Towards Acoustic Modeling of Lithuanian Speech

Darius Šilingas (1), Sigita Laurinčiukaitė (2), Laimutis Telksnys (3)

(1) Vytautas Magnus University, Vileikos 8-409, LT-44404 Kaunas, Lithuania
d.silingas@if.vdu.lt
(2) Institute of Mathematics and Informatics, Goštauto 12-204, LT-01108 Vilnius, Lithuania
sigita.lau@mch.mii.lt
(3) Institute of Mathematics and Informatics, Goštauto 12-205, LT-01108 Vilnius, Lithuania
telksnys@ktl.mii.lt

Abstract

In this paper we present experimental investigation of using various phone sets for acoustic modeling of Lithuanian speech applied to large vocabulary continuous speech recognition. Paper presents specifics of Lithuanian speech acoustics including accentuation, diphthongs, softening and assimilation of consonants. The speech recognition experiments use only acoustic model since effective language modeling for highly inflected Lithuanian language is still under research. Several Lithuanian phone sets are proposed for evaluation in speech recognition experiments. A new Lithuania broadcast news corpus LRN0 is presented. Phone occurrence frequencies in 9 hours speech training data for multiple Lithuanian phone sets are given. Recognition performance for Hidden Markov Models based on multiple proposed simple and contextual phone sets is evaluated using HTK toolkit. Experiment results are presented using figures comparing word error rates for phone sets. Conclusions indicate influence of modeling various linguistic features such as accent, softness, mixed-diphthongs, affricates, and context to recognition performance, recommend a phone set to use for Lithuanian speech recognition, and point the future research directions.

1. Introduction

Human language technologies are becoming more and more important in the modern world where computers and information technologies play major role. Speech recognition, synthesis and language analysis have found many commercial applications in telecommunication services, personal computers, and various devices. However, human language technologies for major languages, especially English, are much more developed than for minor languages like Lithuanian. There are two main reasons: the market size and language structure. Obviously, market size for English language technologies is much larger than for Lithuanian, and there is larger ROI (Return On Investment) for developing commercial products for English. Research and development of human language technologies usually reuse the same fundamental methods like signal processing techniques, Hidden Markov Models (HMM), statistical recognition and analysis frameworks regardless of the target language. However, the details and sophistication often rely heavily on language specifics. There are also many situations when methods applicable for one language are not suitable for another because of different language structure. Minor languages are usually much more complex in nature than major languages that were simplified because of the widespread usage. And this complexity raises many problems, which call for language-specific research. Lithuanian is considered to be the most ancient of the living Indo-European languages and has retained multiple specific properties. The specifics of Lithuanian language relevant to speech recognition are presented in [1]. One of these features, which we aim to investigate in this paper, is the acoustics of Lithuanian language that plays a major role in choosing modeling units for acoustical models used for speech recognition. Importance of phonetics for successful speech recognition is argued in [2]. Some related experiments with large vocabulary Lithuanian speech recognition were reported in [3], phone set for annotating Lithuanian speech corpora was proposed in [4]. There were multiple other works related to acoustical modeling of Lithuanian speech but most of them are either constrained to solving very specific problems and lack sufficient amount of data for statistically significant experiments.

2. Statement of the Problem

We aim at exploring multiple variations of Lithuanian phone sets and evaluating their usage for continuous speech recognition using appropriate amount of training and recognition evaluation data. We have the following goals:

- Define variations of Lithuanian phone sets;
- Compute phone occurrence frequencies in training data;
- Evaluate recognition performance using context-independent phone sets;
- Derive context-dependent phone (triphone) sets from the best performance phone sets;
- Compute triphone occurrence frequencies in training data;
- Evaluate recognition performance using triphone sets.
3. Specifics of Lithuanian Speech Acoustics

Lithuanian acoustics has multiple unique properties. We will shortly introduce those that are relevant for speech recognition. More detailed description of Lithuanian acoustics is given in [5].

Lithuanian accentuation system defines two types of accents – rising pitch accent circumflex, and falling pitch accent acute. In addition, the rising pitch accent may have long or short duration, while the falling pitch accent always has long duration. Thus there are three accents and they may be critical for differentiating between similar words, e.g. words šauk (Imperative of verb fire) and šauk (Imperative of verb shout) differ only in accent. Since there are many similar cases in Lithuanian and accent is well known to influence the acoustics of phone, accent information should be included in the phone sets used for speech recognition.

Lithuanian has multiple diphthongs that are considered by linguists to be closely connected and are suggested to be modeled as a single phone rather than two. There are two main classes of diphthongs: pure diphthongs containing two vowels, and mixed diphthongs containing a vowel and a consonant. Diphthong sounds are pronounced as one unit, pretty different from the same sound pairs when they do not form diphthong. Another reason for modeling diphthongs as a single phonetic unit is the accent, which may fall on diphthongs second part, and in case of mixed diphthongs this part is consonant, which is never accented alone.

There are also multiple effects of Lithuanian phone co-articulation. The two of the most importance for speech recognition are the assimilation of consonants and softening of consonants preceding specific vowels.

We will explore how softness, accent and mixed diphthongs influence speech recognition performance by evaluating phone sets with and without modeling these acoustical features.

4. Experiment Description

4.1. Speech Recognition Framework

The basic speech recognition problem is to find solution for the following formula:

\[ \hat{w} = \arg \max_p P(w|y) = \arg \max_p P(w) P(y|w) \]

Here \( W \) is a possible word sequence and \( Y \) is a feature vector, which is a compact representation of the observed speech waveform. Term \( P(W) \) is modeled by language model and term \( P(Y|W) \) is modeled by acoustic model.

We reuse the recognition framework presented in [6]. However, we use only term \( P(Y|W) \), which is modeled by HMMs, since conventional context-free n-gram language models are not effective for highly-inflected Lithuanian language as pointed out in [1]. Lithuanian language modeling is under research, but to our knowledge no effective solution was implemented or described in a concise way yet. Therefore we focus on investigating efficiency of acoustic modeling.

We use HTK toolkit for extracting features from speech waveforms, training HMMs and evaluating recognition performance. Reader may refer to [7] or [8] for detailed description of HMM-based speech recognition framework and to [9] for description of how it is supported by HTK toolkit. Data for experiments was taken from LRN0 corpus, which is described in the next chapter.

4.2. Speech Corpus

For executing experiments we collected Lithuanian broadcast news speech corpus LRN0 (Lithuanian Radio News version 0). Corpus contains over 9 hours of speech records. Speech samples were recorded directly from broadcasting of Lithuanian Radio first program (LR1) with permission and agreement between Lithuanian Radio and Institute of Mathematic and Informatics. The content of records covers the most important political, economical and sport events of local and foreign areas. The text has large number of specific names. Each original record was of 12 min. duration news read by multiple Lithuanian Ration newreaders. 141 records were recorded for speech corpus and each of it was manually split into sentences, providing sentence level annotations. Specific marks such, as pauses, silence, breathing and mispronunciations were included as well. Speech corpus is accompanied by words-to-phones transcription dictionary. It contains ~18 000 entries, including words with similar orthographic transcriptions and different pronunciations. Phonetic transcriptions and stress markings were created manually referring to [10] and [11]. Original phonetic transcriptions were created taking into account phone duration, softness and stress according to linguistic rules. The resulting phone set included 74 simple phones, 156 diphthongs and 3 special pseudo-phones for pause, silence and breathing – in total 232 phonetic units. Later, we used automatic dictionary transformations to other phone sets that were investigated in experiments. Corpus contains records of 23 male and female speakers, all Lithuanian Radio newreaders speaking with correct and clear pronunciation. The records of 10 speakers (4 females and 6 males) make 89% of the speech corpus. Main characteristics of records: sampling rate – 11 kHz, channels – mono, resolution – 16 bit. Speech corpus was divided into training, development and evaluation data sets. Training set contains over 9 hours (6564 sentences), development set contains about 2 minutes (50 sentences), and evaluation set – about 14 minutes (360 sentences) of speech. Evaluation data do not include the names of specific places or people and is restricted to 5-9 word sentences without mispronunciations.

English speech recognition experiments under similar conditions using corpus WSJCAM0 were reported in [6]. WSJCAM0 corpus is available from
LDC (Linguistic Data Consortium) at http://www.ldc.upenn.edu and its detailed description is given in [12].

4.3. Experiment Design

We will go through the following main steps in our experiment:
1. Selecting phone sets to be used as base for HMM models;
2. Training HMMs based on selected phone sets and evaluating their recognition performance;
3. Deriving contextual phone (triphone) sets from the best performance phone sets;
4. Training HMMs based on selected triphone sets and evaluating their recognition performance.

4.3.1. Phone Sets Description

Phone set #1 defines phones taking into account properties proposed by linguists: softness of consonants, accent information, diphthongs (vowel-vowel) and mixed diphthongs (vowel-consonant) and affricates. This phone set provides the standard simple phone list which is the baseline for evaluating other phone sets, where some property (softness, accent, mixed diphthongs or affricates) is removed, i.e. not taken into account, and measuring the impact of this change in terms of recognition word error rate change.

Phone set #2 removes accent from phone set #1.
Phone set #3 removes softness from phone set #1.
Phone set #4 removes both softness and accent from phone set #1.

Phone set #5 splits mixed diphthongs and removes affricates from phone set #1, because mixed diphthongs and affricates were rarely seen in training data set – all phones that appeared less than 50 times were either mixed diphthongs or affricates. When splitting mixed diphthongs if accent falls on the consonant part it is removed as well (separate consonants can’t be stressed).

Table 1 displays the main phone set properties: presence of softness (Soft), accent (Acc), mixed diphthongs (M-D), affricates (Affr), total number of phones (N), and number and percentage of phones that appear less than 50 times in training data (N50).

<table>
<thead>
<tr>
<th>#</th>
<th>Soft</th>
<th>Acc</th>
<th>M-D</th>
<th>Affr</th>
<th>N</th>
<th>N50</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>229</td>
<td>56 (~24%)</td>
</tr>
<tr>
<td>2</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>139</td>
<td>11 (~8%)</td>
</tr>
<tr>
<td>3</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>140</td>
<td>19 (~14%)</td>
</tr>
<tr>
<td>4</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>86</td>
<td>1 (~1%)</td>
</tr>
<tr>
<td>5</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>87</td>
<td>0 (0%)</td>
</tr>
</tbody>
</table>

4.3.2. Training and Recognition Evaluation for Phone HMMs

The following experiments were executed for all phone sets:
1. Training on whole training set data using phone HMM models with 1-component Gaussian mixture, evaluating recognition on development set data, then increasing number of Gaussian components by 1 and repeating training and recognition for 2-, 3-, and 4-component Gaussian mixtures;
2. Evaluating recognition on evaluation set data for phone HMM models with 4-component Gaussian mixture, which gave the best recognition performance on development set data.

The figures 1 and 2 display the results of these experiments.
Figure 1 – Word Error Rates (WER) for phone HMMs with 1-, 2-, 3-, and 4-component Gaussian Mixtures on Development Set Data

Figure 2 – Word Error Rates (WER) for phone HMMs with 4-component Gaussian Mixtures on Evaluation Set Data
Experiment results suggest that including softness information is of greater influence to recognition performance than accent. Including accents is also beneficial to recognition results. However, including softness and accent info in phone set increases the number of phones and introduces some phones that are very rarely found in training data. Since most of rare phones are variations of mixed diphthongs, splitting mixed diphthongs helps to avoid the problem of rare phones. Also, the results of experiment with evaluation data show that WER using phone set #5 is second lowest and on development data set even outperforms original phone set #1, which contains ~2.5 times larger number of phones. Thus, splitting mixed diphthongs is an efficient solution and should be considered when developing HMMs based on contextual phones.

4.3.3. Additional Experiments with Phone Sets

Since phone set #1, which gave the best recognition performance, had a lot of rarely seen phones, we applied clustering of HMMs using questions tree constructed according to Lithuanian phones’ linguistic properties. This resulted in clustering 90 of 229 phones (40%), mostly mixed diphthongs. However, WER for clustered models did not improve and even increased – from 37.62% to 42.40% on development data set.

We have also tried to use different HMM topology for diphthongs – six state HMMs rather than three state HMMs. However, the recognition performance has slightly decreased thus we suggest using three state HMMs for diphthongs as well.

4.3.4. Triphone Sets Description

Since the best recognition performance was achieved for phone set #1, we have formed the contextual phone (we will use term triphone) set #1 from phone set #1. Consonant softness is basically an influence of contextual vowels, thus it is taken into account implicitly when modeling triphones. Therefore we will use phone set #1, where softness is removed, as a basis for our triphone set #2. The second-best performance was achieved using phone set #5, which also had a desirable property of minimum number of phones with all of them appearing more than 50 times in training data set. Thus we formed triphone set #3 from phone set #5. Similarly, we formed triphone set #4 from phone set #5 with removed softness information.

Table 2 displays the main triphone set properties: base phone set (BPS), total number of triphones (N), number and percentage of triphones that appear more than 50 times in training data (\(N_{50}\)), and number of models after clustering triphones (\(N_c\)). The number of triphone models that have been clustered (i.e. tied to other triphone models) is \(N - N_c\). We have used questions tree for clustering triphone HMMs as suggested in [13] and described in details in [9]. It is obvious from numbers, that some rare triphones were not clustered, thus our question tree for was not fully effective and should be improved for better clustering. Similar problems are researched for other languages like Russian, see [14].

<table>
<thead>
<tr>
<th>#</th>
<th>BPS</th>
<th>N</th>
<th>(N_{50})</th>
<th>(N_c)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>22387</td>
<td>2206</td>
<td>10302</td>
</tr>
<tr>
<td></td>
<td></td>
<td>~10%</td>
<td>~46%</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>18719</td>
<td>2215</td>
<td>9156</td>
</tr>
<tr>
<td></td>
<td></td>
<td>~12%</td>
<td>~49%</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>17911</td>
<td>2349</td>
<td>10134</td>
</tr>
<tr>
<td></td>
<td></td>
<td>~13%</td>
<td>~57%</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>5'</td>
<td>14556</td>
<td>2293</td>
<td>8515</td>
</tr>
<tr>
<td></td>
<td></td>
<td>~16%</td>
<td>~58%</td>
<td></td>
</tr>
</tbody>
</table>

4.3.5. Training and Recognition Evaluation for Triphone HMMs

The following experiments were executed for all triphone sets:

1. Training on whole training set data using HMM models with 1-component Gaussian mixture, evaluating recognition on development set data, then clustering triphones using questions trees made according to Lithuanian linguistic rules, evaluating recognition on development set data, increasing number of Gaussian components by 1, training and evaluating recognition for 2-, 3-, and 4-component Gaussian mixtures;

2. Evaluating recognition performance using clustered triphone HMM models with 4-component Gaussian mixtures on evaluation set data.

The figures 3 and 4 display the results of these experiments.

* Softness marking is removed
Figure 3 – Word Error Rates (WER) for triphone HMMs with 1-Gaussian component and for clustered triphone HMMs with 1-, 2-, 3-, and 4-component Gaussian Mixtures on Development Set Data

Figure 4 – Word Error Rates (WER) for clustered triphone HMMs with 4-component Gaussian Mixtures on Evaluation Set Data
Experiment results show that although phone set #1 gave the best recognition performance for simple phone HMM-based recognition, the triphone set #1 derived from it gave significantly worse recognition results than triphone set #3 derived from phone set #5, which had lower number of triphones and higher percentage of triphones that appeared more than 50 times in training data set. Including softness in phone set from which we need to derive triphones is rather questionable issue since it results in larger number of triphones and difference in recognition performance is very slight and many even be better for triphones derived from phones without softness information.

5. Conclusions
We have presented investigations of acoustical modeling for Lithuanian continuous speech recognition based on simple and contextual phone set variation. From the results of the executed experiments we have drawn the following main conclusions:
- Modeling softness and accent information provides better recognition performance using simple phones;
- Using triphones give much better recognition performance than using phones with rich linguistic info (duration, softness, accent);
- Since mixed diphthongs are relatively rare, splitting them into two separate phones helps to minimize phone/triphone set size and does not result in significant decrease in recognition rate using phones and provides significantly better results using triphones;
- Similar three state left-to-right HMM topologies should be reused for both simple phones and diphthongs;
- Softness information is implicitly modeled by context information in triphones, thus triphones can be constructed from phone sets without softness information;
- Construction of question trees for clustering triphones needs to be further researched in order to cluster rare triphones more efficiently;
- The triphone set derived from phone set including accent and splitting mixed diphthongs gave the best recognition results – thus it is the recommended modeling unit set for using in Lithuanian speech recognition systems.

6. References