Abstract—This paper summarizes our recent efforts for building a Turkish Broadcast News transcription and retrieval system. The agglutinative nature of Turkish leads to a high number of out-of-vocabulary (OOV) words which in turn lower automatic speech recognition (ASR) accuracy. This situation compromises the performance of speech retrieval systems based on ASR output. Therefore using a word based ASR is not adequate for transcribing speech in Turkish. To alleviate this problem, various sub-word based recognition units are utilized. These units solve the OOV problem with moderate size vocabularies and perform even better than a 500 K word vocabulary as far as recognition accuracy is concerned. As a novel approach, the interaction between recognition units, words and sub-words, and discriminative training is explored. Sub-word models benefit from discriminative training more than word models do, especially in the discriminative language modeling framework. For speech retrieval, a spoken term detection system based on automata indexation is utilized. As with transcription, retrieval performance is measured under various schemes incorporating words and sub-words. Best results are obtained using a cascade of word and sub-word indexes together with term-specific thresholding.

Index Terms — speech recognition, spoken term detection, language modeling, discriminative training, morphologically rich languages

I. INTRODUCTION

Turkish, being an agglutinative language with rich morphology, presents a challenge for ASR systems as well as for systems that make use of ASR output, such as speech retrieval systems. The productive morphology of Turkish yields many unique word forms, making it difficult to have a vocabulary with high coverage. Even for vocabulary sizes that would be considered as large for English, the OOV rates for Turkish are quite high. Other morphologically rich languages such as Finnish, Estonian, and Czech also suffer from high OOV rates. Typically, each OOV word causes 1.5 word errors on average [5]. So, high OOV rates directly translate into high word error rates (WERs).

The simplest strategy to decrease the OOV rate is to increase the vocabulary size. However, this also requires more memory and computational power. In addition, including rare words in the vocabulary results in non-robust language model (LM) estimates due to data sparsity. A common solution, especially for agglutinative languages, is to use sub-word units for language modeling. Sub-word units can be obtained by morphological analysis or using statistical methods. Although a morphological analyzer yields sub-words that are linguistically more meaningful, statistical methods have the advantage of not requiring language dependent knowledge.

As with many other machine learning tasks, discriminative training methods have been shown to improve the performance of ASR systems. Both acoustic models (AMs) and LMs have benefited from discriminative techniques that utilize negative examples as well as positive ones. Generating the negative examples for discriminative training makes use of both the AM and the LM. Therefore, discriminative training methods result in a tighter coupling between these models.

As speech recognition technology improves and multimedia archives become more common, providing access to large archives containing spoken content is gaining importance. Speech retrieval integrates ASR and information retrieval (IR) techniques. In addition to spoken document retrieval (SDR) which aims to retrieve news clips related to a query, spoken term detection (STD) which aims to locate occurrences of a term in a spoken archive has recently become an active research topic. Traditionally, in speech retrieval, the ASR component converts speech to text and text retrieval methods are applied on the recognition output. However, when the ASR system does not have high accuracy, this strategy becomes inadequate for speech retrieval. Indexing the alternative ASR hypotheses, as well as the best one, makes the system more robust to recognition errors. However, words that are not in the recognition vocabulary still can not be recognized and located.

As in the case of ASR, a common solution is to use sub-word based language modeling units. Even for English, where OOV rates are usually lower, phonetic indexing has been used to alleviate the effects of OOV queries. However, since phonetic search leads to high false alarm rates, using a hybrid search strategy combining words and subwords is preferred.

In this paper, we compare words and sub-words as language modeling units, and explore the effects of these units on ASR and STD performance. Our investigations include the effects of these units on discriminative training of both the AM and the LM. We find that using sub-word units alleviates the OOV problem and improves the WER with a moderate size vocabulary. We also observed that sub-word units work well with discriminative training methods. Finally, the effect of using sub-word units is more pronounced in the case of STD, since their use is crucial when answering the OOV queries.

The paper is organized as follows: Previous work on related subjects is presented in Section II. In Section III we discuss challenges of Turkish for ASR and introduce the linguistic
tools and data used in this study. We present the baseline language modeling units in Section IV, followed by the discriminatively trained systems in Section V. The methods used for spoken term detection are explained in Section VI. We give our experimental results in Section VII and conclude the paper in Section VIII.

II. PREVIOUS WORK

This paper is an extension of our previous work on building a Turkish Broadcast News (BN) transcription and retrieval system. Previously, we investigated the performance of word and sub-word LMs using various amounts of corpora [1] and introduced lexical stem+ending models [2]. In this paper, we extend our work by experimenting with different representations for sub-word units. In addition, our previous work on discriminative language modeling applied to word hypotheses with morphological features [3] is extended to sub-word hypotheses with sub-word features. On the retrieval side, our initial hybrid STD system that makes use of alternative ASR hypotheses and detection threshold setting [4] is improved to utilize grammatical sub-words on an acoustically diverse corpus. We review the related work in the following sections.

A. Sub-word based ASR

High OOV rates are the main drawback of the word based LMs in ASR of agglutinative and highly inflectional languages. In literature, different sub-word units were explored for those languages to handle this drawback. For Korean, merged morphemes were proposed instead of word-phrases [6]. For Finnish, a statistical method based on the Minimum Description Length (MDL) principle was proposed to split words into sub-word units [7]. Moreover, an optimized sub-word approach was proposed in [8] for Finnish dictation and German street names recognition tasks. Also, a comparison of morpheme based recognition units with words was performed for Czech, which is a highly inflectional language [9]. In [10], stem-ending based language modeling was proposed for agglutinative languages. Recent sub-word approaches applied to Turkish ASR include morpheme based [11], stem-ending based [12], [13], stem based [13], [14], lexical stem-ending based [2] models and a unified model using some of the previous methods together [15]. Post-processing of the sub-word hypotheses using vowel harmony rules [16] and lattice extension technique [17] increased the baseline accuracy. Unsupervised word segmentation with the MDL principle was also applied to Turkish [18]. A detailed comparison of these segments with words for Turkish as well as for other agglutinative languages, revealed promising results [19], [20]. In addition, Turkish speech corpora are becoming publicly available [21].

B. Discriminative Acoustic and Language Modeling

Discriminative training of Hidden Markov Models generally provides significant improvements over the baseline Maximum Likelihood Estimation (MLE) training when a sufficient amount of acoustic data is available. Various criteria, like Maximum Mutual Information (MMI) [22], [23] and Minimum Phone Error (MPE) [24], have been introduced in an effort to represent the discrimination between alternative classes. Extended Baum Welch algorithm [25], [26] is the widely accepted method of optimizing AMs with discriminative criteria and the use of word graphs (lattices) provides an efficient means of collecting discriminative statistics [23], [27], [28].

Discriminatively trained conditional models, such as Conditional Random Fields (CRF) [29], have been successfully applied to language modeling tasks [30], [31], [32]. These sequence modeling approaches have been demonstrated to consistently outperform generative modeling approaches, partly due to improved parameter estimation and partly due to the ease with which many overlapping features can be included in the models. Feature based LMs, for instance, allow for easy integration of relevant knowledge sources such as syntactic and semantic dependencies [31], [33], [34]. In [3] and [35] morphological features as well as word n-grams have been incorporated into word based discriminative language modeling for Turkish and Czech ASR systems respectively.

C. Spoken Term Detection

The NIST STD 2006 Evaluation was an important initiative in STD research [36]. Among several participants, the SRI/OGI system [37] achieved one of the best scores using a word+grapheme system in broadcast news domain and the BBN system [38] achieved the maximum accuracy using word lattices and approximate phonetic transcripts in conversational telephone speech domain.

Several other studies have shown the superiority of combined systems. In [39], cascade of phone and word based indexes was observed to be effective for English. Direct combination of systems, instead of cascading, allows hybrid queries to be searched in both indexes [40]. Word based, phoneme based and particle based (syllable-like data-driven sub-word units) indexes, as well as their hybrids were addressed in [41]. Additional methods, such as query expansion, was applied to combined systems to achieve better scores [42].

Sub-word based approaches appear to be more important for retrieval in agglutinative languages. In [43], [44] and [45] statistical sub-word units were studied in indexing and retrieval for Finnish. Similar methods were applied to Turkish and the combination of word and sub-word indexes with term-specific detection thresholds gave the best results [4].

Using alternative ASR hypotheses, in addition to the best one, increases robustness against recognition errors [4], [39]. These hypotheses can be in the form of lattices or confusion networks. It was reported in [46] that, confusion networks perform better than lattices with smaller size indexes. However no significant difference was observed for Turkish [4].

III. TURKISH: CHARACTERISTICS, DATA AND MORPHOLOGICAL ANALYSIS

A. Characteristics of Turkish

The main characteristic of Turkish is the agglutinative morphology where many new words can be derived from a single stem by addition of several suffixes. There are no prefixes in Turkish. The following examples show concatenated nominal
and verbal inflections. The verbal inflection is more complex than the nominal one.

nominal inflection: ev-im-de-ki-ler-den (among those in my house)

verbal inflection: yap-tir-ma-yabil-iyor-du-k (It was possible that we did not make someone do it)

Although there is not a one to one correspondence between Turkish morphemes and English words, we can say that one Turkish word may correspond to a group of English words. This agglutinative nature causes the vocabulary to expand significantly which is problematic for speech recognition.

Another characteristic of Turkish is the free word order where the order of constituents can be changed without affecting the grammaticality of a sentence. In terms of constituent orders, Turkish can be considered as a Subject-Object-Verb (SOV) type language, however, other constituent orders are also common. Free word order causes data sparseness which leads to non-robust LM estimates.

Vowel harmony is another characteristic of Turkish. According to one of the vowel harmony rules, a stem ending with a back/front vowel takes a suffix starting with a back/front vowel. Vowel harmony is not a problem when word based models are used for speech recognition. However if sub-words are used as language modeling units, we have to take vowel harmony into account since concatenation of sub-words may result in word-like units with incorrect morphophonemics.

B. Data

We have been collecting a large Turkish BN database at Boğaziçi University. In our database, BN programs are recorded daily from a radio channel (VOA) and four different TV channels (CNN Türk, NTV, TRT1 and TRT2). Then these recordings are segmented, transcribed, and verified. The annotation includes topic, speaker and background information. The transcription guidelines were adapted from Hub4 BN transcription guidelines. In this study, we use approximately 194 hours of speech from this database as the acoustic data. This data is partitioned into disjoint training, held-out and test sets. The reference transcriptions of the acoustic data include 1.3 M words. Table I gives the breakdown of the data in terms of acoustic conditions. Here classical Hub4 classes are used: (f0) clean speech, (f1) spontaneous speech, (f2) telephone speech, (f3) background music, (f4) degraded acoustic conditions, and (f5) other.

For building generic LMs for Turkish a general text corpus of 184 M words is collected from the web. This data comes from three major Turkish news portals. After crawling the web and downloading the HTML documents, we extract the text data and apply text normalization and filtering using a morphological parser and some heuristics. The number of words, tokens (words and lexical units such as punctuation marks), and types (distinct tokens) are shown in Table II. After text normalization, the number of word tokens and types reduces to 182.3 M and 1.8 M respectively.

<table>
<thead>
<tr>
<th>Partition</th>
<th>f0</th>
<th>f1</th>
<th>f2</th>
<th>f3</th>
<th>f4</th>
<th>f5</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train</td>
<td>67.2</td>
<td>15.7</td>
<td>8.3</td>
<td>19.8</td>
<td>73.6</td>
<td>3.3</td>
<td>188</td>
</tr>
<tr>
<td>Held-out</td>
<td>1.1</td>
<td>0.1</td>
<td>0.1</td>
<td>0.7</td>
<td>1.3</td>
<td>0.0</td>
<td>3.1</td>
</tr>
<tr>
<td>Test</td>
<td>0.9</td>
<td>0.1</td>
<td>0.1</td>
<td>0.7</td>
<td>1.4</td>
<td>0.1</td>
<td>3.3</td>
</tr>
</tbody>
</table>

C. Morphological Analysis

A morphological parser is required to estimate a LM using morphological units. In this work, we use a morphological parser that was developed for speech recognition and NLP applications [47]. The parser is based on the two-level morphology [48]. The morphophonemic rules and lexicon were adapted from the PC-Kimmo implementation of Kemal Oflazer [49]. A new lexicon of 54,267 root words was compiled based on the Turkish Language Institution dictionary1. The percentage of tokens and types that can be successfully parsed by this parser are indicated in Table II for the web corpus. An example output from the parser for the word “alin” is given in Fig. 1. The English glosses are given in parenthesis for convenience.

\[
alin\text{[Noun]}+[A3sg]+[Pnom]+[Nom] \quad \text{(forehead)}
\]

\[
al\text{[Noun]}+[A3sg]+Hn[P2sg]+[Nom] \quad \text{(your red)}
\]

\[
al\text{[Noun]}+[A3sg]+[Pnom]+[Nom] \quad \text{(of red)}
\]

\[
al\text{[Verb]}-Hn[Verb+Pass]+[Pos]+[Imp]+[A2sg] \quad \text{((you) be taken)}
\]

Fig. 1. Output of the Turkish morphological parser with English glosses. Only 4 out of 8 possible interpretations are given.

The parser output begins with a root and its part-of-speech (PoS) tag in brackets. These are followed by a set of lexical morphemes associated with morphological features (nominal features such as case, person, and number agreement; verbal features such as tense, aspect, modality, and voice information). The capital letters in the lexical morphemes such as H, N and Y are used in two-level morphology to handle some phonetic modifications in the suffixation process. For instance, H in lexical form of the morpheme \( Hn \) can be converted to one of \( i, i, \ddot{u}, \ddot{u} \) in surface form yielding \( \ddot{u}, \ddot{u}, \ddot{u} \), \( \ddot{u} \) respectively due to vowel harmony rule of Turkish or can be omitted in surface form if the preceding word ends with a vowel. The inflectional morphemes start with “+”. The derivational morphemes start with “-” and the first feature

1http://www.tdk.gov.tr
of a derivational morpheme is the PoS of the derived word form. A morphological feature may be appended without any morpheme, indicating that the feature is also applicable to the current word form. The morphological parsing of a word, as shown in Fig. 1, may result in multiple interpretations of that word due to complex morphology. This ambiguity can be resolved using morphological disambiguation tools for Turkish [50], [51], [52].

IV. LANGUAGE MODELING UNITS

In this paper, words, stem+endings, and statistical morphs are utilized as the recognition units. Fig. 2 shows different segmentations of the same Turkish phrase using proposed sub-word units2. Details of these units will be explained in this section.

Words: derneklerinin öncülüğünde
Morphs: dernek lerinin # öncü lüüg nde

Morphemes:

Lex: dernek -lArH -NHn öncü -lHk -SH -NDA
Surf: dernek -leri -nin öncü -lügü -ü -nde

Stem+endings:
Surf: dernek -leri-nin öncü lügü-ü-nde

Fig. 2. Turkish phrase segmented into statistical and grammatical sub-words. The examples showing the lexical and surface form representations of the morphemes are denoted by Lex and Surf abbreviations.

A. Word based Model

Using words as recognition units is a classical approach employed in most state-of-the-art ASR systems. The word model has the advantage of having longer recognition units which results in better acoustic discrimination among vocabulary items. However the vocabulary growth for words is almost unlimited for agglutinative languages and this leads to high OOV rates with moderate size vocabularies. It has been reported that a text corpus with 40 M word tokens results in less than 200 K word types for English and 1.8 M and 1.5 M word types for Finnish and Estonian respectively [53]. The number of word types is 735 K for the same size Turkish corpus.

B. Grammatical sub-words: Morphemes and Stem+endings

In this work, we use the morphological parser outlined in Section III-C to decompose words into grammatical morphemes. To obtain the stem+endings, we first extract the stem from the morphological decomposition and the remaining part of the word is taken as the ending. Stems and surface form endings, (See Fig. 2), are used to generate the LMs. Segmenting the text corpora into stems and endings yields 263.2 M units with 901.2 K distinct stems and 43.7 K distinct surface form endings. The words and endings are respectively composed of 1.7 and 1.5 morphemes on average3. In statistical language modeling, there is a trade-off between using short and long recognition units. Stem+endings are a compromise between words and morphemes. They provide better OOV rate than words, and they lead to more robust LMs than morphemes.

C. Statistically derived sub-words: Morphs

Statistical morphs are morpheme-like units obtained by a data driven approach based on the MDL principle which learns a sub-word lexicon in an unsupervised manner from a training lexicon of words [54]. The main idea is to find an optimal encoding of the data with a concise lexicon and a concise representation of the corpus. The main advantage of this model compared to grammatical models is that it does not require an expert knowledge of the language. Therefore, it can easily be applied to any language.

The Baseline-Morfessor algorithm [54] is used with default settings to automatically segment the word types in the text corpus. For robustness, only the words occurring at least 3 times are used in training, resulting in 50 K morphs. Remaining words are segmented into morphs with the Viterbi algorithm using the initial segmentations. After decoding, morph sequences are converted to word sequences to evaluate the WER. In order to facilitate this conversion, the word boundaries can be marked with a symbol, “#”, as shown in Fig. 2 or a marker, “*”, can be attached to the non-initial morphs like the grammatical morphemes. Since a statistical morph is learned independent of its position in a word, it can occur both as an initial or a non-initial morph. Marking the non-initial morphs results in 76 K morph types and 257 M morph tokens.

V. DISCRIMINATIVE TRAINING

Recent ASR systems utilize discriminative training methods for acoustic and language modeling on top of traditional generative models based on MLE.

A. Discriminative Acoustic Modeling

Discriminative AM training methods aim to maximize posterior model probabilities directly rather than maximizing the likelihood of acoustics as in MLE. MMIE based acoustic training takes alternative hypotheses into account, generally in the form of a lattice, and tries to maximize the mutual information between the reference transcription and the acoustic feature sequence. This is achieved via decreasing the acoustic likelihood of the alternatives available in the recognition lattice while increasing that of reference. Since recognition lattices are products of baseline AM and LM, the performance of discriminative methods may be altered by the chosen units depending on the amount of confusion introduced by the corresponding LM.

2The parser output is simplified by removing PoS tags and morphological features. The distinction between inflectional and derivational morphemes is not taken into account and all morphemes are preceded with “*”.

3Unparsed words are counted as single morphemes.
B. Discriminative Language Modeling

Discriminative language modeling (DLM) is a complementary approach to the existing baseline LM. In contrast to the generative LM, it is trained on acoustic sequences with their transcripts to optimize discriminative objective functions using both positive (reference transcriptions) and negative (recognition errors) examples.

The first step in DLM is to generate the training data which consists of lattices or *N*-best lists. The DLM training data is generated by breaking the acoustic training data into *k* folds, and recognizing the utterances in each fold using the baseline AM (trained on all of the utterances) and an *n*-gram LM trained on the other *k*–1 folds to alleviate over-training of the LMs. AM training data is not typically controlled in the same way since baseline AM training is more expensive and less prone to over-training than *n*-gram LM training [32].

In the DLM training data, each candidate hypothesis is represented as a feature vector. This vector, $\Phi(x, y)$, is defined as a function of the acoustic input, $x$, and the candidate hypothesis, $y$. The first element of the feature vector, $\Phi_0(x, y)$, is the contribution of our baseline system to the DLM and defined as the “log-probability of $y$ in the lattice produced by the baseline recognizer for utterance $x$”. The word *n*-grams [32], morphological relations [3], [35] and syntactic dependencies [31] can be used as the other features. Those features are defined as the number of times a particular *n*-gram, relation or dependency is seen in the candidate hypothesis.

$\alpha$ is the vector of parameters associated with the features. Under this model, the best hypothesis maximizes the inner product of the feature and the parameter vectors, $\Phi(x, y) \cdot \alpha$. The parameters are estimated using the perceptron algorithm which penalizes features associated with the current 1-best hypothesis, and rewards features associated with the gold-standard hypothesis (reference or lowest-WER hypothesis).

VI. SPOKEN TERM DETECTION

Spoken Term Detection aims to locate the occurrences of query terms. Although it is similar to traditional keyword spotting in purpose, it has major differences in method. Keyword spotting requires scanning the whole archive for retrieval, whereas STD eliminates this inefficiency using an index. The index is built offline (before queries are seen) and keeps the necessary information about all terms in the collection. In the retrieval stage, which operates online, the submitted query is located in the previously built index. It is important to note that, queries are directly searched in the index, not in the collection. This makes search time less dependent on database size and provides efficient retrieval. In addition, unlike keyword spotting, queries are not required to be known prior to search.

A. Indexing

Weighted automata indexation is an efficient method for the retrieval of uncertain inputs [55]. In our case, alternative ASR hypotheses, together with their probabilities, are represented as weighted automata. These automata are processed to extract all of the possible substrings that are contained in the automata. In this process the automata are turned into transducers where the inputs are the original labels of the automata and the outputs are the utterance numbers. Next, these transducers are combined by taking their union. The final transducer is optimized using weighted transducer determination, resulting in optimal search complexity — linear in the length of the input string. The weights in the index transducer correspond to expected counts.

B. Retrieval

The queries presented to the system are also represented as finite state automata and the search is performed by composing these automata with the index transducer. The output contains the list of all utterances containing the query and the corresponding expected counts. The utterances are ranked using the expected counts and those exceeding a threshold are selected [55]. In case of sub-word units, the retrieval strategy is modified to retrieve an occurrence if no suffix is seen after the query. After obtaining the utterance indices, the exact location of each term is determined by forced alignment [4].

It is important to note that, by varying the threshold on the expected count, different operating points can be obtained and the system can be tuned for various applications. For instance, it would be useful to set a high threshold in a sign language tutoring tool since a few correct samples are adequate to observe the sign [56]. On the other hand, monitoring applications may require to detect all occurrences. In this case, setting a smaller threshold reduces the miss rate.

Setting a global threshold for all query terms is the simplest method. A recently proposed approach identifies term-specific thresholds [38], which optimize a metric defined in NIST STD 2006 Evaluation.

VII. EXPERIMENTAL RESULTS

A. Baseline ASR Systems

Baseline AM and LMs were built using the Turkish acoustic and text databases explained in Section III-B. AMs were speaker-independent and they utilized no special adaptation technique. We used decision-free state clustered cross-word triphone models with 10843 HMM states. Since Turkish is almost a phonetic language, graphemes were utilized instead of phonemes. Each HMM state had a GMM with 11 mixture components. The silence model had a GMM with 23 mixtures. The HTK” frontend was used to get the MFCC based acoustic features. The training and decoding tasks were performed using the AT&T tools.

Generative LMs with interpolated Kneser-Ney smoothing as well as entropy based pruning [57] were built using the SRILM toolkit. In order to reduce the effect of pruning on the recognition accuracy, the first-pass lattice outputs were re-scored with unpruned LMs. LMs built with the web corpus and the reference transcriptions were linearly interpolated to reduce the effect of out-of-domain data. The interpolation

4http://htk.eng.cam.ac.uk/
5http://www.speech.sri.com/projects/srilm/
6http://www.research.att.com/~fsmtools/{fsm, dcd}
constant and the $n$-gram order for each model were optimized over the held-out data.

ASR experiments for various units were performed using the same AM and the same real-time factor (RTF $\approx 1.5$). The lexicon sizes, average unit length in each lexicon (AUL), units per word (UPW) and the WERs are given in Table III. For the word based model, the effect of the OOV rate was investigated by changing the vocabulary size. We conducted experiments with large (50 K, 76 K) and very large (200 K, 300 K, 500 K) vocabulary sizes. The aim of using 50 K and 76 K vocabularies was to compare the performances of word and statistical sub-word systems having the same vocabulary sizes. The best results are obtained with the 3-gram LMs. Increasing the vocabulary size from 50 K to 500 K increases the coverage by 7.3% and decreases the WER by 5.7% on the test set. The coverage is calculated over the word tokens. Significant relative improvements at the level of $p < 0.001$ as measured by the NIST MAPSSWE significance test are obtained with the increased vocabulary size until 200 K. The improvement from 200 K to 300 K vocabulary is statistically significant at $p < 0.01$. No significant improvement is obtained with 500 K vocabulary compared to 300 K vocabulary. In the further experiments, 200 K system was used as the baseline-word model to balance the trade-off between recognition performance and LM complexity.

As a consequence of the complex Turkish morphology, a word can yield several morphological analyses as given in Fig. 1. Removing the morphological ambiguity can be crucial in dependency parsing and word sense disambiguation. However, the effect of this ambiguity on the WER has not been investigated yet. In this research, we could not disambiguate the multiple analyses of a word since the parser outputs used in our experiments are not compatible with the input expected by the available disambiguation tools. Therefore, we first experimented with language modeling without disambiguation. Here each parse was assumed to be equally likely and a LM was built using all the ambiguous parses. Second, we chose the parse with the least number of morphemes and built the LM with the selected parses. Finally, random parses for each word token and type were selected for language modeling. In Table IV, the WER results obtained with the stem+ending LMs estimated from only the reference transcriptions of the acoustic data are given. No significant difference is obtained between the first two approaches and they are significantly better than random. Due to its simplicity, we used the second approach in the stem+ending experiments.

In the stem+ending model, the most frequent 76 K and 200 K units out of 945 K surface form stem and ending types were utilized as the vocabulary items. The reason for selecting these vocabulary sizes was to make the stem+ending system comparable with the morph and the word systems in terms of vocabulary size. The selected vocabularies cover 99.6% and 99.8% of the word tokens on the test set$^7$ and reduce the WER to 23.2% and 23.1% respectively. Since no significant gain is obtained by increasing the vocabulary size, the system with 76 K vocabulary was used as the baseline model.

In the morph based model, we explored four different ways of converting the morph sequences to word-like units:

1) Use a word boundary (WB) morph, e.g. “dernek lerinin # öncü lügünü nde”
2) Use no WB morph in the first-pass recognition. Perform re-scoring with a LM containing WBs, e.g. “dernek lerinin öncü lügünü nde”
3) Mark non-initial morphs with “.”, e.g. “dernek -lerinin öncü -lüğü -nde”
4) Concatenate non-initial morphs and mark with “.”, e.g. “dernek -lerinin öncü -lüğü -nde”

In the first scenario, all the morph types (50 K morphs) and the WB morph were utilized as the vocabulary items. In the text corpora, the ratio of morph tokens to word tokens is calculated as 2.4 including the WB symbol. This suggests higher order $n$-gram LMs and the best result is obtained with the 5-gram LM. In the second scenario, first-pass recognition was performed with a 4-gram morph LM built without the WB symbol. Then optional WB symbols were added to every node in the lattice and this lattice was re-scored with a 5-gram LM.

---

TABLE III

<table>
<thead>
<tr>
<th>Recognition Units</th>
<th>Lexicon Size</th>
<th>UPW</th>
<th>n-gram</th>
<th>Coverage (%)</th>
<th>WER (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Words</td>
<td>50 K</td>
<td>9.4</td>
<td>3-gram</td>
<td>92.7</td>
<td>29.9</td>
</tr>
<tr>
<td></td>
<td>76 K</td>
<td>9.7</td>
<td>3-gram</td>
<td>94.9</td>
<td>27.7</td>
</tr>
<tr>
<td></td>
<td>200 K</td>
<td>10.4</td>
<td>3-gram</td>
<td>98.0</td>
<td>25.5</td>
</tr>
<tr>
<td></td>
<td>300 K</td>
<td>10.6</td>
<td>3-gram</td>
<td>98.7</td>
<td>25.1</td>
</tr>
<tr>
<td></td>
<td>500 K</td>
<td>10.9</td>
<td>3-gram</td>
<td>99.1</td>
<td>25.1</td>
</tr>
</tbody>
</table>

**TABLE IV**

<table>
<thead>
<tr>
<th>Methods</th>
<th>WER (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiple parses</td>
<td>35.2</td>
</tr>
<tr>
<td>Parses with the least number of morphemes</td>
<td>35.9</td>
</tr>
<tr>
<td>Random parses for each word token</td>
<td>33.8</td>
</tr>
<tr>
<td>Random parses for each word type</td>
<td>36.6</td>
</tr>
</tbody>
</table>

---

$^7$When calculating word coverage using sub-word units, a word is considered as an OOV word if it cannot be generated by any combination of a stem and an ending or an initial unit followed by non-initial units.
B. Discriminative Acoustic Modeling

The morph based model yields significant improvements (is significantly better than 50 K, 76 K and 200 K word models. obtained on the test set with the baseline stem+ending model model in terms of the recognition accuracy. The improvement with smaller vocabulary sizes and outperform the word based accuracy of the IV words. Since the number of IV words is pass or re-scoring degrade the accuracy of the in-vocabular scenarios); 2) Scenarios that use the WB morph in the first-to 77% for stem+ending and to 71% on average for morph approaches reduce the WERs for OOV words (from 100% approach. This analysis revealed that 1) All the sub-word accuracy significantly. The best result is obtained with the

<table>
<thead>
<tr>
<th>Acoustic Models</th>
<th>200 K Word LM</th>
<th>76 K Morph LM</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLE (baseline)</td>
<td>25.5</td>
<td>24.0</td>
</tr>
<tr>
<td>MMIE (with 200 K word LMs)</td>
<td>24.3</td>
<td>22.8</td>
</tr>
<tr>
<td>MMIE (with 76 K morph LMs)</td>
<td>24.5</td>
<td>22.6</td>
</tr>
</tbody>
</table>

In this paper, the effectiveness of DLM was investigated on word based and sub-word based approaches. Here baseline word (200 K) and morph (76 K, non-initials marked with “-”) models were used as the generative systems. To generate the word and morph training data and to train the feature parameters, we followed the findings of the previous research on DLM [3], [32]: 1) Only the LM training was controlled via 12-fold cross validation; 2) 50-best lists were utilized instead of lattices as the DLM training data; 3) Oracle best path was used as the gold standard (50-best oracle error rates are 13.6% for words and 12.6% for morphs); 4) The perceptron algorithm (as presented in [32]) was utilized for training; 5) Averaged perceptron parameters were used for testing.

During training, the weight of the baseline model (the weight associated with \( \Phi_0(x, y) \)) was not updated. Therefore, 10 different models were trained with \( \alpha_0 \) constants from 0 to 20. On the held-out set, \( \alpha_0 = 2 \) and \( \alpha_0 = 4 \) yield the best models for the word and morph DLM experiments respectively. These models converge in 1-3 iterations of the perceptron algorithm. Word and morph unigram and bigram features were utilized in the experiments with word and morph DLM data. In order to use word \( n \)-gram features in morph hypotheses, morph sequences were concatenated to obtain word-like units. In the generation of the morph \( n \)-gram features from word hypotheses, words were segmented into the morphs using the Viterbi algorithm. Examples of word and morph bigram features are:

\[
\Phi_j(\text{word})(x, y) = \text{Number of times "derneklerinin öncülüğünde" is seen in } y
\]

\[
\Phi_k(\text{morph})(x, y) = \text{Number of times "dernek -lerinin" is seen in } y
\]

The results for the DLM experiments are given in Table VI where “Feats” represents the number of features extracted from the 50-best lists and “Feats (Act.)” represents the percentage of the features with non-zero weights after the parameter training. In contrast to [3], we do not get any significant improvements with DLM over the baseline word model (200 K). This can be due to increased amount of data and larger vocabulary size. However, DLMs on the 50-best output of the morph based system give 0.5-0.7% absolute improvements over the baseline model. These improvements are statistically significant at \( p < 0.001 \). For the morph based system, the same improvement is obtained whether the DLM parameters are trained to optimize the Morph Error Rate or the WER. In addition, neither incorporating trigrams into the feature sets nor increasing the \( N \)-best size to 1000 give further improvements on the reported results which is consistent with the results reported in [3], [35].

D. Spoken Term Detection

In this part, we employed several language modeling units for STD and compared their performances. The effect of using multiple indexes and setting term-specific thresholds was also investigated.

Our evaluation is based on the NIST STD 2006 Evaluation Plan [36]. System performance is measured using the Spoken
Term Detection Evaluation Toolkit (STDEval). The fundamental metric is Term-Weighted Value (TWV) defined as:

$$TWV(\theta) = 1 - \frac{1}{Q} \sum_{k=1}^{Q} \left\{ P_{miss}(q_k, \theta) + \beta.P_{FA}(q_k, \theta) \right\}$$ (1)

where $\theta$ is the detection threshold and $\beta$ is a user defined parameter to adjust the trade-off between miss and false alarm probabilities ($P_{miss}$ and $P_{FA}$). Maximum Term-Weighted Value (MTWV) is the maximum of TWV found over the range of all possible $\theta$ values. The query set was constructed by the NIST Term Selection Tool. This software randomly selects terms based on manual transcriptions. Our query set consisted of 1627 queries, including single and multi-word terms as well as foreign names and acronyms.

Based on the ASR results given in Table III, we experimented with baseline word (200 K), stem+ending (76 K) and morph (76 K, non-initials marked with “-”) models. The results are summarized in Table VII. The third column shows the percentage of OOV queries in the query set. An OOV query includes one or more OOV units; as a result it can be neither recognized by the ASR system nor detected by the STD system. The STD performance is measured over three subsets of the query set: all queries, IV queries and OOV queries.

As presented in Table VII, sub-word based indexes yield better scores than the word based indexes. Sub-words are quite helpful in detecting OOV queries, with a cost of reduced performance over IV queries. However, the decrease is very small compared to the gain over the OOV set. This gain is strongly dependent on the OOV query percentage. Overall, morphs give the best scores.

Word based index is the best performer in case of IV queries and morph based index is the best in case of OOV queries. To utilize the advantages of both, we cascaded the word based and morph based indexes as follows. First, the query was searched in the word based index. If no results were returned, it was searched in the morph based index [39]. As can be seen in Figure 3 and Table VII, the cascade of word and morph based indexes is superior to both of the individual indexes.

In the last experiment, term-specific thresholds were used in detection and compared with global thresholds. As proposed in [38], term-specific thresholds were chosen to maximize the TWV metric. As shown in Table VII, using term-specific thresholds provides 5-7% absolute increase in MTWV for all indexing schemes.

Applying a specific threshold for each term, a single operating point can be obtained. To obtain different operating points, we changed the $\beta$ parameter in Equation 1. Incrementing $\beta$, the penalty of false alarms is increased and higher precision points can be obtained. DET curves for term-specific thresholding (indicated by +TermTh) are shown in Fig. 3 together with those for global thresholding.

In conclusion, sub-word units are observed to be extremely useful in locating the OOV queries. Term-specific thresholding, on the other hand, improves the MTWV results over IV queries. Therefore, the best scores are achieved with a word+morph cascaded index and term-specific thresholding in detection. With the application of these methods, the MTWV increases from 65.1% (Words) to 79.4% (Word+Morph Cascade with term-specific thresholding).

VIII. CONCLUSION

This paper presents ASR challenges in Turkish and our efforts to overcome them. The main focus is on the investigation of positive contributions introduced by incorporating
sub-word units, whether they be towards better acoustic and language modeling or towards higher precision and recall rates in retrieval tasks. We introduce discriminative models employing sub-word units and compare their performance with baseline solutions. In case of STD, we combine sub-word and word based indexing to cope with the high OOV rate and utilize term-specific thresholds to improve the MTTW metric.

Baseline experiments bring into sharp relief the superiority of sub-word units in modeling Turkish language. Sub-words solve the OOV problem introduced by the agglutinative language characteristics and result in lower WERs. Statistical sub-word approach is significantly better than the word based approach even for very large vocabulary sizes.

Further, the interaction between recognition units, words and statistical morphs, and discriminative training is explored. Discriminative models trained with morph recognition lattices outperform their counterparts constructed from word lattices. The effect of using sub-word lattices in discriminative training is more pronounced in the language modeling framework while still leaving room for further improvements with the complementary feature sets.

Finally, employing the morph based indexes in STD makes it possible to locate the OOV queries. With this contribution, morph approach achieves the highest individual score. Hybrid indexes further improve the STD performance by alleviating the OOV problem of word based indexes and the high false alarm rate of sub-word based indexes. Independent of the OOV problem and WER, setting term-specific thresholds in detection provides further improvement.

ACKNOWLEDGMENT

The authors would like to thank AT&T Labs – Research for the software and Brian Roark for useful discussions about DLM and his help with the coding of the perceptron algorithm.

REFERENCES


Ebru Arsoy received her B.S. and M.S. degrees from the Electrical and Electronics Engineering Department at Bogazici University in 2002 and 2004 respectively. She has been pursuing her Ph.D. degree in the same department since 2004. In her Ph.D. thesis, she is currently working on statistical language modeling for LVCSR of Turkish. Her main research interests include automatic speech recognition and statistical language modeling for agglutinative languages. She is currently a board member of the ISCA Student Advisory Committee.

Dogan Can received his B.S. degree in 2006 from the Electrical and Electronics Engineering Department at Bogazici University. He was a graduate student of the Electrical, Computer and Systems Engineering Department at Rensselaer Polytechnic Institute during the year following his graduation. He has been pursuing his M.S. degree at the Electrical and Electronics Engineering Department of Bogazici University since 2007. Currently he is working on discriminative methods for speech recognition and spoken term detection.

Siddika Parlak received her B.S. degree from the Electrical and Electronics Engineering Department at Bogazici University in 2006. She is currently an M.S. student at the same department and research assistant in BUSIM (Bogazici University Signal and Image Processing) Laboratory. Her main interests are in the field of Speech and Language Processing, focusing on Automatic Speech Recognition and Spoken Information Retrieval.

Hasim Sak received his B.S. degree in 2000 from the Computer Engineering department at Bilkent University, and M.S. degree in 2004 from the Computer Engineering department at Bogazici University, where he is now pursuing his Ph.D. degree. On his thesis, he is focusing on the language modeling and speech decoding challenges associated with agglutinative languages and rich morphology. From 2000 to 2005, he worked in a company developing speech technologies for Turkish. His main research interests include speech recognition, speech synthesis, statistical language modeling, morphological parsing, spelling correction, and morphological disambiguation.

Murat Saraclar received his B.S degree in 1994 from the Electrical and Electronics Engineering Department at Bilkent University, M.S.E. degree in 1997 and Ph.D. degree in 2001 from the Electrical and Computer Engineering Department at the Johns Hopkins University. He is currently an assistant professor at the Electrical and Electronic Engineering Department of Bogazici University. From 2000 to 2005, he was with the multimedia services department at the AT&T Labs – Research. His main research interests include all aspects of speech recognition, its applications, as well as related fields such as speech and language processing, human-computer interaction and machine learning. Dr. Saraclar is currently a member of the IEEE Signal Processing Society Speech and Language Technical Committee.


