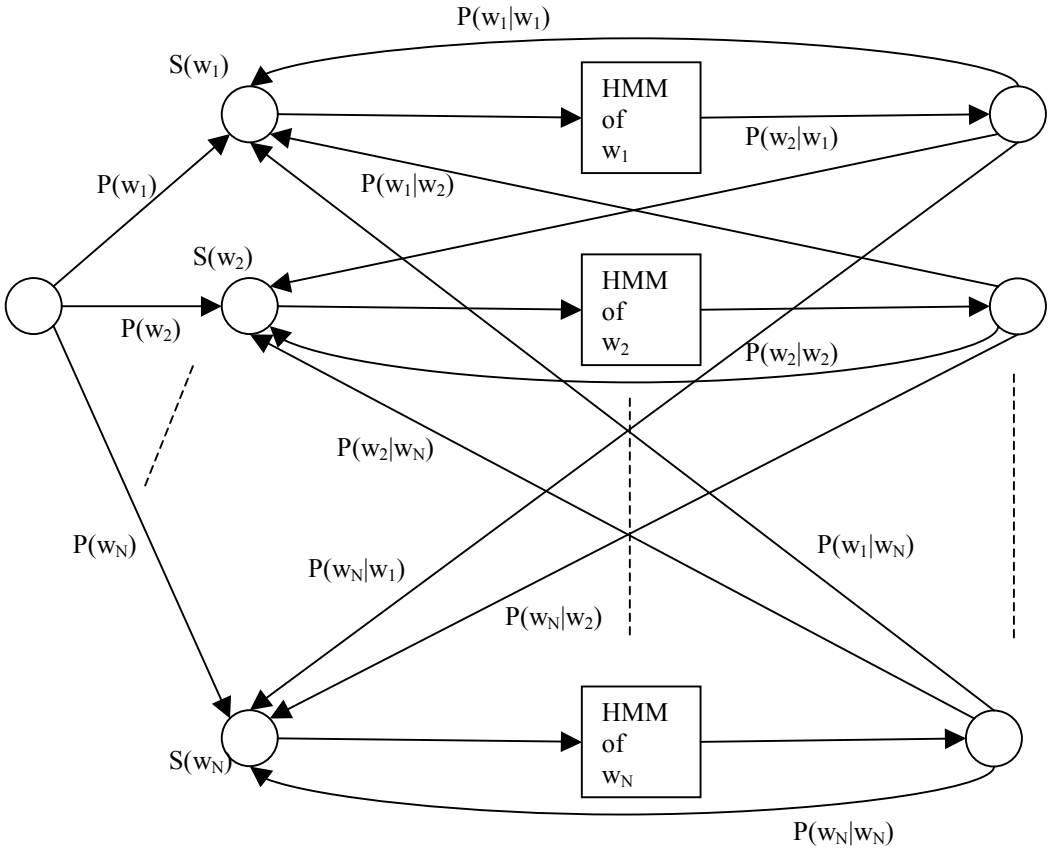

Figure 2-9 Search network with a bi-gram language model

For tri-gram case, search network becomes more complicated. As a result, the speed of recognizer decreases drastically with increasing order of N-gram language model. In addition memory requirements increase. Therefore usually bi-gram or tri-gram language models are used in most of modern speech recognition applications.

2.7 Speech Recognition Software Toolkits

There are available tools to obtain acoustic and language models for speech recognition. We extensively used two of these, Hidden Markov Toolkit (HTK) [29] and AT&T's GRMTOOLS [32]. HTK is developed in Cambridge University to implement

HMMs for general-purposes but it found most applications in speech recognition. AT&T's tools are developed in AT&T laboratories to implement general finite state machines (FSMs). This toolkit implements acoustic and language models with WFSMs. We also used Cambridge Statistical Language Modeling Toolkit Version 2 (CMU SLM) to evaluate the goodness of statistical language models. This toolkit is a set of unix software tools designed to facilitate the construction and testing of statistical language models [36].

2.8 General Evaluation Metrics

There are some known metrics to evaluate the performance of a speech recogniser. Before calculating these metrics, an optimal string match must be performed between the hypothesis and reference sentences. Dynamic programming is used to find the minimum edit distance alignment of hypothesis and reference. Once the optimal alignment is found, percentage correct and percentage accuracy rates can be obtained with the following formulas:

$$\text{Percent Correct} = \frac{N - D - S}{N} * 100\% \tag{2-15}$$

$$\text{Percent Accuracy} = \frac{N - D - S - I}{N} * 100\% \tag{2-16}$$

In these equations, N, S, D and I represent the total number of labels in reference file, substitution errors, deletion errors and insertion errors respectively [29].

Similarly word error rate can be found using the formula below if the recognition lexicon consists of words:

$$\text{Word Error Rate} = \frac{D + S + I}{N} * 100\% \tag{2-17}$$

Word error rate (WER) is a commonly used evaluation metric to measure the performance of a recognizer. When different base units are used as vocabulary of the speech recognizer, decoder outputs the hypothesized sentence in terms of these units. In order to have a fair comparison between the experiments, which are based on different lexicon entries, error rates should be calculated with respect to the same base unit. For this purpose, conversion of sub-word unit sequences to word units is needed. Extra symbols must be appended to lexicon entries of sub-word experiments in order to obtain this conversion.

We also feel the need to define new evaluation metrics that depend on recognizing sub-words because of agglutinative nature of Turkish. These new evaluation metrics are stem error rates (STER) and half-word error rates (HWER). The details of these new metrics are discussed in 4.4.

3 PROPERTIES OF TURKISH LANGUAGE

3.1 Overview

Turkish is an agglutinative language spoken in various parts of the world. Most of native speakers of Turkish live in Turkey. There are other places where Turkish speakers live such as Bulgaria, Uzbekistan, Kazakhstan, Kyrgyzstan, Tajikistan, Azerbaijan, Cyprus, Bulgaria, Macedonia, Greece and Germany. The experiments in this thesis are conducted for modern Turkish spoken in Turkey.

3.2 Agglutination

Turkish is an agglutinative language (like Finnish, Hungarian, etc…). In such languages, adding various suffixes to stems indicates grammatical functions. In other words, morphemes are attached to one or more free morphemes like beads on a string [37]. Each morpheme in the constituted word encodes linguistic information. Hankamer estimates a few million forms per verbal root based on generative capacity of derivations [38]. Besides over 250K words 6000 distinct morphological feature combinations were observed [40]. Because of this productivity, a steady vocabulary growth rate with no saturation is expected in Turkish language if words are chosen as vocabulary units.

The agglutinative nature of Turkish makes the language modelling of LVCSR more challenging compared to inflectional languages like English. Because of morphological productivity, any number of words as lexicon entries results in coverage problem of test words. The uncovered test words in the vocabulary do not have a chance to be recognized by the decoder. Therefore they worsen the recognition performance. The use of words as base units in lexicon does not achieve successful recognition performance due to this coverage problem, although the same units achieve 90-95% accuracy in English dictation system.

In addition to the coverage problem, the need for larger lexicon size degrades the speed of decoder in word-based speech recognition. It is noticed that when a smaller

dictionary is used in LVCSR with the help of sub-word units, 30% speed improvement can be achieved in recognition process [16].

After noticing these properties of Turkish, it can be said that base-units different from words must be utilized in Turkish LVCSR to solve the coverage problem, to improve the accuracy and to speed up the recogniser.

3.3 Free Word Order

Another important property of Turkish in terms of language modelling is free word order. Subject-Object -Verb word order in Turkish is a typical characteristic, but other orders are possible under certain discourse conditions. The perplexity of N-gram language models increase, since a sentence can be constructed with different order of words without changing the meaning. As a consequence, large amount of training data is needed in order to reliably train language model parameters.

3.4 Regularity

Morphotactics, which state how morphemes are combined and morphophonemic processes, are very regular with minor exceptions in Turkish [20]. This regularity of Turkish language can be utilized in speech recognition by incorporating different linguistic rules to the recogniser.

One of the regular rules in Turkish is vowel harmony. The vowel type in Turkish depends on the position of the tongue (front (F) or back (B)) and lips (rounded(R) or un-rounded (U)). In Turkish, there are eight vowels “a, ı, e, i, o, u, ö, ü”. The first two (a, ı) are back and un-rounded (BU), the next two (e, i) are front and un-rounded (FU), followed by two back and rounded (o, u) (BR) and two front and rounded (ö, ü) (FR) vowels. In Turkish, vowels in suffixes agree in certain phonological features with the last vowel of the stem. This property is called vowel harmony. For example suffixes –lar and –ler both make plural nouns, but we have to choose the one that has vowel harmony with the last vowel of preceding stem. There are some minor exceptions to this rule when for example a palatal “l” is the last consonant in the stem (some examples: futbol / football, alkol / alcohol, ampul / bulb, kemal/ a name in Turkish). In that case, even if the last vowel is a back vowel, it acts as a front vowel. Despite these types of minor exceptions most of the stem + ending pairs obey the vowel harmony rule.

Another rule in Turkish is verb-subject agreement in case and number, which is also obeyed in most of the sentences. In Turkish, the suffix appended to the end of verb must correlate with the subject of the sentence. These types of regular rules in a language can be utilized during the speech recognition process. In this research, we benefit from vowel harmony rule to improve the accuracy of recogniser. We realise vowel harmony rule using a WFSM. Details of this WFSM is discussed in chapter 4. Verb-subject agreement can also be implemented using a similar approach.

4 LVCSR FOR TURKISH: OUR IMPLEMENTATION

4.1 Acoustic Modelling for Turkish

4.1.1 Phonemes

We have used 29 different acoustic units in pronunciation lexicon of speech recogniser. Each acoustic unit corresponds to a letter in Turkish alphabet. We don't expect to see significant performance degradation due to this simple construction of pronunciation lexicon. This expectation is valid because of the fact that Turkish has an almost one-to-one mapping between graphemic representation of words and their pronunciation [37]. There are few pronunciation variations in Turkish letters. In addition this pronunciation mapping is the fastest and easiest way of obtaining pronunciation dictionary for Turkish. We have also trained "ğ" as an individual acoustic unit although its main function is known as to lengthen the preceding vowel.

Phonemes in this thesis can be listed as follows: a, b, c, ç, d, e, f, g, ğ ,h , ı, i, j, k, l, m, n, o, ö, p, r, s, ş, t, u, ü, v, y, z,

4.1.2 Decision tree for tri-phone clustering

Context dependent, tri-phone acoustic models are obtained from the trained mono-phone models. During the estimation of these models, many parameters can be under-estimated because of insufficient data associated with many of the states. As a result, tying states within tri-phone sets is needed in order to share data and make robust parameter estimates.

We have defined context questions in Turkish with respect to acoustic similarities, as in the case of English. The context questions are defined using the following clusters of phonemes:

Categories	Context question groups for consonants
Resonance	‘b’,’c’,’d’,’g’,’ğ’,’j’,’l’,’m’,’n’,’r’,’v’,’y’,’z’
Non-resonance	‘ç’,’f’,’h’,’k’,’p’,’s’,’ş’,’t’
Stop	‘p’,’t’,’ç’,’k’,’b’,’d’,’c’,’g’
Nasal	‘m’,’n’
Fricative	‘f’,’s’,’ş’,’v’,’z’,’j’
Liquid	‘l’,’r’
Glide	‘y’,’ğ’,’h’
Voiced	‘b’,’d’,’g’,’v’,’z’,’j’,’c’
Unvoiced	‘p’,’t’,’k’,’f’,’s’,’ş’,’ç’
Double-lip	’p’,’b’,’m’
Up-teeth	‘f’,’v’
Back-teeth	‘t’,’d’
Gum	‘s’,’z’,’n’,’r’,’l’,’ç’,’c’
Palate	‘k’,’g’
Hard-palate	‘ş’,’j’,’y’
Larynx	‘h’
Burst-palate	‘b’,’p’
Gum-palate	‘s’,’z’
Leakage-palate	‘ş’,’j’
Burst-gum-palate	‘c’,’ç’

Table 4-1 Decision-tree context question groups for clustering of consonants in Turkish

Categories	Context question groups for vowels
Back	‘a’,’ı’,’o’,’u’
Front	‘e’,’i’,’ö’,’ü’
Un-rounded	‘a’,’e’,’ı’,’i’
Rounded	‘o’,’ö’,’u’,’ü’
Low	‘a’,’e’,’o’,’ö’
High	‘u’,’ü’,’ı’,’i’

Table 4-2 Decision-tree context question groups for clustering of vowels in Turkish

Then states of context dependent HMMs are clustered by asking right and left context questions. In the final step, some pairs of clusters, which decrease the log likelihood of observation vectors less than a threshold, are merged. With decision tree tri-phone clustering, more robust parameters for tri-phone HMMs are obtained.

4.2 Splitting Words into Sub-words

Usage of sub-word units like syllables, morphological parts (stem and endings) as lexicon entries are proposed to overcome OOV rate problem encountered with word units [1-15]. There is more than one way to split a word into its parts especially in Turkish and other agglutinative languages. Some splitting options can be listed as follows:

1. Full-word (no split)
2. Stem + ending
3. Stem + morph1 + morph2
4. Syllabifying

In this research, we split the words using first, second, and fourth techniques. After this splitting procedure, speech recogniser's lexicon will include the following units:

1. Stems (used as a full-word or a half-word, examples: ev/house, sokak/street, duygu/emotion)
2. Endings (used as final half-words only, examples: -ler, -lar, -lerde, -imizin, -imizdekiler, -ından)
3. Syllables (<a>, <e>, <de>, <bak>, <kır>, <trak>,<ler#>)

Some syllables can also act as single-syllable stems, like "bak/look". In order to distinguish the stem "bak" from syllable "bak", an extra symbol is needed. Syllable units are contained within angular brackets to avoid confusion. We also append a "#" symbol to indicate word-final syllables to enable conversion of syllable sequences to word sequences at the output. Similarly "-" symbol is added in front of endings in order to prevent confusion with other units and to enable word reconstruction at the output.

Despite solving the coverage problem, using sub-word units in LVCSR for agglutinative languages does not always achieve acceptable performance [1, 2, 3, 5,15]. A reason for bad performance is the shortness of sub-word units as compared to full-words. As expected, shorter units become more acoustically confusable with each other.

Also, LVCSR decoders have a tendency to insert short units into wherever they can since matching short units to acoustic data is easier. Another reason for the performance drop in sub-word systems is the smaller effective language model history size while using N-gram language models. In a sub-word system, the language model effectively uses a shorter history as compared to a full-word based system. The smaller the sub-words, the shorter the effective history one uses in an N-gram language model. As an example for the word sequence "siyah kalem/black pen" with bi-gram language model when using words as language model entries, the model sees the probability $P(\text{kalem} \mid \text{siyah})$. However if the words are syllabified as "<si> <yah#> <ka> <lem#>", the last syllable just sees the probability $P(\text{lem\#} \mid \text{ka})$. As a consequence, sub-word units may not achieve good performance and they must be chosen very carefully to escape from aforementioned problems.

Choosing appropriate sub-word units is crucial for LVCSR systems. The main criteria in this choice should be to cover the words in the language as much as possible while maintaining small acoustic confusability between units not to decrease the recognition rates. In other words, unnecessary and short units should not be inserted into the lexicon of the speech recogniser while words in the language are covered with an acceptable rate. In this research, we propose some new ideas to construct a lexicon of half-word and full-word units for an LVCSR system in Turkish. We will not use the third splitting choice listed above since this splitting procedure as a morphological analysis of the word introduces unnecessary, acoustically confusable, small units to the lexicon. We will also use hybrid lexicon entries in which all the possible sub-word units will be included. In forming a hybrid lexicon, a word will be included in the vocabulary as a full word if it can be found among most frequent stems, if this is not possible we will try to split the word into two half-words as stem + ending. If this split is also not possible, then the word will be syllabified. Eventually the used lexicon entries during the experiment can be listed as below:

1. Full words
2. Stem + endings
3. Syllables
4. Hybrid

Our final lexicon contains all frequent stems, endings and syllables found in the training text corpus of language model. The results in this research will be given for

these lexicon entry units and the experiments will be called with lexicon entry's names. We will call the second experiment as half-word since the words in this experiment are split approximately from their middle points. We expect to see better performance from this experiment since the trade-off between coverage problem and acoustically confusable, small units is most suitably solved with half-word lexicon entries.

An important potential problem while using sub-word units in speech recognition is that, the recognizer may output a sequence of sub-words that may not form a legitimate word sequence in the language. For example vowel harmony rule between the stem and ending of a word may be violated at the output of the decoder. In this research we propose to use a rule-based weighted finite state machine (WFSM) to accept only allowable sub-word unit sequences. Although we specifically realize vowel harmony rule between stem and endings of a word with this rule based WFSM, other rules in Turkish can also be realized with the same approach. This WFSM can be composed with an N-gram language model to improve the overall language model and reduce its size. With this composition the new language model not only use the statistical knowledge obtained from large text corpus but also block the grammatically impossible recognition units. Modeling the language constraints with a WFSM is a new approach and with this approach not only language model size can be decreased but also recognition rates can be improved. The influences of proposed rule based WFSM on speech recognition performance will be shown in chapter 7.

4.3 A Novel Sub-Word Language Model Using a Rule-Based WFSM

N-gram language models can be represented by a weighted finite state machine (WFSM) [39]. Software packages are available to train N-gram language models and convert them into WFSMs like AT&T's GRM library [32]. The resultant N-gram WFSM is a large and complex one usually. The weights in this WFSM correspond to N-gram probabilities mentioned in language modelling section and they are trained statistically. In order to obtain reliable weights, a text corpus containing millions even billions of words is needed. This text corpus should be cleaned, tokenised and checked for typos before being used in N-gram language model training. Since this cleaning procedure is usually not perfect, some noise is always left in the training data and this noise causes imperfect weights in the WFSM. Also, to avoid zero probabilities and to enable detection of word sequences that were never seen in training data, linguistically

impossible word sequences receive a small but non-zero probability in the language model. If acoustic evidence strongly prefers a word sequence that was never seen in training data, the decoder can emit that sequence.

When only an N-gram statistical language model is used in a sub-word recogniser, the decoder can output linguistically impossible sentences in the language of interest. For example when the hybrid lexicon introduced in the experiment section is used, decoder could output such a sentence in Turkish LVCSR:

<a> <lis#> ev -inde -ler <a> <vi> <za#> bul –ındırmayı sev -mez.

Assume that the correct utterance is "Alice evinde avize bulundurmayı sevmez. / Alice does not like to have a chandelier in her house." There are linguistically impossible sub-word sequences in the decoder output. Errors in this recognition result can be listed as follows:

1. Two endings consequently recognized (Example: –inde -ler)
2. The ending following a stem does not obey the Turkish vowel harmony rule (Example: "bul –ındırmayı", should be "bul –undurmayı")
3. Although the word "avize/chandelier" is included in the lexicon of the speech recogniser, a sequence of syllables similar to the word with a better acoustic match to the acoustic data may be preferred.

In fact when the statistical language model is well trained, these types of problems will be reduced, but never avoided totally. In order to remove the problems listed above we are proposing to use a new rule-based WFSM that accepts only linguistically acceptable sub-word sequences.

The main role of the WFSM is to enforce vowel harmony between stems and endings. As Turkish is very regular language in which most of the rules are obeyed despite some exceptions, we expect to see an increase in speech recognition performance with this WFSM. We also take care of the words, which violates vowel harmony rule, and list these words carefully. This WFSM also enforces the correct ordering of stems and endings to form a valid word sequence. This property also improves the recognition accuracy as it eliminates the wrong ordered recognized sub-words.

The rule based WFSM is depicted in Figure 4-1. The WFSM contains parallel alternative branches to form a single allowable word that obeys vowel harmony rule in Turkish and accepts a sequence of such words.

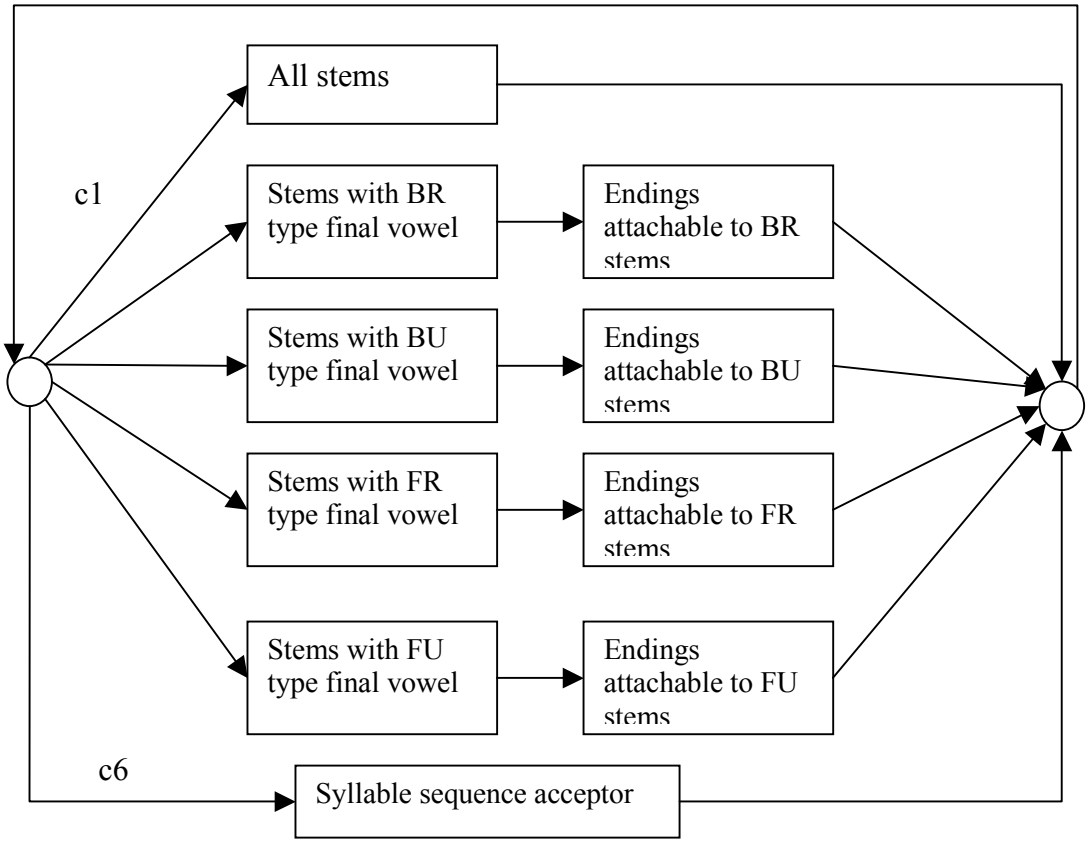

Figure 4-1 Rule-based WFSM, which accepts sub-word sequences that obey vowel harmony in Turkish

To realize the WFSM, we form four distinct classes of stems depending on their final vowel type. In Turkish, when attaching suffixes to stems, vowels of suffixes may change according to the last vowel of the stem they are attached to. This property is called vowel harmony as mentioned in 3.4. There are some rare exceptions to this rule such as alkol / alcohol, ampul / bulb, gol / goal, kemal/ a name in Turkish, saat / watch and kabahat / fault. In these cases, even if the last vowel is a back vowel, it acts as a front vowel. We incorporate such exceptions into our rules easily by including such stems into the front vowel classes instead of the back vowel classes. The allowable endings that attach to the class of stems can be found from the training text corpus of N-gram language model. We group together all endings that were attached to a class of stems in the training data. Even though the stem classes are distinct, there is overlap between ending classes since some suffixes do not change form according to vowel harmony (-ki, -ken) and some suffixes only change form according to being front or back regardless of the roundness attribute (-lar or –ler, -da or -de).

In Figure 4-1, the weights in the branches are shown as c1,...,c6. By using a small weight for the syllable branch, the probability of choosing this branch can be decreased.

Thus, the third problem listed above will not be observed too frequently. In addition there can be added weights to the branch between stems and attachable endings to these stems by counting the occurrences of individual stem+ endings pairs. Although we did not add these weights in this study, it may improve recognition performance slightly.

When WFSM shown in Figure 4-1 is composed with an N-gram WFSM, a new WFSM is obtained which not only uses the statistical information obtained from a large text corpus but also enforces vowel harmony and a linguistically correct half-word order. In the combined WFSM, decoder will never produce a recognition result which is not accepted by the rule based WFSM in Figure 4-1, effectively setting their language model probabilities to zero.

The WFSM displayed in Figure 4-1 may also be used in a syllable-based experiment in which the lexicon only contains syllables. In that case, WFSM can be designed to accept syllable sequences, which give meaningful words and eliminate the syllable sequences, which result in meaningless ones. When an N-gram syllable language model is composed with the rule based WFSM, a more powerful language model can be obtained since this language model will block meaningless syllable sequences. When such a rule based WFSM is not used, syllable-based experiments usually generate meaningless syllable sequences. As a result, a big improvement can be expected when such a WFSM is used in the language modelling of syllable experiment.

In some of our experiments, we set the weight for the syllable branch to 0 (c6=0 in Figure 4-1). This effectively disables syllable units and we use only half-words and full-words in recognition. In that case, we call our system a “half-word system”, since we only use full or half-words as lexicon units. When we set all c1, c2 ,…,c6 to be equal, we call this system “hybrid system” since syllable sequences are also possible in addition to half-word sequences. In the test-set that we worked with, half-word system yielded a better recognition result as compared to the hybrid system. This is due to the fact that we did not have a coverage problem with the half-word system, so we did not need the syllables to get increased coverage. We also did not optimise the weights c1 through c6 in our hybrid experiment and used equal values for them. We expect better results with the hybrid system when we optimise the weights and when we test our system on different test-sets, which have reduced coverage while using only half-words. We present our detailed results in chapter 7.

4.4 New Evaluation Metrics for Agglutinative Languages

Although word-error-rate (WER) is a well-accepted method for evaluating speech recognition performance, for agglutinative languages like Turkish, we feel the need to define new metrics that depend on recognizing sub-words. Usually decoder may recognize the stem of the word correctly but fail in recognizing the ending part. When we use WER as a metric, we count this as a single substitution error. However since the stem of the word is recognized correctly, we may wish to count it as a correct recognition and a substitution (or deletion) error. For example:

Reference sentence: bu savaş *takımı* olumlu etkileyecek (this struggle will positively affect the team)

Hypothesis sentence: bu savaş *takım* olumlu etkileyecek

In this example the root "takım / team" is correctly recognized but as the ending part "_ı" is not recognized, full word is taken as substiutiton error. In order to get rid of these types of accuracy calculation problems, we split the words in reference and hypothesis sentences into their morphological parts and obtain new recognition results with these morphological units. The morphological parts of words are found using two-level morphological analyser as in half-word experiment [20]. In this case, endings that contain one phoneme are also allowed contrary to half-word experiment. As a result, the example above is changed as follows:

Reference sentence: bu savaş *takım* ı olumlu etkile yecek (this struggle will positively affect the team)

Hypothesis sentence: bu savaş *takım* olumlu etkile yecek

After these changes while "takım / team" is counted as correct recognition, "_ı" is counted as deletion error. We think that this accuracy calculation gives more meaningful rates in speech recognition task of an agglutinative language.

Another alternative is to remove the ending of words and calculate error-rate using only the stems. This also gives an idea about the accuracy of the speech recogniser in recognizing the main part (stem) of the words, which is the most semantically informative part.

So, we use three different metrics in evaluating our speech recognition systems.

WER: word-error-rate. We form word sequences from our recognized sub-word sequences (this is not usually ambiguous since a stem + ending is a word and a syllable sequence that ends in a syllable with # symbol forms a word. For statistical-only

language models, we may have un-allowed sequences of sub-words. In those cases, we combine parts that we can combine to form words and the parts that cannot be combined are left as is.)

HWER: half-word-error-rate. We re-split the words into two (stem + ending) in both reference and recognized (hypothesis) word sequences. We use the same splitting algorithm as we used during language model training of half-word experiment with one exception that the ending part is allowed to be of length one character (one phoneme) in this case. We calculate the error-rate among these recognized half-words.

STER: stem-error-rate. After dividing the words into two parts, we delete the second part (if the second part exists) for both the reference and the hypothesis texts and calculate the error-rate among stems only.

In our experiments, we provide error-rates using the above three metrics.

5 EXPERIMENTAL SETUP, TRAIN and TEST DATA

5.1 Properties of Text Corpus

5.1.1 Nature and Amount of Train Text Data

Text corpus used in language modelling is collected using automatic text collection programs in php. The subjects of collected text can be divided into four categories: Daily life, sports, authors, e-books. We have also used extra text data whose subject is not known. Number of sentences and words for these four categories can be summarised with the following table:

Categories	Number of sentences	Number of words
Daily Life	886.451	14.630.764
Sports	701.037	10.390.687
Authors	1.786.722	24.363.470
E-books	214.494	2.486.198
Extra text data	1.967.745	29.028.794

Table 5-1 Properties of training text corpus

As a result during the training of language model we have used 5.556.449 different sentences and 80.899.913 total words collected from four major subjects. From this training text corpus, 1.170.526 unique words are obtained. This large number of unique words shows OOV rate problem in Turkish LVCSR.

5.1.2 Nature and Amount of Test Text Data

Perplexity and bi-gram hits are calculated using two test texts whose recorded speech files are also used in recognition experiments. One of these test sets contains 106 sentences from Yasar Kemal's well-known book, "Memed, My Hawk". Yasar Kemal makes use of distinct Turkish in his books. He intentionally uses some words, which are locally used in some parts of Anatolia to better reflect the people living in these regions. In addition, his sentence structure is different from regular, modern Turkish to represent

dialect of Anatolian villagers. As a result, this test set has challenging sentence structure in terms of language modeling and also contains infrequent words such as “çakırdikeni / a plant in Anatolia”, “köre / ant hole”, “yornuk / necessity”. This test set is especially chosen to see the brittleness of statistical language models to domain changes and the effects of this brittleness to speech recognition performance. The second test set contains 88 sports news sentences collected from different newspapers. We expect to represent sports news test better with available language model since sentences in sports news domain are also used during the training of statistical language models. There are total 920 words in the test texts and both test texts are not used in the training of language models. In this thesis, former test data is called as “literary novel test”, latter one is called as “sports news test”.

5.1.3 Modifications in Train and Test Corpus

The large training text corpus contains just full words when it is first collected. Before using it in half-word, syllable and hybrid experiments, it should be modified properly. The text corpus should be syllabified for syllable experiment and bi-gram model should be trained using this syllabified text corpus. In the half-word experiment, words are split into stem + endings by using two-level morphological analyser [20]. When there is more than one morphological split possible, the one with the longest stem is chosen. To avoid using acoustically confusable short endings, at least 2 phonemes-long endings are also required. After all the words in training corpus are split to their morphological parts, they are replaced with these parts in training text corpus. This modified text corpus is used in the training of bi-gram language model for half-word experiment. In hybrid experiment the most frequent stems and endings in half-word experiment are used. For this purpose, specified number of most frequent stems and endings in half word experiment are listed. Then a word stays as a full word in text corpus if it can be found among the listed most frequent stems. If this is not possible, the word is split into half-words as stem + ending. In order to perform this split, the listed most frequent stems and endings must include particular stem and ending. If this half-word split choice is also not possible, the word is syllabified.

Test texts are also modified by applying the same rules in training text corpus modification in order to obtain unit recognition rates. As a result slightly modified texts for language model training and accuracy calculation are obtained for each experiment.

5.1.4 Vocabulary Size and OOV Rate

For each test a vocabulary size is determined and this size is used throughout the experiments. In the determination of this vocabulary size, special care is given to cover test units as much as possible not to worsen speech recognition performance significantly. However acoustic confusability between vocabulary units also increases with the increasing vocabulary size. There is a trade-off between vocabulary size and acoustic confusability. We choose the vocabulary sizes for each experiment to balance this trade-off for the benefit of speech recognition accuracy. Eventually vocabulary in experiments contains most frequent stems, endings and syllables in the training text data. The vocabulary sizes and OOV rate with these sizes can be given in the following table:

Lexicon Type	Vocabulary size	Out of vocabulary (OOV)
Word	30132	18.16%
Half-word	13073	1.66%
Syllable	2000	6.04%
Hybrid	23053	0.66%

Table 5-2 Vocabulary sizes and OOV rate

In half-word experiment vocabulary contains 10073 most frequent stems and 3000 most frequent endings. Likewise in hybrid experiment the vocabulary contains 18053 stems, 3000 endings and 2000 syllables. For these experiments, we manually added stems that are in our test text data but are not in the most frequent stems in the training text data to the ASR acoustic dictionary for all lexicon types. This procedure added 132 extra entries to the word lexicon, 73 extra entries in the half-word stem lexicon and 53 extra entries in the hybrid stem lexicon. These different numbers are due to different initial lexicon sizes for each lexicon type. This gives us a performance result that mimics an ideal scenario where we have a stem dictionary available and all test words are morphologically derived from these stem words in the dictionary. In real life, of course this is not realistic and there will always be words that are not in the lexicon nor they can be derived from a word in the lexicon. However, to reduce the effects of unknown words, we performed such a manual adjustment.

As seen from the table, low OOV rate is obtained with small lexicon size in syllable experiment. Half-word and hybrid experiments also give small OOV rates.

Nevertheless when words are used as lexicon entries, OOV rate becomes high although vocabulary size is chosen higher in this experiment compared to other ones.

5.1.5 Perplexity and bi-gram hits for different lexical units

The evaluation of different language models is made using CMU SLM toolkit [36]. Bi-gram language models are obtained for each test using the modified training text corpus mentioned in 5.1.3.

With aforementioned vocabulary sizes in 5.1.4, following perplexity and bi-gram hits are obtained for each bi-gram language models in sports news and literary novel test texts:

Lexicon Type	Perplexity	Bi-gram hits
Word	191.52	95.41%
Half-word	119.62	95.16%
Syllable	31.36	99.95%
Hybrid	91.85	98.38%

Table 5-3 Evaluation of bi-gram language models using sports news test

Lexicon Type	Perplexity	Bi-gram hits
Word	489.75	85.12%
Half-word	426.02	85.82%
Syllable	68.51	99.82%
Hybrid	253.17	91.97%

Table 5-4 Evaluation of bi-gram language models using literary novel test

As we see from the evaluation results language models better represent the sports news test than literary novel test. For example while perplexity with sports news test is 119.62 for half-word experiment, it becomes 426.02 with literary novel test. This result was expected since the training text corpus contains sports sentences. In addition as mentioned earlier literary novel test has very difficult sentence structure in terms of language modelling. As a result it can be concluded that, the result coincides with the brittleness of statistical language models across domains. Since sports news test is better captured with bi-gram language model, we expect to see more accurate recognition results for this test compared to literary novel test.

The best perplexity and bi-gram hits are obtained for syllable experiment. This is due to the smaller vocabulary size of this experiment. Although syllable based bi-gram language model is better than the other models, we may not get very accurate recognition results with these small units since recogniser can easily confuse the units of this experiment with each other. In order to improve the recognition performance of the experiment, the acoustic information in syllable units must be increased. Some frequent syllable sequences can be merged to increase the acoustic information. Looking at the most frequent bi-gram or tri-gram syllable sequences in training text corpus can make the choice of the sequences that will be merged. This merging idea will also be applied in syllable experiment and the performance of merged syllable experiment will be discussed in section 6.2.

Eventually we should point out that half-word units give slightly better bi-gram language model compared to full word units. In addition OOV rate of half-word units is also much smaller than word units. This superiority of half-word units might affect the recognition performances on behalf of half-word experiment.

5.1.6 Illustration of the coverage problem in Turkish

We obtain number of unique units versus number of sentences graph in order to see vocabulary size increase with different units. To draw this graph different number of sentences from training text corpus is taken, and number of unique units in these sentences are calculated. This graph can show us which vocabulary unit grows faster compared to other units. We may expect high OOV rate with the fastest growing unit and conclude that it is not suitable choice as lexicon entry of speech recogniser.

We should also point out that since there is a lot of noise in the training text corpus, some of the words are not syllabified or split in to their morphological parts. If a word could not be processed during morphological split or syllabification, it is taken as full word. These noisy words increase the number of syllables and morphological parts more than expected. Especially the number of stems in half-word experiment increases because of these noisy words since they are considered as stem in language modelling. Most of these noisy words are exclamation, foreign or misspelled words. Regardless of this increase; we can see the desired result from this graph, which is seeing the fast growing lexicon units with increasing number of sentences. Obtained graph can be seen below:

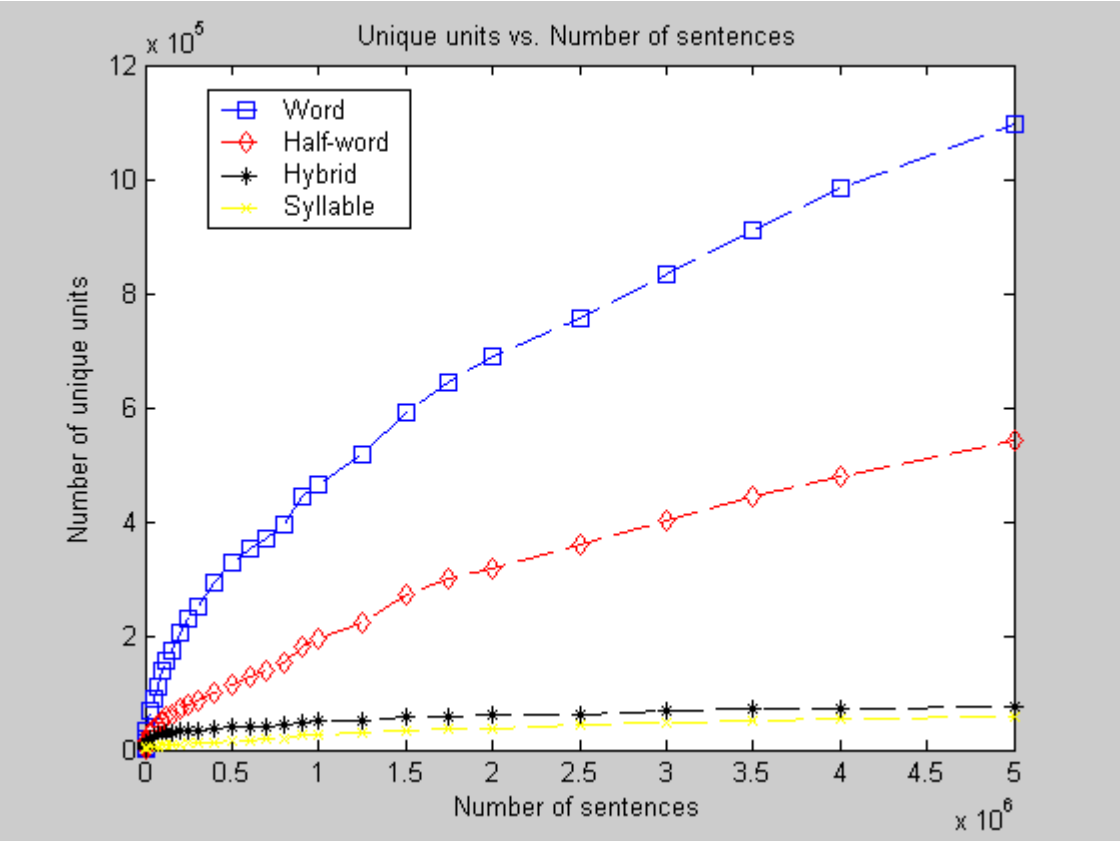

Figure 5-1 Number of sentences vs. number of unique units

As seen from the graph, the fastest unique unit increase is observed with words, and the slowest increase is observed with syllables. Number of words increase with high slope even for very large number of sentences. This is due to agglutinative nature of Turkish and OOV rate problem with the use of words as lexicon entries becomes more evident with this graph. Number of unique hybrid units is very close to number of unique syllable units for the same number of sentences. Another important observation from this graph is that the rate of increase in half-word units is approximately half of the rate in word units. This is a good property in terms of speech recognition and this property of half-word units shows that these units can be used to overcome OOV rate problem.

We also obtained number of added new units versus number of sentences graph. This graph shows how many new units are added to on hand unique units when more text is available. As this graph is very sensitive to domain changes, it is obtained for a subset of training text corpus in which no domain change occurs. The conclusion obtained with previous graph is also verified with this graph.

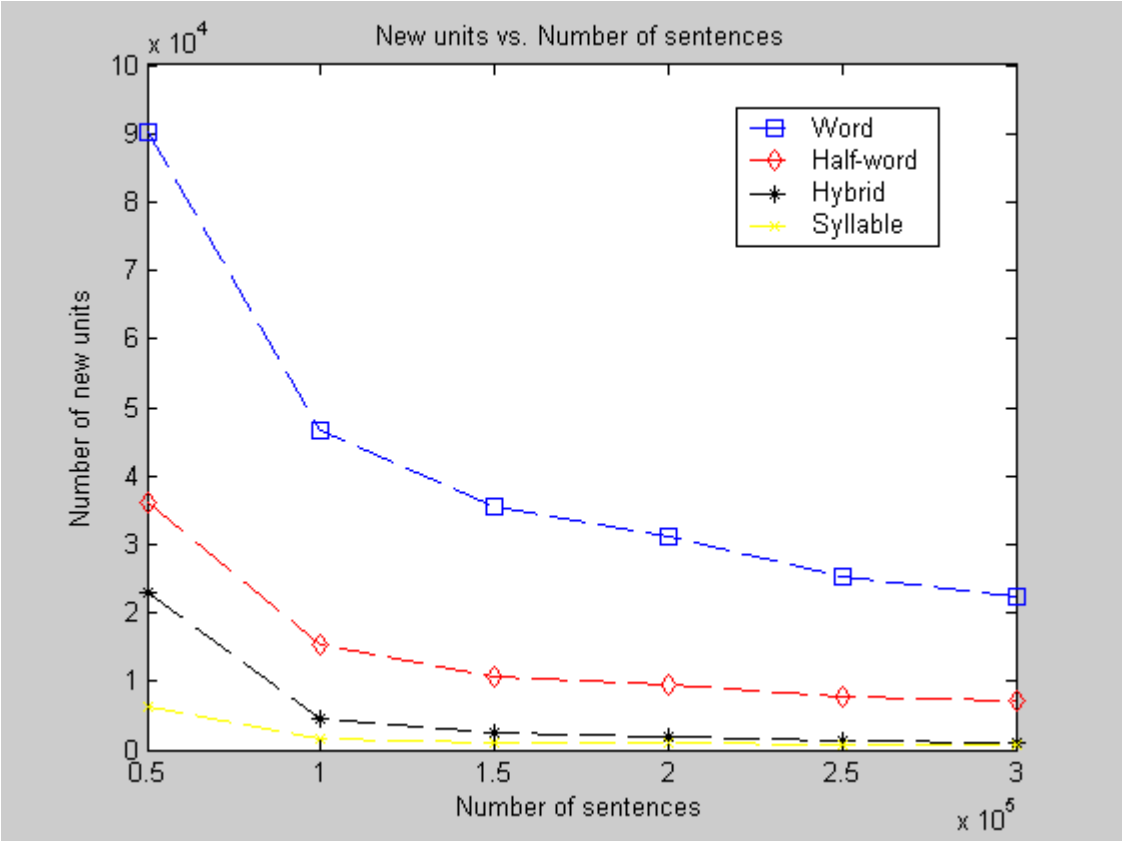

Figure 5-2 Number of sentences vs. number of new units

5.2 Properties of the Speech Data

5.2.1 Nature and amount of the speech data

During the experiments of LVCSR for Turkish we have used various different sources of speech data. We have collected speech data from 356 different speakers with the help of freshman students in Sabanci University. The number of males and females are approximately equal in this set of people. Each person reads nearly 100 phonetically balanced sentences ranging in its subjects. In this speech collection process, 32 hours of data is obtained. In addition to this source some part of Middle East Technical University (METU) speech data, which is collected by Dr. Tolga Çiloğlu, is also used [18]. This data contains 5 hours 26 minutes of recordings. As a result, our speech database contains 37 hours and 26 minutes of speech data. All speech data is sampled at 16 kHz and 16 bits are used during the quantisation. Aforementioned properties can be summarised in the following table:

Categories	Hours of data	Size of data (in MB)
Sabanci University	32	3158
METU	5,26	702
Total	37,26	3860

Table 5-5 Properties of speech data

This amount of speech data is divided into two categories: test and train. Test speech data is not used during the training of acoustic models and set aside for testing the accuracy of recogniser. 28 hours and 40 minutes of total speech data is used for the training of acoustic models. In order to test the recogniser 3 hours and 18 minutes of speech data is used. Test data can also be divided into two categories: sports news and literary novel. In sports news test data, there are 88 sentences collected from sports news of different newspapers. These sentences are recorded with 16 different speakers. While 9 of these speakers are male, remaining 7 of them are female. Literary novel test contains 106 different sentences taken from famous Turkish writer Yasar Kemal's well-known book "Memed, My Hawk". 18 different people are recorded the sentences of this test set. While 11 of these speakers are male, remaining 7 of them are female. As a consequence, using recorded speech of 34 different speakers in two different subjects makes speaker independent test of recogniser. Mentioned properties of test and train speech data can be summarized with the following table:

Categories	Hours of data	Size of data (in MB)	Number of people
Train	34.08	3448	322
Test	3.18	412	34

Table 5-6 Division of speech data to test and train

As a result of aforementioned distinct properties of literary novel test, we might expect that the language model trained with a general text corpus may not be able to capture the properties of this test as much as sports news test. This language modeling difference naturally affects the recognition performances for two test sets.

5.2.2 Acoustic Model Training

HMMs are used to train acoustic models. For this purpose we have been benefit from HTK. We utilized MFCC feature extraction method in the training of acoustic models. Window size and period between each parameter vector are chosen 25 msec

and 10 msec, respectively. Number of filter-bank channels and number of cepstral coefficients are chosen 26 and 12, respectively. Cepstral mean normalization is employed to remove the effects of a transmission channel on the input speech waveform. Energy measure, delta and acceleration coefficients discussed in section 2.3 are also appended to the parameter vectors to enhance system's performance.

29 acoustic units for each letter in Turkish are trained using the speech data explained in section 5.2.1. We also trained one short pause model to better capture the short silence between words and one silence model to be able to recognize the silences at the beginning and end of sentences. The training is made for two different cases: syllable and word based training. Syllable-based acoustic models are used when syllables are used as the lexicon entries. Word-based acoustic models are used when the other units are used as lexicon entries. In word-based training a short pause is assumed between words. Nevertheless a short pause is assumed between syllables in syllable-based training. HTK optionally inserts a short pause between units. It inserts the short pause, if it can find a suitable acoustic structure for it. Nonetheless if it cannot match short pause structure, it does insert nothing. As a result of this optional insertion, two training approaches may resemble to each other in this respect. On the other hand we may expect to encounter problems in syllable based training since most of the syllables in Turkish are two phonemes long. This may cause underestimation of some tri-phone acoustic units. It may also cause inaccurate model estimates.

First of all mono-phone training for acoustic units is made. Context-dependent tri-phones are also obtained by simply cloning mono-phones and then re-estimating using tri-phone transcriptions [29]. Tri-phone transcriptions are obtained for each training sentences beforehand. In addition all possible tri-phones in a large text corpus are obtained for syllable and word based cases separately. In this way, 19558 different tri-phones for word based training and 15684 different tri-phones for syllable-based training is obtained from 24389 possible tri-phones in Turkish. Tied-state tri-phones are also obtained from set of untied tri-phone models. This step is necessary and beneficial as discussed in section 4.1.2. Three states for each phoneme are used in training. At most 12 Gaussian mixtures are employed for each state. All the recognition results in this report are given for tri-phone training with 3 states for each phone and 12 Gaussian mixtures for each state.

In order to see the sufficiency of available speech data to train tri-phones, we have obtained a number of statistics for training. After the tying of tri-phones, 10650

different tri-phones for word-based and 4103 different tri-phones for syllable-based training are used during the estimation of HMM parameters. In word-based training 5859 of total 10650 tri-phones never occurred in tri-phone transcriptions of training sentences. This is 55% of obtained total tri-phones from the large text corpus. 151 of these tri-phones occurred once (≅1.42% of total tri-phones) and 146 of these tri-phones occurred twice (≅1.37% of total tri-phones). In syllable-based training 2585 of total 4103 tri-phones never occurred in tri-phone transcriptions of training sentences. This is 63% of obtained total tri-phones from the large text corpus. 35 of these tri-phones occurred once (≅0.85% of total tri-phones) and 51 of these tri-phones are occurred twice (≅1.24% of total tri-phones). Average number of occurrences for tri-phones, which occurred at least once in training transcriptions, is 141 in word-based training. The same number is 321 for syllable-based training. Total number of states in word-based training is 6701 after tri-phones are tied. This number is 3204 for syllable-based training. Each state is trained using approximately 100 acoustic evidences in word-based training and 152 in syllable-based training. These numbers obtained from the training of HMMs carry important information about the sufficiency of training speech data and goodness of acoustic models.

6 EXPERIMENTAL RESULTS

6.1 Explanation of Experimental Procedure

6.1.1 Overview

Recognition experiments are conducted for four different lexicon units: word, half-word, syllable and hybrid. As explained above bi-gram language models are obtained in HTK using aforementioned large text corpus modified with respect to lexicon unit. Higher order language models like tri-gram could not be used because of language model restriction of HTK to bi-gram. HMMs trained with speech data in section 5.2, is used as acoustic model during the experiments. While syllable-based acoustic training is used for syllable experiment, word-based acoustic training is used in the other experiments.

6.1.2 Parameters of Speech Recogniser

Recognition experiments are conducted for the test sentences described in section 5.1.2. For the test of different experiments, we used the available tools in HTK. There are some important parameters in HTK, which significantly affect the performance of recogniser. One of these parameters is defined with –s option. This parameter sets the grammar scale factor and post-multiplies the language model likelihood from the word lattices. Second parameter is defined with –t option. This parameter enables beam-searching algorithm such that any model whose maximum log probability token falls more than specified value below the maximum for all models is deactivated. Setting this parameter to zero disables the beam search mechanism. Last factor is given to HTK with –p option. This factor sets word insertion log probability. This option specifies a value that is added whenever a path enters a new word. For example paths containing N words have this value added N times [29]. In order to obtain the best performance from the recogniser, these parameters should be set optimally.

6.2 Recognition Results

In our experiments, we especially want to see the effect of language modelling factor in speech recognition. Therefore we set beam searching (-t option) and word insertion (-p option) parameters to 120 and 0 respectively. We obtain recognition results for three different grammar scale factors (-s option): 5, 7, and 10. In this way, we can see the effects when just grammar scale factor is increased in recognition experiments. Detailed speech recognition results for word, half-word, syllable and hybrid lexicon units are provided in appendix.

The most important problem in syllable experiment is the shortness of lexicon entries. These small lexicon units do not contain enough acoustic information and this increases acoustic confusability between units. Because of this problem, we cannot get very accurate results in syllable experiment, although bi-gram perplexity and OOV rate of these units are smaller compared to other units.

In order to increase the acoustic information in syllable units, we merged some of syllable sequences and obtain longer lexicon entries. For this purpose inside word tri-gram and bi-gram frequencies of syllables are found from the text corpus of syllable experiment. Most frequent 10000 tri-gram and 5000 bi-gram syllable sequences are chosen. Then these syllable sequences are merged inside the word. During the merging procedure, if a word has three or more syllable units, a merged tri-gram syllable sequence is searched inside the syllables of the word. If a match for a merged tri-gram syllable sequence is found, the syllables inside the word are merged and a new word is obtained. Then a new merged tri-gram syllable sequence is searched inside the new syllables of the word. This procedure is repeated until reaching to the last syllable of the word. When the last syllable of the word is reached, the same search procedure is applied for merged bi-gram syllable sequences.

After merging some syllable units, bi-gram language model with these new units is trained in HTK. In the vocabulary of this experiment we have chosen 10000-merged most frequent tri-gram syllable sequences, 5000-merged most frequent bi-gram sequences. In addition, we inserted 200-syllables to lexicon in order to increase the coverage. As a result we have obtained 15200 units in the vocabulary of merged syllable experiment. Word-based acoustic units are used as acoustic model in this experiment. As we increase the acoustic information in lexicon units by merging some frequent syllables, we may expect to improve the accuracy compared to syllable

experiment. Detailed recognition result for merged-syllable experiment is provided in the Appendix.

The recognition results can be summarized with following two graphs. In these graphs unit accuracy rates are shown with respect to changing s-value:

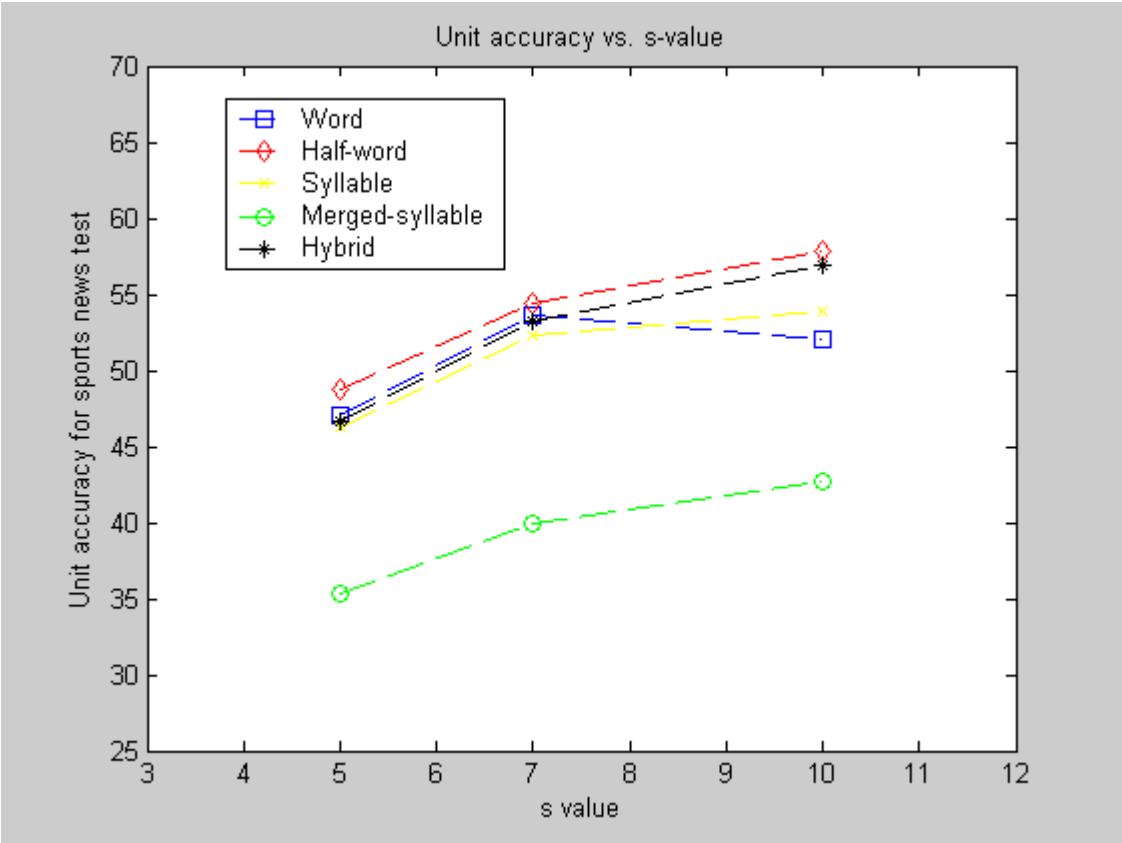

Figure 6-1 Accuracy vs. grammar scale factor for sports news test

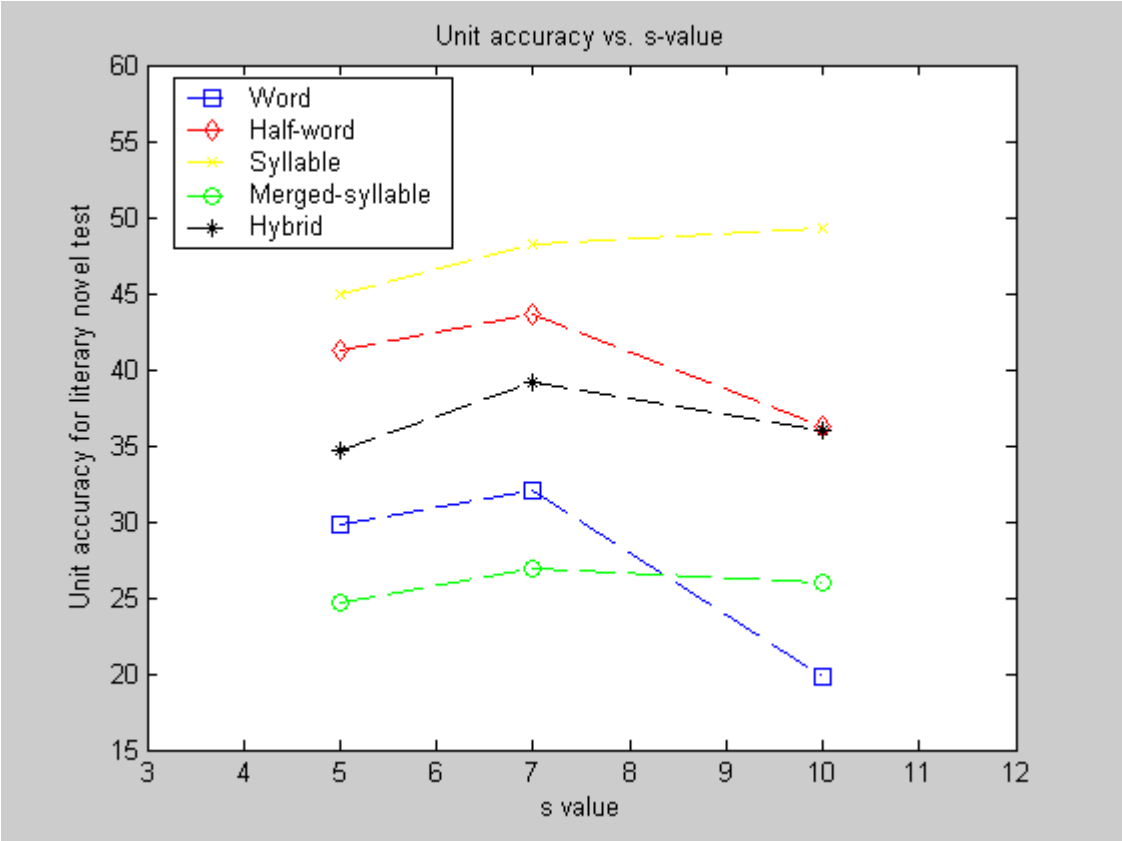

Figure 6-2 Accuracy vs. grammar scale factor for literary novel test

These results are calculated for lexicon units in speech recognition. As lengths of units are different from each other, we can't compare the experiments efficiently by just looking at these recognition rates. In order to have a better comparison between the experiments, we provide word error rates. For this purpose, we have obtained word units from the recognized sub-word unit sequences. Obtaining words from recognized sub-word units is not usually difficult since we use extra symbols to represent the characteristic of sub-word units. For example syllable units are contained within angular brackets and also a "#" symbol is appended to indicate word-final syllables. Similarly "_" symbol is added in front of endings. These extra symbols enable conversion of syllable sequences to word sequences at the output. At the output of half-word experiment, a stem + endings is converted to a word. Similarly a syllable sequence that ends in a syllable with # symbol forms a word in syllable experiment. For statistical-only language models, we may have un-allowed sequences of sub-words (e.g. in half-word experiment we may have two endings adjacent to each other although this is an un-allowed sub-word sequence). In those cases, we combine parts that we can combine to form words and the parts that cannot be combined are left as they are.

In addition to word error rate (WER), we give half-word error rate (HWER) and stem error rate (STER) of the experiments. These evaluation metrics are discussed in section 4.4. We do not obtain these evaluation metric results for hybrid experiment since better unit recognition performances are obtained in half-word experiment compared to hybrid experiment. In addition to this performance superiority of similar experiment, un-allowed sub-word sequences are frequently observed in hybrid experiment. As these un-allowed sub-word sequences will not be converted to words, some of the sub-word units will be left as they are. These unconverted sub-word sequences will cause deceptive results for hybrid experiment. Because of these reasons we did not obtain WER, HWER and STER for hybrid experiment.

WER, HWER, STER are obtained for the grammar scale factor (-s option) which gives the best unit recognition accuracy in each experiment. Error rates can be seen in the following tables:

Lexicon Type	Sentence correct %	WER	HWER	STER
Word	15.20	46.44	37.58	36.96
Half-word	11.99	45.11	40.12	36.30
Syllable	2.57	66.58	63.22	60.33
Merged-syllable	1.79	61.35	53.73	52.89

Table 6-1 Comparable percentage error-rates for different lexicon types in sports news test

Lexicon Type	Sentence correct %	WER	HWER	STER
Word	6.64	68.01	57.98	59.85
Half-word	7.79	58.66	55.08	53.70
Syllable	3.15	72.62	67.84	69.16
Merged-syllable	2.63	79.98	73.20	75.38

Table 6-2 Comparable percentage error-rates for different lexicon types in literary novel test

To better compare the recognition performances of the experiments following figure is provided for WER and STER:

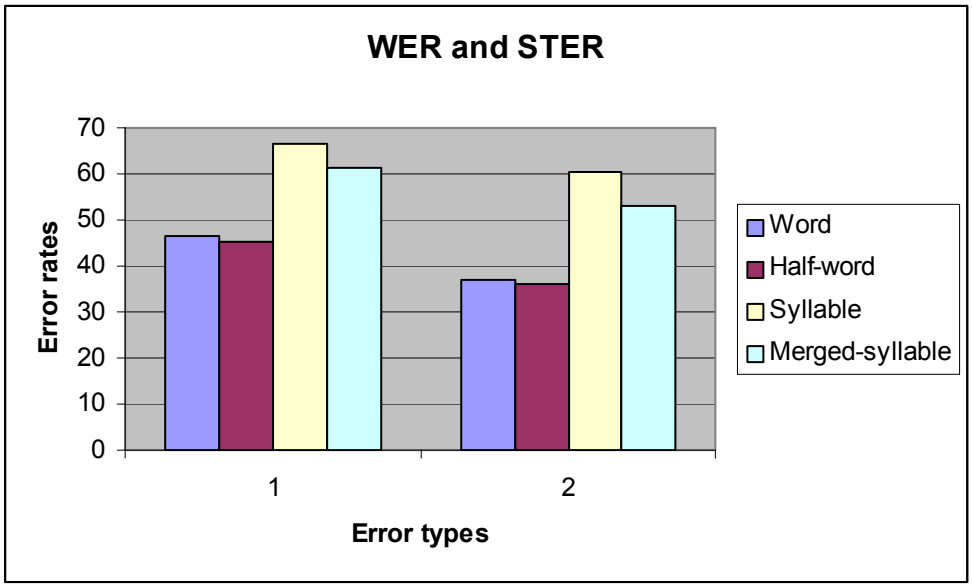

Figure 6-3 WER and STER for sports news test

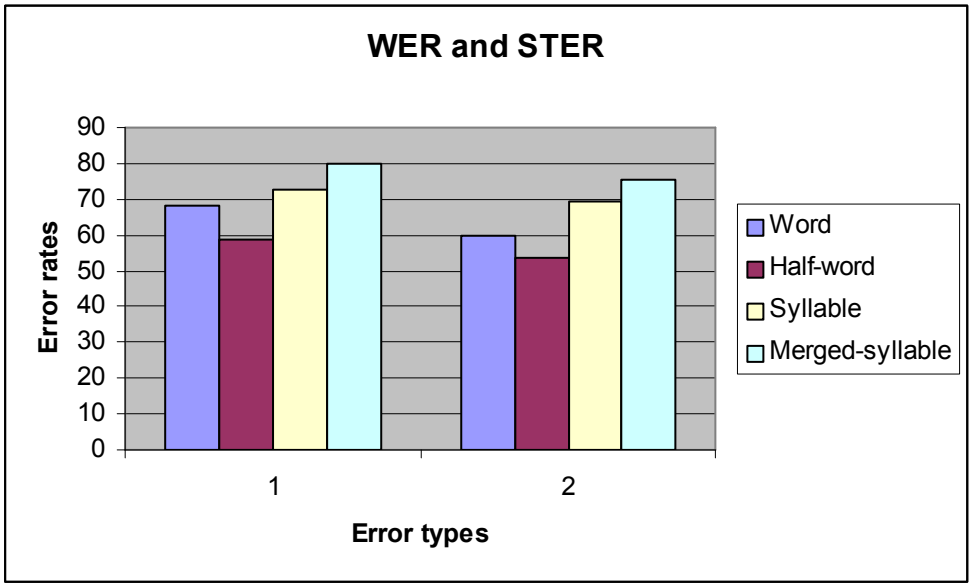

Figure 6-4 WER and STER for literary novel test

6.3 Discussion and Comparison of Results

As seen from Table 6-1 and Table 6-2, better recognition rates are obtained for sports news test compared to literary novel test for all base units. This is not a surprising result since approximately 700.000 sentences in sports domain are included in the training corpus of bi-gram language models. In addition, as stated in section 5.1.2 literary novel test has challenging sentence structure in terms of language modelling and contains infrequent words, which may make difficult to obtain an accurate language model with a general training text corpus for this test set. The perplexity and bi-gram

hits show that sports news test is better modelled with the available training text corpus compared to literary novel test. The perplexity of literary novel test is approximately four times larger than the perplexity of sports news test in half-word and word experiments. There is also significant difference between the bi-gram hits of literary novel and sports news tests in the same experiments. For example while perplexity of sports news is 191.52, perplexity of literary novel is 489.75 in word-based experiment. Moreover while bi-gram hits of sports news are 95.41%, bi-gram hits of literary novel are 85.12% in the same experiment. On the other hand, syllable based bi-gram language model closely represent both test sets. In this experiment, perplexity of literary novel is twice more than perplexity of sports news and bi-gram hits are very close to each other.

This language modelling difference between two test sets shows the brittleness of language models across domains. As argued in section 2.4.5.2, language models are extremely sensitive to changes in the style and topic of the text on which they are trained if especially words are used as base unit [33]. This brittleness also affects the recognition performances. In all of lexicon types used, sports news test gives better accuracies. For example there is 21.44% difference between WER of two test sets in word experiment. There is 13.55% difference between WER in half-word experiment. The WER difference between two test sets is closer to each other in syllable experiment compared to word and half-word experiments. This closeness is also expected since the lexicon entries are very small in syllable experiment. These small lexicon entries can model both of test sets closely as seen from perplexity and bi-gram hits.

Another important observation from the results is the effect of grammar scale factor on recognition performance. Figure 6-1 indicates that recognition accuracies are increased with increasing language model post-multiplier for sports news. On the contrary, Figure 6-2 indicates that recognition accuracies decrease when this parameter is increased to 10 in literary novel. This observation is due to the fact that bi-gram language models better capture the characteristics of sports news. Because of this better representation, the recognition accuracies in sports news are improved when the weights of language model on large search network is increased.

Recognition accuracies clearly show that sports news test is better modelled with the available training text corpus compared to literary novel test. As a result, more accurate recognition performances are obtained in this test set for all lexicon types. Because of this property, a subset of sports news is chosen as test set to see the improvements in speech recognition performance with higher order or more

sophisticated language models. The results obtained when we incorporate rule based WFSM and tri-gram language model to speech recogniser will be given for this test set as the influence of these more advanced language models on recognition performance can be more accurately seen with this set.

As discussed in sections 5.1.4, 5.1.5 and 5.1.6, half-word units give better perplexity and coverage compared to word units. Nevertheless half-word units contain less acoustic information as they contain fewer phonemes per lexicon unit. In most frequent 30132 words used as vocabulary of word experiment, average number of phonemes per word is 7.91. This number is close to the number reported in [2]. Average number of phonemes per unit is 6.26 in half-word experiment. In addition as discussed in 4.2, smaller units see shorter language model history in N-gram language models. These two problems might result in bad performance of sub-word based experiment. In our experiments although we can't improve the recognition results significantly, half-word experiment achieves 1.33% WER reduction in sports news compared to word experiment. WER reduction is 9.35% in literary novel. The most important reason for this improvement is high coverage of half-word experiment. As given in section 5.1.4, coverage is about 82% in word experiment. On the other hand coverage is 98% in half-word experiment with fewer lexicon entries.

Syllable experiment does not achieve comparable recognition performances to word and half-word experiments despite the superiority of syllable based bi-gram language model. WER of syllable experiment is 21.47% worse than half-word experiment in sports news test. WER is 13.96% worse in literary novel. First reason of this worse performance is that syllables are not constrained to form a valid word when using statistical language models. Second reason is the smallness of syllable units. In most frequent 2000 syllables used as vocabulary of this experiment, average number of phonemes per syllable is 3.25. This number is much less than the average number of letters per unit in half-word and word experiments. High acoustic confusability between these small lexicon units degrades recognition performance. In this study we merged some frequent syllable sequences to improve recognition accuracy of this experiment.

Merged syllable experiment achieves 5.23% WER reduction compared to syllable experiment in sports news test. Despite this improvement, the recognition accuracy is not comparable to word and half-word experiments. WER of merged syllable experiment is 16.24% more than half-word experiment. However the accuracy is not improved in literary novel when some of frequent syllables are merged, even

performance is worsened. There are some reasons for this performance degradation of merged syllable experiment. First reason is acoustically similar units. Especially merged syllable sequences, which resemble to a meaningful word, are mixed with the real word in recognition experiments. Some examples of these word are “trab-zon-spo /resembles the word trabzonspor which is the name of a Turkish Premier League team” and “is-tan-buls / resembles the name of the biggest city in Turkey, istanbul”. Second reason of performance drop is the short pause (sp) insertion problem. Some examples for this type of mis-recognized units are “biliyorzaman”, “yediyirmi”, and “istediönemli”. If a short pause between “biliyor” and “zaman”, “yedi” and “yirmi”, “istedi” and “önemli” could be inserted; these units would be correctly recognized. This error type occurs frequently in merged-syllable experiment. The third reason is having more than one way to merge syllable sequences in a word. In other words, we have several possible merged syllable representations of most words and we choose one of them. As a result of these reasons, recognition performance of merged syllable experiment degrades. I think that if these problems can be solved efficiently with new approaches, the accuracy can be improved significantly.

7 RULE-BASED WEIGHTED FINITE STATE MACHINE

7.1 Explanation of Experimental Procedure

7.1.1 Overview

While testing the effects of rule based WFSM on speech recognition performance, we set the weight for the syllable branch to 0 (c_6=0 in Figure 4-1). This effectively disables syllable units and only half-words and full-words are used in WFSM. As a result of this, we can use recognition units of half-word experiment as the test set of rule based WFSM, since only full and half-words are used in this experiment, too. We intentionally chose half-word based experiment, since it yielded slightly better recognition performance as compared to the other experiments. In addition the rule-based WFSM can improve accuracy of half-word experiment more than the other experiments as it forces the vowel harmony between the stem and endings of words.

As indicated in chapter 6, better performance is obtained for sports news test compared to literary novel test for all lexicon types. This is due to the fact that bi-gram language model better models sports news compared to literary novel. As a result of this fact, we tested the recognition accuracy of rule based WFSM and tri-gram language model on a subset of sports news.

7.1.2 Lattice Re-scoring Paradigm

In order to see the improvements with rule-based WFSM on speech recognition performance, general lattice re-scoring technique is employed. To achieve this, N-best lattices (N is chosen 1000) with bi-gram language model in half-word experiment is obtained using HTK. These lattices are produced for the recordings of three speakers chosen from sports news test set. During lattice generation, grammar scale factor (-s option), beam-searching threshold (-t option) and word insertion log probability (-p option) are set 10, 120, and 0 respectively since these parameters give the best performance in half-word experiment. These lattices are then converted to AT&T's

FSM format. Tri-gram language model is also trained using AT&T's GRM library with 2 million sentences from the same text corpus discussed in section 5.1. Rule based WFSM in Figure 4-1 is obtained in AT&T's FSM format by setting weight c6 to zero. The lattice re-scoring experiment can be summarized with the following steps

- Generation of 1000-best lattices with bi-gram language model in HTK.
- Converting these lattices to AT&T's FSM format with their weights.
- Obtaining tri-gram language model with 2 million sentences in AT&T's FSM format.
- Realizing rule-based WFSM in AT&T's FSM format.
- FSM composition of these WFSMs.
- Finding new best path in re-scored lattices.

The paradigm can be seen in the in the figure depicted below:

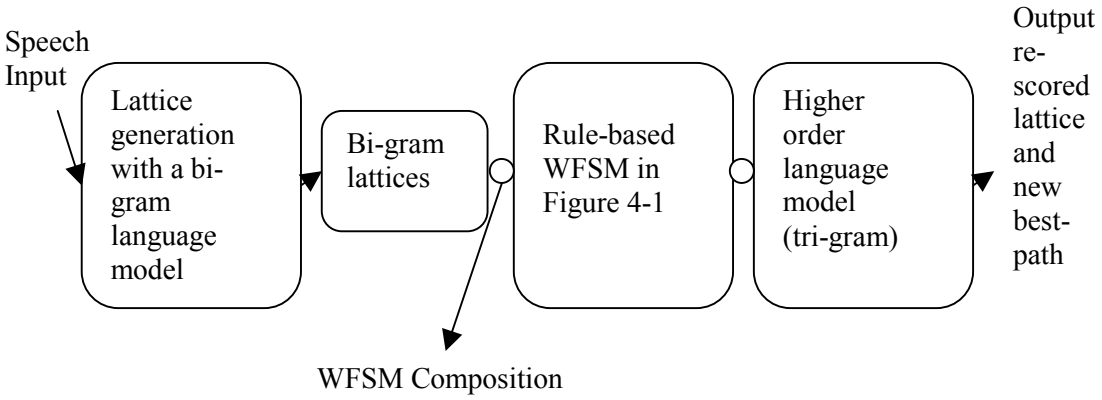

Figure 7-1 Lattice re-scoring paradigm used in testing rule-based WFSM and tri-gram language models.

7.2 Recognition Results

The detailed results obtained with this lattice re-scoring approach can be summarized with the following table:

Language Model used with half-word lexicon	Sentence correct %	WER	HWER	STER
Bi-gram	11.74	40.51	36.51	31.53
Rule-based WFSM	11.79	39.66	34.89	30.26
Tri-gram	19.70	33.06	30.19	25.83
Rule-based + Tri-gram	19.77	32.54	29.10	25.28

N-best bi-gram (oracle)	31.44	21,25	19.20	15.14

Table 7-1 Percentage error-rates obtained by various language models

In this table the first row shows bi-gram error rates in test sentences. Following two rows show the error rates when vowel-harmony (or in other words rule-based) WFSM or tri-gram language model is re-scored the bi-gram lattices individually. The fourth row shows the error rates of re-scored lattices with both rule based and tri-gram WFSMs. The last row is the best result than can theoretically be attained from bi-gram lattices.

WER and STER for various language models can be summarized with the following graph:

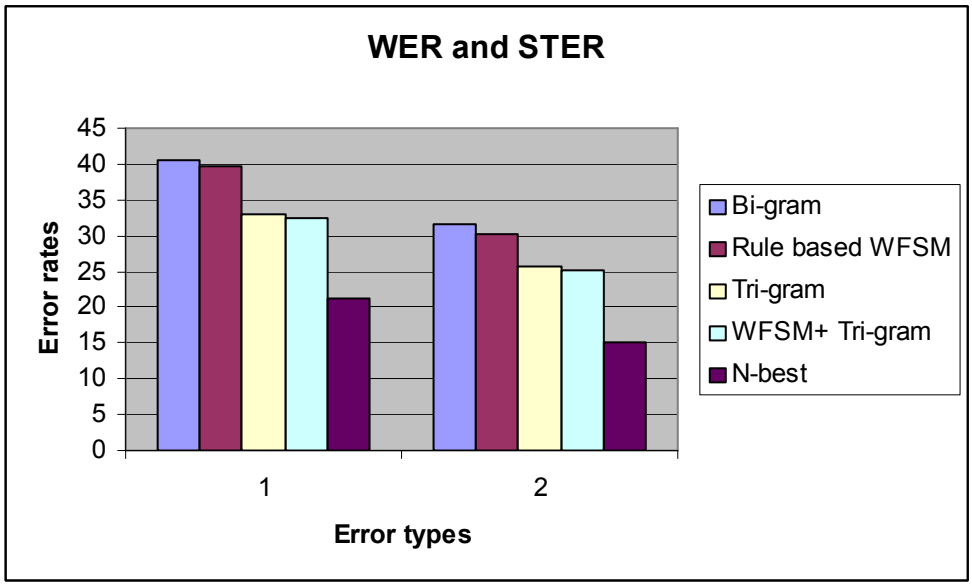

Figure 7-2 WER and STER for various language models

7.3 Discussion of Results

As seen from results, significant speech recognition accuracy improvement is obtained with the new language models. 2.1% relative performance improvement in WER has been observed when rule-based WFSM is applied to bi-gram lattices. In addition 18.4% reduction in WER has been seen with tri-gram language model. When both rule-based WFSM and tri-gram language model is applied to bi-gram lattices, 19.68% improvement is achieved in WER. These improvement rates are higher for STER. When rule-based WFSM is applied to bi-gram lattices, 4.02% improvement in STER is obtained. When tri-gram language model is applied, the improvement is 18%. When both WFSMs are re-scored bi-gram lattices, we achieve 19.8% STER reduction.

Another observation from these results is that applying rule-based WFSM directly to re-score bi-gram lattices results in a bigger improvement than applying it after tri-gram re-scoring. Because bi-gram language model is sub-optimal compared to tri-gram and returns more half-word sequences that are disallowed according to linguistic rules. When rule-based WFSM is applied to bi-gram lattices directly we get 2.1% improvement. On the other hand when it is applied after tri-gram re-scoring, 1.57% improvement is achieved.

In the light of obtained recognition performances, we can say that tri-gram language model significantly improve the recognition accuracy. Its effect is 18% relative advance of recognition performance. Rule-based WFSM also improves the results. We get approximately 3% improvement when this WFSM is used in speech recognition. Moreover when both rule-based and tri-gram language models are incorporated into recognizer, approximately 20% improvement is achieved. As a result, the best performance is obtained with half-word lexicon units when rule based WFSM and tri-gram language models are re-scored bi-gram lattice in sports domain. Because of this performance improvement we can conclude that, the recognition performances can be improved by incorporating language constraints into speech recognizer. In addition we can say that, this study has given promising results for LVCSR in agglutinative language, Turkish.

8 CONCLUSIONS and FUTURE WORK

In this study, we have investigated large vocabulary continuous speech recognition (LVCSR) task for Turkish. For this purpose we have collected approximately 37 hours of speech data. While 34 hours of this data is used for training acoustic of models, 3 hours is used for testing. We have used approximately 5.5 million sentences in the training of statistical language models. This amount of text and audio corpora is one of the largest training corpora used in Turkish LVCSR so far.

As Turkish is an agglutinative language, coverage problem occurs when just words are used as the lexicon entries of speech recognizer. In order to solve this problem, usage of sub-word units is proposed. In this study words, syllables, morphological parts and hybrid of these three units as lexicon entries are examined and speech recognition performances based on different lexicon units are compared. When morphological parts of the words are used as lexicon entries, we have approximately divided the words from their half as stem + endings. Because of that, this experiment is called as half-word experiment during the study.

In order to see the performance of recognizers based on different lexicon units, we have used two different test sets: literary novel and sports news. We have also introduced two new evaluation metrics stem error rates (STER), and half-word error rates (HWER) in addition to word error rates (WER). We have seen that better recognition accuracies are obtained for sports news test for each lexicon entries, since statistical language model better represents this test set. We have also seen that word and half-word experiments give much better performances compared to the other experiments. In addition we have observed that half-word experiment gives slightly better recognition accuracies compared to word experiment due to small out of vocabulary rate.

In order to improve recognition performances, we have used higher order statistical language models (tri-grams). Moreover we have implemented a rule based weighted finite state machine (WFSM) to incorporate language constraints into speech recognizer. We have chosen to implement vowel harmony rule in Turkish language with this rule based WFSM. We have used bi-gram N-best lattices of sports news test in half-

word experiment to test these new language models. As a result, we get approximately 3% relative improvement when rule based WFSM is incorporated to speech recognizer. Tri-gram language model also improved the accuracies with approximately 18% relative. When both rule-based and tri-gram language models are used, approximately 20% relative reduction in error rates is achieved. In conclusion, 32.54% WER and 25.28% STER are obtained as the best recognition performance in sports news domain. We think that these recognition accuracies are very promising for Turkish LVCSR.

In future, the performance of syllable experiment can be improved by using a WFSM, which just allows the syllable sequences that give a meaningful word. When this type of WFSM is composed with a statistical language model, significant improvement can be obtained in syllable experiment. This WFSM can also be used in merged-syllable experiment in which some syllables are merged to increase the acoustic information in lexicon units. In addition, implemented WFSM for vowel harmony rule in Turkish can be optimized by inserting weights, which are trained from the large training text corpus, to the WFSM. Hybrid based experiment can also be studied more extensively for the cases in which the coverage of test units with other lexicon entries is very difficult. In order to see the performance of hybrid experiment more accurately, a rule based WFSM similar to the one used in this research must be incorporated to the recognizer during decoding. In this case, decoder can eliminate any illegal output sequences in search network. As a result the main source of errors in hybrid-based experiment can be removed. Lastly more speech and text corpora can be used in experiments as the amount of available sources play crucial role in the performance of recognizer.

REFERENCES

1. E. Arisoy, *Turkish dictation system for radiology and broadcast news applications,* M.S. Thesis, Bogazici University, 2004.
2. Kenan Çarki, Petra Geutner, and Tanja Schultz, "Turkish LVCSR: Towards better Speech Recognition for Agglutinative Languages," *ICASSP*, 2000.
3. Ciro Martins, João P. Neto, Luís B. Almeida, "Using Partial Morphological Analysis in Language Modeling Estimation for Large Vocabulary Portuguese Speech Recognition," *Eurospeech*, 1999.
4. T. Rotovnik, M. S. Maučec, B. Horvat, Z. Kačič, "Large Vocabulary Speech Recognition of Slovenian Language Using Data-Driven Morphological Models," *Lecture Notes in Computer Science Vol. 2448*, 2002.
5. M. Maucec, T. Rotovnik, Z. Kacic, B. Horvat, "Large vocabulary speech recognition of Slovenian language using morphological models," *EUROCON*, 2003.
6. Daniel Kiecza, Tanja Schultz, Alex Waibel, "Data-Driven Determination of Appropriate Dictionary Units for Korean LVCSR", *ICSP*, 1999.
7. Oh-Wook Kwon, "Performance of LVCSR with word-based and syllable-based recognition units," *ICASSP*, 2000.
8. Oh-Wook Kwon, Kyuwoong Hwang, Juan Park, "Korean Large Vocabulary Continuous Speech Recognition of Newspaper Articles", *ICSP*, 1999.
9. Oh-Wook Kwon, Kyuwoong Hwang, Juan Park, "Korean large vocabulary continuous speech recognition using pseudomorpheme units," *Eurospeech*, 1999.
10. Kadri Hacioglu, Bryan Pellom, Tolga Ciloglu, Ozlem Ozturk, Mikko Kurimo and Mathias Creutz, "Word splitting for Turkish LVCSR," *SIU,* 2003.
11. Kadri Hacioglu, Bryan Pellom, Tolga Ciloglu, Ozlem Ozturk, Mikko Kurimo and Mathias Creutz, "On lexicon creation for Turkish LVCSR," *Eurospeech,* 2003.

12. Vesa Siivola, Teemu Hirsimaki, Mathias Creutz and Mikko Kurimo, "Unlimited vocabulary speech recognition based on morphs discovered in an unsupervised manner," *Eurospeech*, 2003.
13. Teemu Hirsimäki, Mathias Creutz, Vesa Siivola and Mikko Kurimo, "Morphologically Motivated Language Models in Speech Recognition" *International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning (AKRR)*, 2005.
14. Helin Dutagaci and Levent M Arslan, "A comparison of four language models for large vocabulary Turkish speech recognition," *ICSLP,* 2002.
15. W. Byrne, J. Hajie, P. Ircing, F. Jelinek, S. Khudanpur, P. Krbec and J. Psutka, "On large vocabulary continuous speech recognition of highly inflectional language - Czech," *Eurospeech,* 2001.
16. P. Geutner, "Using morphology towards better large vocabulary speech recognition systems," *ICASSP*, 1995.
17. Ebru Arısoy, Levent M. Arslan, "Turkish Radiology Dictation System", *SPECOM*, 2004.
18. Ozgur Salor, Bryan Pellom, Tolga Ciloglu, Kadri Hacioglu, Mubeccel Demirekler, "On Developing New Text and Audio Corpora and Speech Recognition Tools for the Turkish Language", *ICSLP*, 2002.
19. Mathias Creutz and Krista Lagus, "Unsupervised discovery of morphemes," *Workshop on Morphological and Phonological Learning of ACL*, 2002.
20. K. Oflazer, "Two-level Description of Turkish Morphology," *Literary and Linguistic Computing, Vol.9 No.2*, 1994.
21. Mathias Creutz "Unsupervised segmentation of words using prior distributions of morph length and frequency," *ACL-03, the 41st Annual Meeting of the Association of Computational Linguistics*, 2003.
22. K. Koskenniemi, *Two-level morphology: A general computational model for word-form recognition and production,* Ph.D. thesis, University of Helsinki, 1983.
23. V. Siivola, M. Kurimo and K. Lagus, "Large Vocabulary Statistical Language Modeling for Continuous Speech Recognition in Finnish", *Eurospeech*, 2001.
24. Peter Scheytt, Petra Geutner, Alex Waibel, "Serbo-Croatian LVCSR On The Dictation And Broadcast News Domain", *ICASSP*, 1998.

25. Petra Geutner, “Introducing Linguistic Constraints into Statistical Language Modeling,” *ICSLP*, 1996.
26. R. Isotani and S. Matsunaga “Speech recognition using a stochastic language model integrating local and global constraints,” *ARPA SLT Workshop*, 1994.
27. Ebru Arısoy, Levent M. Arslan, “Dayanıklı Konuşma Tanıma Uygulamaları için Evrensel bir İnsan-Makine Dilinin Geliştirilmesi”, *SIU*, 2003.
28. Frederick Jelinek, *Statistical Methods for Speech Recognition*, MIT Press, Cambridge, Massachusetts, 1999.
29. S. Young, D. Ollason, V. Valtchev and P. Woodland, *The HTK Book (for HTK Version 3.2),* Entropic Cambridge Research Laboratory, 2002.
30. Peter Linz, *An Introduction to Formal Languages and Automata,* Jones and Bartlett Publishers International, London, 1997.
31. Daniel Jurafsky and James H Martin, *Speech and language processing,* Prentice Hall, New Jersey, 2000.
32. AT&T grm tools library, http://www.research.att.com/projects/mohri/grm
33. R. Rosenfeld, “Two decades of Statistical Language Modelling: Where do we go from here?” *Proceedings of the IEEE, Vol. 88, pp. 1270-1278*, August 2000.
34. Lawrence Rabiner and Biing-Hwang Juang, *Fundamentals of Speech Recognition,* Prentice Hall, New Jersey, 1993.
35. A. J. Viterbi, “Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm,” *IEEE Trans. Inf. Theory, vol. IT13, pp. 260-269*, April 1967.
36. P.R. Clarkson and R. Rosenfeld, “Statistical Language Modeling Using the CMU-Cambridge Toolkit,” *Eurospeech*, 1997.
37. Kemal Oflazer and Sharon Inkelas, “A Finite State Pronunciation Lexicon for Turkish,” *Proceedings of the EACL Workshop on Finite State Methods in NLP*, April 2003.
38. Jorge Hankamer, “Morphological Parsing and Lexicon,” In W. Marslen-Wilson editor, *Lexical Representation and Process*, MIT Press.
39. M. Mohri and M. Riley, “Weighted finite-state transducers in speech recognition (tutorial),” *ICSLP*, 2002.
40. Dilek Zeynep Hakkani-Tür. *Statistical Language Modelling for Agglutinative Languages.* Ph.D. Thesis, Department of Computer Engineering, Bilkent University, August 2000, Ankara, Turkey.

APPENDIX

With the parameter set described in 6.2 and two different test categories (sports news and literary novel), we have obtained the following unit recognition performances for different lexicon units. The performances are calculated using the formulas in section 2.8. These results are summarized in Figure 6-1 and Figure 6-2:

-s	Sentence correct %	Unit correct %	Unit accuracy %
5	11.93	60.57	47.05
7	15.20	64.02	53.56
10	16.14	62.43	52.09

Table A 1 Word based recognition results for sports news test

-s	Sentence correct %	Unit correct %	Unit accuracy %
5	6.21	51.48	29.77
7	6.64	51.45	31.99
10	5.27	42.56	19.77

Table A 2 Word based recognition results for literary novel test

-s	Sentence correct %	Unit correct %	Unit accuracy %
5	6.58	52.83	48.72
7	8.95	57.34	54.45
10	10.70	60.57	57.88

Table A 3 Half-word based recognition results for sports news test

-s	Sentence correct %	Unit correct %	Unit accuracy %
5	5.32	49.11	41.25
7	6.95	50.15	43.68
10	5.69	45.15	36.19

Table A 4 Half-word based recognition results for literary novel test

-s	Sentence correct %	Unit correct %	Unit accuracy %
5	1.81	48.10	46.27
7	2.85	53.61	52.34
10	2.57	54.75	53.88

Table A 5 Syllable based recognition results for sports news test

-s	Sentence correct %	Unit correct %	Unit accuracy %
5	1.68	50.23	44.98
7	2.36	52.17	48.28
10	3.15	52.36	49.33

Table A 6 Syllable based recognition results for literary novel test

-s	Sentence correct %	Unit correct %	Unit accuracy %
5	7.43	50.69	46.62
7	10.01	55.96	53.27
10	12.26	59.41	56.85

Table A 7 Hybrid based recognition results for sports news test

-s	Sentence correct %	Unit correct %	Unit accuracy %
5	4.21	39.31	34.72
7	6.37	42.33	39.21
10	7.02	41.10	36.02

Table A 8 Hybrid based recognition results for literary novel test

-s	Sentence correct %	Unit correct %	Unit accuracy %
5	1.50	38.41	35.31
7	2.07	41.93	39.90
10	1.79	44.03	42.65

Table A 9 Merged syllable recognition results for sports news test

-s	Sentence correct %	Unit correct %	Unit accuracy %
5	1.74	27.83	24.71
7	2.37	28.93	26.86

10	2.96	27.67	26.02

Table A 10 Merged syllable recognition results for literary novel test