

Turkish Radiology Dictation System

Ebru Arisoy, Levent M. Arslan

Boğaziçi University, Electrical and Electronic Engineering Department, 34342, Bebek, İstanbul, Turkey

arisoyeb@boun.edu.tr, arslanle@boun.edu.tr

Abstract

We have designed a Turkish dictation system for Radiology applications. Turkish is an agglutinative language with free word order. These characteristics of the language result in the vocabulary explosion and the complexity of the N-gram language models in speech recognition. In order to alleviate this problem, we propose a task-specific, radiology, dictation system. Using words as recognition units, we achieve 87.06 % recognition performance with a small vocabulary size in a speaker independent system. The most common reason of errors during the recognition is due to the pronunciation variations across speakers, and also due to the inaccurate modeling. In this paper, to overcome these problems, we proposed a pronunciation modeling technique in which variation is modeled at the lexicon level. The pronunciation variants are selected by learning the common mistakes of our speech recognition system. As a proof of the concept, firstly we apply this method to the isolated recognition of small vocabulary radiological words. Our preliminary results show that, 24.74 % error rate reduction can be achieved for isolated word recognition. This idea can also be generalized to continuous speech recognition problem with a moderate vocabulary size.

1. Introduction

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

Turkish is a difficult language for speech recognition applications. Firstly, Turkish is an agglutinative language, i.e., from the same root very

high number of words can be formed by suffixation [1]. Secondly, Turkish has a free word order. Five different types of constituent orders can be commonly used in conversational speech [2], which increases the perplexity of the language. However, the specific vocabulary of radiological terminology and systematic arrangement of words in sentence formation, make the radiology area suitable for the dictation applications. In Turkish Radiological dictation system, the vocabulary size can be reduced to only several thousand words, and the perplexity can be very small.

Our primary objective is to decrease the word error rate (WER) in Turkish radiology dictation task. The reasons of errors in speaker independent speech recognition are due to inaccurate modeling of the acoustical and language models and the variation of pronunciation between speakers. Adding different pronunciations to the vocabulary is a solution to the speaker variability. However, this may not be suitable solution to the inaccurate modeling. In this paper, our approach is that, instead of selecting the pronunciation variants using the characteristics of the language itself, we choose the variants from common mistakes of the speech recognition system.

This paper is organized as follows: In Section 2, the recognizer, the statistics of our corpus and the recognition experiments of the radiological dictation system are introduced. Section 3 describes the proposed pronunciation modeling technique. Section 4 describes the experiments and results of the proposed technique. Discussions and further research ideas based on these results are given in Section 5.

2. Turkish radiology dictation system

In this section we will describe the details of our speech recognition system for radiology applications. The reasons that radiology area is selected for this dictation system are as follows:

- ◆ Using keyboards for input entry is not appropriate for hands-busy and eyes-busy applications, and radiological diagnosis is a specific example of this type.
- ◆ Large vocabulary sizes degrade the recognition performance; however, in radiology area vocabulary size is limited to a few thousand.
- ◆ Although Turkish is a difficult language for general dictation applications, in radiology,

there is a systematic arrangement of words in sentence formation.

In light of all these facts, radiology is a suitable area to start with for a Turkish dictation system. In this section, firstly an overview of our recognizer is given. Then, the statistics of the radiology text corpus and the recognition results are explained.

2.1. Recognizer overview

In this paper, Hidden Markov Model Toolkit (HTK) [3] is used for the design of the speech recognizer. The training can be categorized into two major components: the acoustic model training and language model training. The details of the training will be explained in the following sections. The lexicon used in the recognizer consists of the phonemic representations of the radiology words. Finally, the decoder decides the best word sequence and transcribes the acoustic signal into text.

2.1.1. Acoustical training

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

We used labeled recordings of 10 speakers, each uttering 149 different phonetically balanced words and sentences for training the initial estimates of our monophone models. We generate 29 monophone models with three states and six mixtures. Then we generate our own training database recordings for radiological dictation system. We select 95 sentences covering the most frequent triphones from our radiology reports. Speech data from these sentences are recorded from 16 different speakers. These unlabeled data are labeled using the initial monophone estimates by force alignment using Viterbi algorithm, and these recordings are used for the final estimates of our monophone models. In addition, a three state silence model, for modeling the beginning and the ending of the utterances, and a one state short pause model, for representing the pauses between words, are generated.

The next step is making context-dependent triphones from monophone models and re-estimating them. This task is also accomplished using the HMM training tools of HTK. We trained 1680 triphone models and we decreased this number to 1650 by applying data driven clustering. Also, triphones that appeared in radiological report entry, but not in the acoustical training data are mapped to these 1650 physical models using the acoustical similarity criterion of neighboring phonemes.

Finally, we obtained 1650 physical models that are used in our radiological dictation system.

2.1.2. Language model training

The language modeling library of HTK can support general N-grams. However, constructing and using N-grams are limited to bigram level. Bigram probabilities with back-off smoothing are calculated using HTK. Next, we generated our recognition network with back-off bigram language model probabilities. In addition, we generated sub lattice networks for month, date and digits entries which are appended to the radiology words network. We assign small back-off bigram probabilities to these additional networks.

One important point in dictation is to make the transcript document as close as possible to the original one. Therefore the punctuation marks are very important. However, during the recording of radiological reports some doctors preferred to utter all of the punctuation marks while others were not as thorough. Another problem was that the reports collected from the hospital were very limited. Therefore, we trained the language models using the text of the same reports with all the punctuation marks and without the punctuation marks.

2.2. Statistics of the radiology corpus

The radiological reports needed for the training of the speech recognition system are collected from Hacettepe University Radiology Department. All of them are ultrasonography reports belonging to 28 different areas. We have collected 507 radiological ultrasonography reports. There are 437 reports belonging to 28 different radiological domains in the training corpus and there are 60 reports belonging to 17 different radiological domains in the test data. The analysis of the radiology training corpus in terms of number of words and the number of distinct words is given in Table 1.

Table 1: Statistics of the training corpus

	Number of words	Number of distinct words
Training Corpus	91469	1562

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

The analysis of the test data in terms of coverage and bigram perplexity is given in Table 2. As shown from the table, the coverage is very high and the perplexity is very small especially compared to the

general Turkish. This is the indication of the specific and limited vocabulary of the radiology area.

Table 2: Statistics of the test data

	Coverage (%)	Perplexity
Test Data	95.54	13.62

2.3. Recognition experiments

We perform recognition experiments with the recordings of the test data. The test data are recorded from six female and four male speakers, only two of them are doctors. The pronunciation of radiological terminology is not easy for an unfamiliar speaker. Therefore, the pronunciation variability between speakers is very high. Also the recordings are taken in two different ways, as reading reports slowly (Recordings-1) and very fast (Recordings-2). The recognition results are listed in Table 3.

Table 3: Recognition Results

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

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

During the recognition experiments, we observe that some of the recordings are consistently misrecognized in most of the utterances. These words are confused with acoustically similar other words in the lexicon, and this is the main source of recognition errors. In this paper, we aim to increase the recognition performance by handling this problem. Therefore, firstly we try to learn the common mistakes of our recognizer which are caused from the pronunciation variability of the speakers and the system itself. Then we try to correct them automatically. It is the main idea behind the proposed method for pronunciation modeling.

3. Proposed method for pronunciation modeling

Pronunciation variation constitutes a major problem for speech recognition systems. Inter-speaker pronunciation variability is addressed by employing speaker adaptation techniques. Intra-speaker pronunciation variability is another source of problem for speech recognition systems.

The reasons that result in the variation of the pronunciation among speakers or within a speaker are summarized as [4]:

- ◆ Interlocutor, which means that a speaker is influenced by the speaking style of the person they are speaking.
- ◆ In all the languages, there are free variations for the pronunciation of the same word, and the speakers are free to choose one of them during their conversation.
- ◆ There can be accent or dialect variations among speakers.

The goal of pronunciation modeling is to handle all these factors for better speech recognition performance.

The first step in pronunciation modeling is to find the pronunciation variants. In some languages, pronunciation dictionaries exist which contain possible different pronunciations of words. However, this kind of linguistic studies may not be available in all the languages. For example, pronunciation dictionaries for Turkish language are not available yet [5]. Also, pronunciation dictionaries can not contain all the variants that occur in the conversational speech, and they can give less common form of the pronunciation variant.

Data-driven methods can also be used for the selection of pronunciation variants. An approach is to generate all possible variants automatically using phonetic rules of the language [6], and then to select the most likely ones using statistical models [7]. In addition, decision tree approach has been used to generate the pronunciation variants using phonological rules [8].

A common way of using the pronunciation variants in speech recognition is to add them directly to the lexicon, as multiple transcriptions of the same word increase the chance of correct recognition of the particular word. However, adding large number of variants to the lexicon increases the acoustic confusability between lexicon entries and results in poorer recognition performance. Therefore, only the variants that give improvement in the system performance or that occur frequently have to be taken into account. In addition to that, finding the pronunciation variants available in the training data, and then re-training the acoustic models with the data containing these variants is a way of considering the pronunciation variation at the acoustic model level.

Our proposed pronunciation modeling method is also a data-driven approach and new variants of the words are added to the speech recognizer at the lexicon

level. However, our pronunciation variant generation is completely based on the available text data and the available acoustic models. The main idea is to find the common errors during recognition, and try to find alternatives of the words by repeating the same experiment with a larger lexicon compared to the previous one. The flowchart of our approach is shown in Figure 1.

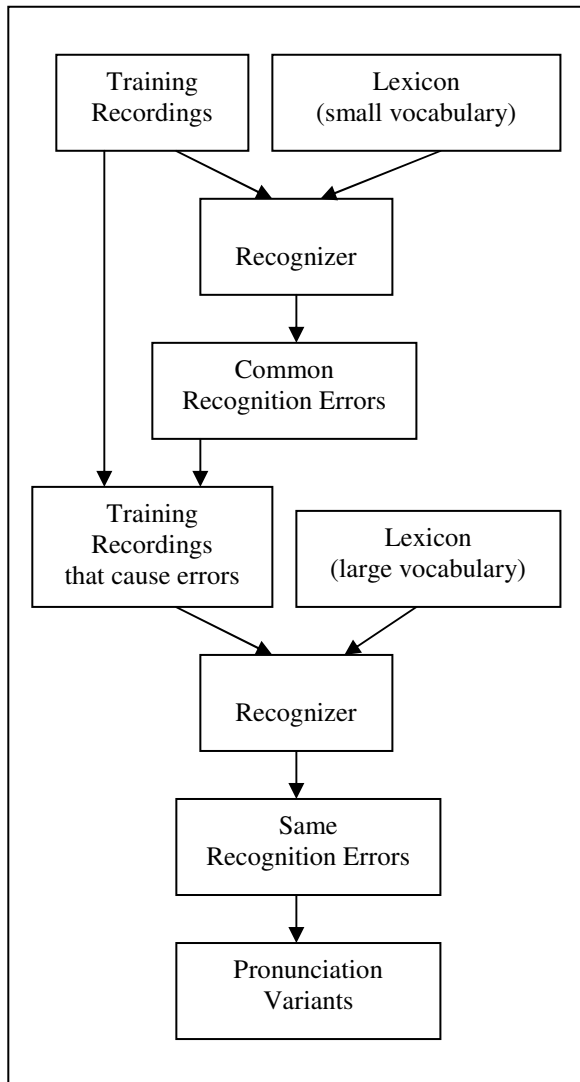


Figure 1. Flowchart showing pronunciation variant generation algorithm

As shown in Figure 1, firstly recognition experiments are performed using the training recordings with the original lexicon. Then the common recognition errors are reported. If the same word is recognized wrongly in most of the recordings, then a new recognition experiment is performed with all the recordings of this word using a very large lexicon which includes the original lexicon. For example, if word A is misrecognized in the small lexicon, and if word A is consistently confused with word B in the large lexicon, then word B is registered as a pronunciation variant of

word A. In our proposed method more than the linguistic knowledge, the important criterion in pronunciation variant selection is recognizer behavior. Due to the inaccurate modeling in HMM's, two completely different words in meaning can be selected as pronunciation variant pairs of each other. For example, the pronunciation variants of the word "loj" are selected as the word "yoliş" and the word "borç" which are completely different in meaning.

As a proof of concept, we apply this procedure to the isolated word recognition of the radiological terminology. This method is tested only on a small set of vocabulary and very encouraging results are obtained. The details of the experiment are given in the next section.

4. Experiments and results

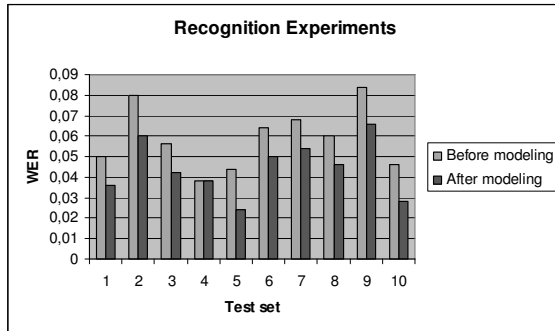
We perform the experiments using the recordings of 100 words in the radiology training corpus. The reason for using a smaller set of words is to see the performance of our proposed approach. The recordings of these words are taken from 15 different speakers, 7 females and 8 males. The number of recordings in the experiment is limited; therefore, we apply cross-validation. We perform the experiments 10 times and in each time, different combination of recordings of speakers is used. The recordings of 10 speakers are used to generate the pronunciations variants and the recordings of 5 speakers are used for testing purposes.

During the experiments, the first step is to decide on the pronunciation variants. We perform recognition experiments with the lexicon having 100 radiology words using the same acoustical models in the radiology dictation system and we observe the common recognition errors. If a word is misrecognized at least in two of the recordings out of 10, then we perform the same experiment on all the recordings of this word with a lexicon containing 100000 Turkish words containing the radiology lexicon. If at least in two of the recordings, same recognition errors occur, new words except from the radiology area are recognized, we add this word to the radiology lexicon as the pronunciation variant of the wrongly recognized word in the first experiment. By this way, instead of finding pronunciation variants to all the words in the vocabulary, we find alternative pronunciations to only problematic words for the speech recognition system.

The second step is to see the effect of adding these variants to the radiology lexicon. We perform recognition experiments on the test recordings using the lexicon having the 100 words and the appended pronunciation variants. The recognition results for different test set recordings before and after the modeling is shown in Table 4.

It is clear from Table 4; pronunciation modeling significantly decreases the WER. After the pronunciation modeling, the WER for the entire test set decreases to 4.44% from 5.90%. There is a reduction of 24.74% in the WER.

Table 4: Comparison of recognition results before and after the pronunciation modeling



5. Discussion and future work

In this paper, we propose a method for modeling pronunciation variation in speech recognition. The main idea behind this method is to learn how the recognizer behaves during misrecognitions. We aim to correct these errors using the recognition system automatically.

One of the main sources of errors is the pronunciation variation between speakers. Also there will be some errors during the modeling part. Training recordings may be uttered wrongly or in a different pronunciation. It is impossible to label all of these recordings manually and automatic labeling can not handle this problem. In addition, the training data is not large enough to generate accurate statistics. Both of these result in inaccurate modeling in the speech recognition system which is the main source of recognition errors.

We propose a method to overcome those problems. Our method differs from previous approaches such that no linguistic information regarding phonetic rules or pronunciation dictionaries is required. Using a large lexicon and the recognition system, we can generate alternative pronunciations of the words. Also we do not need to generate the pronunciation variants of all the words in the lexicon. Only the pronunciation variants of the words which are commonly confused are generated and added to the lexicon.

In a previous research study [8], it has been found that, pronunciation model adaptation at both the lexicon and acoustic model levels gives better results than adaptation only at the lexicon level. In our method, instead of optimizing acoustic models, we optimize our pronunciation variants with the acoustic models. By this way, adding our pronunciation variants directly to the lexicon contains pronunciation modeling both at the lexicon and the acoustic model levels.

Like the other data-driven approaches, the main drawback of this method is that, if any changes occur in the data, the variants have to be updated. This means if the lexicon or the acoustic models change, then the whole pronunciation variant selection process has to be repeated. This is not a big problem for the small

vocabulary systems; however for large vocabulary systems this process can be computationally expensive.

The results show that, our method performs well for small vocabulary isolated word recognition. This method can also be used in digit recognition applications or in dictation systems where words are uttered with long silence intervals.

As a further research, this idea can be tried on continuous speech recognition. At that time, computational issues can be a problem because it will be difficult to report the common errors and recognize continuous utterances with a huge lexicon. However, if these problems are alleviated, this method may also be a solution to the pronunciation variation problem for continuous speech recognition.

6. References

- [1] D. Hakkani-Tür, K. Oflazer and G. Tür, "Statistical morphological disambiguation for agglutinative languages", Bilkent University, Technical Report, 2000.
- [2] E. E. Erguvanlı, "The function of word order in Turkish grammar", Ph.D. dissertation, University of California, Los Angeles, 1979.
- [3] S. Young, D. Ollason, V. Valtchev, P. Woodland, "The HTK book (for HTK Version 3.2)", Entropic Cambridge Research Laboratory, 2002.
- [4] H. Strik, C. Cucchiari, "Modeling pronunciation variation for ASR: a survey of the literature", *Speech Communication*, vol. 29, pp. 225-246, 1999.
- [5] K. Çarkı, P. Geutner and T. Schultz., "Turkish LVCSR: towards better speech recognition for agglutinative languages", in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP-2000*, İstanbul, Turkey.
- [6] M.-B. Wesenick, "Automatic generation of German pronunciation variants", in *Proceedings of ICSLP-96*, Philadelphia, 1996, pp. 125-128.
- [7] A. Kipp, M.-B. Wesenick, F. Schiel, "Automatic detection and segmentation of pronunciation variants in German speech corpora", in *Proceedings of ICSLP'96*, Philadelphia, 1996.
- [8] M. Riley, W. Byrne, M. Finke, S. Khudanpur, A. Ljolje, J. McDonough, H. Nock, M. Sarachar, C. Wooters, G. Zavalagkos, "Stochastic pronunciation modelling from hand-labelled phonetic corpora", *Speech Communication*, vol. 29, 209-224, 1999.