

# Language Modeling for Automatic Turkish Broadcast News Transcription

*Ebru Arısoy, Haşim Sak, and Murat Saraçlar*

Electrical and Electronic Engineering Department  
Boğaziçi University, 34342 Bebek, Istanbul, Turkey  
{arisoeyeb, hasim.sak, murat.saraclar}@boun.edu.tr

## Abstract

The aim of this study is to develop a speech recognition system for Turkish broadcast news. State-of-the-art speech recognition systems utilize statistical models. A large amount of data is required to reliably estimate these models. For this study, a large Turkish Broadcast News database, consisting of the speech signal and corresponding transcriptions, is being collected. In this paper, information about this database and experiments performed using the system developed on the collected data are presented. In addition to the baseline system, various sub-word language models are investigated. Lexical stem-endings are proposed as a novel unit for language modeling and are shown to perform better than surface stem-endings and morphs. Currently, our best systems have lower than 20% error on clean speech.

**Index Terms:** speech recognition, agglutinative languages, sub-word units, language modeling.

## 1. Introduction

Turkish is an agglutinative language with a productive inflectional and derivational morphology. Many new word forms can be derived from a single stem by concatenation of suffixes. Although there is not a one to one correlation between Turkish suffixes and English words, one Turkish word may correspond to a group of English words. Therefore, using an English speech recognition setup for an agglutinative language does not yield a comparable performance. Agglutinative languages need a huge vocabulary to achieve the same coverage with an English system. In our system, Out-of-Vocabulary (OOV) rate is 9.3% for a 50K Turkish lexicon. Also for other agglutinative languages, Finnish and Estonian, OOV rates are around 15% for 69K lexicon and 10% for 60K lexicon respectively [1, 2]. However, for English, OOV rates are less than 1% for similar vocabulary sizes. On the one hand, using a moderate vocabulary size results in large number of OOV words and on the other hand, using a huge vocabulary size suffers from non-robust language model estimates. So sub-word units are essential to alleviate the data sparseness and OOV problem in agglutinative languages. Stems and morphemes or their groupings are natural choices since they are capable of handling the OOV problem with smaller units which are still meaningful for speech recognition. Therefore, sub-word units like stem-endings, grammatical morphemes, and statistical morphs are proposed for those languages [3, 4, 5].

One of the drawbacks of sub-word units is that they may result in non-grammatical items. Although statistical morphs give better performance than words for Turkish, Finnish, and Estonian [2], they result in over generation. In [6], it was shown that a simple mapping from concatenated morph sequences to grammatically correct words gives 1.1% absolute improvement over

the baseline. In addition, post-processing of sub-word recognition output for Turkish with vowel harmony rules, gave a significant improvement [7].

In previous studies, using sub-word units in speech recognition was investigated with surface form realizations. However, same stems and morphemes in lexical form may have different phonetic realizations due to some phenomena such as gemination and vowel harmony. If we consider these two words; *evler* (houses) and *kitaplar* (books), they are decomposed into their morphemes as *ev-ler* and *kitap-lar* respectively. Although, both of the words have the same morpheme in the lexical form, the vowel of the plural morpheme is modified according to the last preceding vowel of the stem during the suffixation process. Therefore, sub-word units in surface form may reveal many units that correspond to the same stem or morpheme groups in lexical form. In this work, in addition to surface form realizations of all proposed units, we also experimented with stem plus lexical morpheme ending model. The problem with using morphemes in lexical form in n-gram based models is that they are relatively small and we have to use longer unit histories to cover enough word histories to estimate an accurate model.

The main difference of this study from previous sub-word approaches are: i) all proposed techniques are compared in the same and large acoustic and text database and ii) in addition to surface form representations, lexical form sub-word recognition units are used in recognition. Lexical to surface form mapping ensures correct surface form alternations and gives better performance than recognition with surface form representations.

This paper is organized as follows: In the next section details of the data collection and the statistics of the data are given. Section 3 describes the morphological parser and Section 4 describes the word and sub-word systems. Section 5 explains the experiments and discusses the results, and finally Section 6 concludes the paper.

## 2. Data Collection

We have been collecting Turkish Broadcast News database in Boğaziçi University. In our database, Broadcast News programs are recorded daily from four different TV and a radio channel. Next, these recordings are screened for content and audio quality and the remaining are segmented, transcribed, and verified.

The recordings are first processed by an automatic segmentation program<sup>1</sup>. The output is manually corrected and topic, speaker and background information is added. The segments to be transcribed are also decided at this stage. The open source Transcriber<sup>2</sup> program is used for manual segmentation and an-

<sup>1</sup>SESTEK: <http://www.sestek.com.tr>

<sup>2</sup>Transcriber: <http://trans.sourceforge.net>

notation. This program is also used during the transcription process. The transcription guidelines were adapted from Hub4 Broadcast News transcription guidelines.

Once the data is processed, the acoustic data is converted to 16kHz 16-bit PCM WAV format, and segmentation, speaker and text information is converted to the NIST STM format.

### 2.1. Statistics of the data

Approximately 100 hours of data has been collected for this research. Out of this, 71 hours were selected for transcription. This data was partitioned into training (68.6 hours) and test (2.5 hours) sets. The training data does not overlap with the data in terms of selected dates.

Table 1 gives the breakdown of 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. Only 38% of the data is marked as clean.

Among the recorded programs is a program called “news for the hearing impaired” which contains sign language videos as well as clearly articulated slow speech. The data from this program is 7.7 hours in duration. We mark this data as HI in the experimental results and analysis.

Table 1: Amount of data for various acoustic conditions (in hours)

Partition	f0	f1	f2	f3	f4	fx	Total
Train	25.9	7.0	1.8	6.2	26.4	1.3	68.6
Test	1.3	0.1	0.1	0.2	0.8	0.03	2.5

## 3. Morphological Analysis

To estimate a language model using morphological units, we need a morphological parser. In this work, we used a morphological parser that one of the authors is currently developing for speech recognition and other NLP applications. The parser is based on the two level morphology [8]. The morphophonemic rules and lexicon have been adapted from PC-Kimmo implementation of Kemal Oflazer [9]. The implementation currently uses the AT&T FSM library for finite-state operations [10]. An example output from the parser for the word *çocukları* is given below. The English gloss is given in parenthesis for convenience.

```

çocuk+Noun-lAr+A3pl-SH+P3sg+Nom (her/his children)
çocuk+Noun-lAr+A3pl-SH+P3pl+Nom (their children)
çocuk+Noun-lAr+A3pl+Pnon-YH+Acc (the children)
çocuk+Noun+A3sg-lArH+P3pl+Nom (their child)

```

In the analysis, the first part is the stem, *çocuk* (child), and morphological features like Noun is shown starting with a plus sign. The morphemes like lAr (plural) start with minus sign. The capital letters in the morphemes such as A, S and H are used in two-level morphology to handle some phonetic modifications in suffixation process. For instance, A in lexical form of the morpheme lAr can be converted to a or e in surface form yielding lar or ler due to vowel harmony rule of Turkish. In this work, we did not use the morphological features, so we have three different parse results for this example.

```

çocuk-lAr-SH
çocuk-lAr-YH
çocuk-lArH

```

The morphological parsing of a word as shown in the example above may result in multiple interpretations of that word due to complex morphology. Although morphological disambiguation is required for Turkish to find the correct morphological parse of a word in a given context using a system such as in [11], we selected the parse with the minimum number of morphemes in this work.

We used a morphophonemic finite state transducer encoding two-level rules for Turkish phonology to convert the lexical forms to surface forms to get the pronunciations of the units in the language model.

## 4. Language Modeling

Various language models are built with word and sub-word recognition units. Statistical morphs, grammatical morphemes and stem-endings with both surface and lexical form representations are used as sub-word units. The example below shows the same sentence segmented into possible recognition units. In order to easily recover words from sub-word sequences, a word boundary symbol, #, is added between each sub-word unit. Last line of the example shows the lexical stem-ending model. We do not remove the morpheme boundaries for the morphophonemic transducer.

```

Words: kesildiği andan itibaren
Morphs: kesil diği # a ndan # itibar en
Morphemes: kes il diğ i # an dan # itibaren
Stem-endings:
Surf: kes ildiği # an dan # itibaren
Lex: kes -Hl-DHk-SH # an -DAn # itibaren

```

### 4.1. Words

Using words as recognition units is a classical approach employed in most state-of-the-art recognition systems. These systems require limited vocabularies and it is not possible to use all the words as vocabulary items. Therefore, most frequent words are selected to balance the trade-off between OOV rate and system complexity. In our system, using a 96.4M words corpus, approximately 1.4M unique words are generated. The most frequent 50K words are added to the recognition lexicon. This gives an OOV rate of 9.3% over the test data.

### 4.2. Statistical morphs

The morph model is a statistical approach where a recursive Minimum Description Length (MDL) algorithm learns a sub-word lexicon in an unsupervised manner from a training lexicon of words [12]. Using the same 96.4M words text corpus, 34.7K morph types are generated. In order to eliminate huge number of morph types, rare words (occurring less than 5 times) are not used for morph generation. Later, they are separated into morphs or letter sequences with the Viterbi algorithm using the generated morph types.

### 4.3. Grammatical morphemes

We also tried using grammatical morphemes to compare their performance with statistical ones. For the generation of grammatical morphemes, we used our morphological parser containing 28602 stems. In order to obtain a full coverage with the same vocabulary size of morphs using grammatical morphemes, we split unparsed rare words into their letters. Since our parser does not have a huge lexicon of stems, unparsed words occurring less than 75 times in the corpus are split into their letters

and 34.7K morpheme types are obtained from the same corpus. However, only 46% of the word types in our corpus are split into their morphemes and this technique introduced too many letter sequences which are not meaningful as recognition units.

#### 4.4. Stem-endings

In this model, the same morphological parser is used to obtain the stems and endings. Both surface and lexical form realizations of endings are used to generate language models. The advantage of lexical forms to surface forms is that surface form may reveal many units that correspond to the same stem or morpheme groups in lexical form. Therefore, in statistical language models, lexical forms can capture the suffixation process better. Also, in lexical to surface mapping, compatibility of vowels in terms of vowel harmony is enforced. However, note that using the stems followed by endings, the language model can produce morphotactically invalid sequences and the recognition results may also contain invalid sequences. In the case of lexical stem-endings, we produce the surface forms of these invalid sequences using the morphophonemic regardless of morphotactics. In order to make this model comparable with the word based model, we select the most frequent 50K units from the corpus. For the surface stem-ending model, this corresponds to the most frequent 40.4K roots and 9.6K endings. The OOV rates in terms of words are around 2% for training and 2.5% for the test data. For the lexical stem-ending model, the most frequent 45K roots and 5K lexical endings are enrolled in the lexicon. The word OOV rates are 1.7% and 2.2% for training and test data. The advantage of these units compared to the other sub-words is that we have longer recognition units with an acceptable OOV rate.

## 5. Experiments

In this section, recognition results with the baseline and channel adapted acoustic models are given for different language modelling units in terms of Word Error Rate (WER).

### 5.1. Experimental Setup

In this research, we have used a text corpus of 96.4M words which was collected from the web. The corpus contains text from various sources: online books, newspapers, journals, magazines, etc. For the acoustic models, we used Broadcast News data with acoustic signals and their transcriptions. The details of the acoustic data is given in Section 2. Statistical language models are generated using SRI Language Modelling toolkit [13]. The recognition tasks are performed using the AT&T Decoder [10]. We use decision-tree state clustered cross-word triphone models with approximately 7500 HMM states. Instead of using letter to phoneme rules, the acoustic models are based directly on letters. Each state of the speaker independent HMMs has a GMM with 11 mixture components. The baseline acoustic models are adapted to each TV/Radio channel using supervised MAP adaptation on the training data, giving us the channel adapted acoustic models.

In order to reduce the effect of out-of-domain data in language modelling, transcriptions of acoustic training data are used in addition to the text corpus. A simple linear interpolation approach is applied for domain adaptation. This gives an absolute reduction of 2.8% in the error rate for word models. In all of the experiments, adapted language models are generated with the same interpolation constant and same acoustic models are used.

For the word-based model, 3-gram language models give the best performance. However, for sub-word-based approaches higher order n-grams are required to track the n-gram word statistics and this results in more complicated language models. To be able to handle computational limitations, entropy-based pruning [14] is applied to all the language models.

### 5.2. Results

The recognition results for the experiments are given in Table 2. In this study, experiments were performed without any time constraints.

Four different experiment scenarios were performed. First is the baseline recognition experiment for each recognition unit with the same acoustic model and unit specific language models. The size of the language models is set with entropy-based pruning according to the computational limitations.

The second scenario is the re-scoring strategy, which is only applied to sub-word units since pruning effects sub-word models more than words, and previous experiments showed that re-scoring does not improve the results for word models. In this framework, lattice output of the recognizer is rescored with a same order n-gram language model pruned with a smaller pruning constant. Experiments with rescoring are labeled with “\_rescore” in the table.

As was mentioned in Section 2, acoustic training data are collected from 4 different channels. Since the acoustic conditions of the channels are different from each other, acoustic models are adapted for each channel. Supervised MAP technique is used for channel adaptation. Experiments with the channel adapted acoustic models are labeled with “\_map\_sup” in the table.

Not included in the table is the experiment with grammatical morphemes. This model yields 43% WER with baseline acoustic models with lattice rescoring. Since this is worse than the word model, no further experiments were performed.

For the stem-ending models, since the output may contain invalid sequences, a simple restriction is applied. It is not possible to handle those sequences without a detailed grammatical and semantic analysis. However, our aim in this restriction is just to enforce the decoder not to generate consecutive ending sequences. This restriction is implemented as a finite-state acceptor that is intersected with the lattices. This restriction does not yield any improvement for the surface stem-ending model. However, for the lexical stem-ending model, the restriction decreases the error rate by 0.1-0.2% for the baseline and adapted acoustic models respectively. For this experiment, re-scoring results are reported after composing with the restriction acceptor.

In Table 2, the first group of rows shows the experiments using the baseline acoustic models and the second group of rows shows the experiments with channel adapted acoustic models. It is clear that adapted acoustic models perform better than baseline acoustic models. However, the improvement is most pronounced for telephone speech. This is expected since we do not yet have special acoustic models for telephone speech in our system.

As can be seen in the table, the best results are obtained by the lexical stem-ending models. The WER rate is improved by 0.8% over the previous best model using statistical morphs. The improvement is statistically significant at  $p=0.02$  as measured by the NIST MAPSSWE significance test.

Table 2: Recognition Results

Experiments	f0	f1	f2	f3	f4	fx	Avg.	HI
Words	27.7	66.9	75.4	45.9	49.4	76.2	41.4	20.4
Morphs_resc	22.4	67.4	76.9	41.6	47.2	75.6	37.9	16.6
Stem-ending_rescore	24.7	63.4	75.9	41.8	47.3	78.0	38.8	15.9
Stem-ending-lexical_rescore	21.1	66.7	74.7	40.8	47.0	74.4	37.0	13.0
Words_map_sup	26.3	65.1	64.4	43.2	48.4	77.4	39.6	20.8
Morphs_map_sup_rescore	19.9	63.1	63.2	38.3	46.6	74.4	35.4	13.1
Stem-ending_map_sup_rescore	23.1	62.2	62.2	37.9	45.8	75.0	36.5	15.5
Stem-ending-lexical_map_sup_rescore	19.4	64.0	60.7	36.2	45.8	75.6	34.6	12.4

## 6. Conclusions

In this research, a Turkish Broadcast News transcription system is developed. To be able to generate reliable acoustic and language models, broadcast news programs are collected and transcribed by untrained native speakers according to the guidelines. In addition to the baseline word model, sub-word approaches like statistical morphs, stem-endings are used as language modelling units. Those units solve the data sparseness and large number of OOV problem caused by the agglutinative structure of Turkish with a reasonable vocabulary size. As a novel approach, lexical representations of endings are introduced in the stem-ending model. Since, vowel harmony determines the rule in suffixation process, lexical stem-ending model reduces the number of endings significantly and improves language model estimation. This approach is shown to improve the system performance over our previous best approach using statistical morphs.

## 7. Acknowledgements

The authors would like to thank Sabanci and ODTU universities for the Turkish text data and AT&T Labs – Research for the software. This research is partially supported by TÜBİTAK (Scientific and Technological Research Council of Turkey) (Project code: 105E102) and Boğaziçi University Research Fund (Project code: 05HA202). The first author is supported by SIMILAR Network of Excellence and TÜBİTAK BDP. The second author is supported by TÜBİTAK BİDEB.

## 8. References

- [1] T. Hirsimäki, M. Creutz, V. Siivola, M. Kurimo, S. Virpioja, and J. Pyllkkönen, “Unlimited vocabulary speech recognition with morph language models applied to Finnish,” *Computer, Speech and Language, Elsevier*, vol. 20, no. 4, pp. 515–541, 2006.
- [2] M. Kurimo, A. Puurula, E. Arisoy, V. Siivola, T. Hirsimäki, J. Pyllkkönen, T. Alumäe, and M. Saraçlar, “Unlimited vocabulary speech recognition for agglutinative languages,” in *Proc. HLT-NAACL 2006*, New York, USA, 2006.
- [3] O.-W. Kwon and J. Park, “Korean large vocabulary continuous speech recognition with morpheme-based recognition units,” *Speech Communication*, vol. 39, pp. 287–300, 2003.
- [4] V. Siivola, T. Hirsimäki, M. Creutz, and M. Kurimo, “Unlimited vocabulary speech recognition based on morphs discovered in an unsupervised manner,” in *Proc. EUROSPEECH 2003*, Geneva, Switzerland, 2003, pp. 2293–2296.
- [5] J. Kneissler and D. Klakow, “Speech recognition for huge vocabularies by using optimized sub-word units,” in *Proc. EUROSPEECH 2001*, Aalborg, Denmark, 2001, pp. 69–73.
- [6] E. Arisoy and M. Saraçlar, “Lattice extension and rescoring based approaches for LVCSR of Turkish,” in *International Conference on Spoken Language Processing - Interspeech2006 ICSLP*, Pittsburg PA, USA, 2006.
- [7] H. Erdogan, O. Buyuk, and K. Oflazer, “Incorporating language constraints in sub-word based speech recognition,” in *Proc. ASRU 2005*, Cancun, Mexico, 2005.
- [8] K. Koskenniemi, “A general computational model for word-form recognition and production,” in *Proceedings of the 22nd annual meeting on Association for Computational Linguistics*, 1984, pp. 178–181.
- [9] K. Oflazer, “Two-level description of Turkish morphology,” *Literary and Linguistic Computing*, vol. 9, pp. 137–148, 1994.
- [10] M. Mohri and M. D. Riley, “Dcd library, speech recognition decoder library, AT&T Labs - Research. <http://www.research.att.com/sw/tools/dcd/>,” 2002.
- [11] H. Sak, T. Güngör, and M. Saraçlar, “Morphological disambiguation of Turkish text with perceptron algorithm,” in *CICLing 2007, LNCS 4394*, 2007, pp. 107–118.
- [12] M. Creutz and K. Lagus, “Unsupervised morpheme segmentation and morphology induction from text corpora using Morfessor 1.0. Publications in Computer and Information Science, Report A81, Helsinki University of Technology, March.” 2005.
- [13] A. Stolcke, “Srlm – An extensible language modeling toolkit,” in *Proc. ICSLP 2002*, vol. 2, Denver, 2002, pp. 901–904.
- [14] —, “Entropy-based pruning of backoff language models,” in *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, Lansdowne, VA, 1998, pp. 270–274.