On the Adaptation of Foreign Language Speech Recognition Engines for Lithuanian Speech Recognition

Vytautas Rudzionis[1], Rytis Maskeliunas[2], Algimantas Rudzionis[2], and Kastytis Ratkevicius[2]

[1] Vilnius University Kaunas faculty, Kaunas, Lithuania
[2] Kaunas University of Technology, Kaunas, Lithuania
vytautas.rudzionis@vukhf.lt

Abstract. This paper presents some of our activities trying to adapt the foreign language based speech recognition engines for the recognition of the Lithuanian speech commands. In recent years several quiet successful speech recognition engines became available for the most popular languages (such as English, French, Spanish, German, etc.). The speakers of a less widely used languages (such as Lithuanian) have several choices: to develop own speech recognition engines or to try adapting speech recognition models developed and trained for the foreign languages to the task of recognition of their native spoken language. First approach is expensive in time, financial and human resources sense. Second approach can lead to faster implementation of Lithuanian speech recognition modules into some practical tasks but proper adaptation and optimization procedures should be found and investigated.

Keywords: speech recognition, speech engine, transcriptions, phonetic.

1 Introduction

Speech recognition is long awaited mode for human computer interaction that needs to be introduced into many modern devices. In recent years pressure to develop faster voice based dialogue interfaces rises due to more and more widespread use of mobile devices such as mobile phones (many of them has capabilities significantly overcoming capabilities of traditional phones), tablet PCs, pocket PCs, etc. The characteristic property of similar devices is their portability which means reduced screen and keyboard. Small keyboard and screen introduces significant inconveniences for user to use them efficiently. Voice based dialogue interface often may become efficient substitute for traditional GUI based interface or in some cases may become single possible solution for mobile users.

Importance of voice based interfaces is stressed additionally by the fact that many business consulting companies predict that speech will remain dominant component in total data flow over wired and wireless networks in coming years.

From the advent of speech recognition research and the appearance of first commercial applications the main efforts were devoted to the recognition of widely used languages, particularly English language. The reason of such behavior is very clear – widely used languages have bigger market potential for practical applications. So

W. Abramowicz and D. Flejter (Eds.): BIS 2009 Workshop, LNBIP 37, pp. 113–118, 2009.

looking at the general trend in the development of commercial speech recognition applications and tools for the development of speech recognition, such sequence could be observed: first version of speech recognition engine oriented to the recognition of English (and particularly US English) is released, then that system is supplemented with the engines for the other widely used languages (most often Spanish, French, German and several others) and sometimes but not necessarily with recognition modules of some other relatively widely used languages (in example Dutch, Italian, Turkish, Polish, etc.). Many other less widely used languages remains out of the scope of interest for the major speech recognition solution providers.

In such situation businesses and state institutions, in countries were such less popular languages are used as a main source of spoken language communication, faces a challenge of development of own speech recognition tools. Two major ways for solution are as follows:

- to develop own speech recognition engine from scratch;
- to adapt the foreign language based engine for the recognition of your native language.

The first approach has potentially higher capabilities to exploit peculiarities of selected language and hence to achieve higher recognition accuracy. But the drawback of such approach is that the providers of major speech technologies avoid the implementation of such languages in their products – high costs in the general sense of this word.

The second approach has the potential to achieve some practically acceptable results faster than developing entirely new speech recognition engine. Another advantage of this approach is potential to achieve faster compatibility with the existing technological platforms. Such advantage is often important for business customers, since they need to follow various technical specifications in order to guarantee consistent functioning of the enterprise. But this approach also requires careful investigation of the ways of adapting and optimizing adaptation algorithms.

This paper will present some of our activities trying to adapt speech recognition engine oriented to the recognition of spoken English for the recognition of several Lithuanian voice commands.

2 Expert – Driven and Data-Driven Approaches

The topic of the multilingual and crosslingual speech recognition is very important because of the reasons mentioned above. The importance of this research topic is expressed especially in Europe as this region is a multi-cultural society with many languages used in parallel. As the costs of generating a non-existing speech database and training acoustic models from that database can be very high, one possibility is to use crosslingual speech recognition. The idea behind this is to transfer the existing source acoustic models from source language to the target language without using speech corpora in that language and without full retraining of the speech recognition system.

Similarity measures used to transfer the source acoustic models to a target language can be divided into two major groups [4]: expert-driven methods and data-driven crosslingual speech recognition approaches.

In expert-driven methods mapping from one language to another is performed using human knowledge and is typically based on some acoustic-phonetic characteristics. One of the most frequently used methods is the use of so called IPA scheme. Expert knowledge of all included languages is needed. Such approach could be very difficult if many different languages were included in the system or must be used for the optimization. It was also observed that some subjective expert influence from the mapping can be expected.

IPA scheme [3] we will describe in more detail. For each phoneme in target language an equivalent phoneme in the source language was searched for. As an equivalent phoneme with the same IPA symbol is selected often. The ratio of the equivalent phonemes depends on the acoustic similarity of languages and on the number of phonemes in the all involved languages. In the case when IPA equivalent is non-existent in the target language the most similar phoneme according to IPA scheme is looked for. The search for the most similar candidate can be performed in horizontal or vertical direction through the IPA scheme. The main advantage of described scheme is that it can be applied without any speech material in the target language. Disadvantage of such approach is that expert knowledge should be obtained somehow and this knowledge also has subjective influence introduced by the expert. Data-driven crosslingual speech recognition approaches are based on data-driven similarity measures. In these methods the similarity measure is applied during mapping. Similarity measure itself is obtained from some data applying some algorithm. Data-driven approach with the phoneme confusion matrix will be described below in more details. The idea behind this method is that similar phonemes are confused during speech recognition by a phoneme recognizer. The basic characteristic of such recognizer is that it recognizes phoneme sequences instead of words from a vocabulary. For generating crosslingual confusion matrix, acoustic models of one of the source languages were applied on speech utterances of the target language. The recognized sequence of the source phonemes was then aligned to the reference sequence of the target phonemes. The output of this alignment was the crosslingual phoneme confusion matrix M. At this stage for each target phoneme f_{trg} the best corresponding source phoneme f_{src} should be looked for. As similarity measure, the number of phoneme confusions $c(f_{trg}, f_{src})$ is often selected.

So for each target phoneme f_{trg} the source phoneme f_{src} with the highest number of confusions c is selected in this schema. If two or more source phonemes has the same highest number of confusions it was proposed to leave decision for the expert which one of source phonemes should represent target phoneme f_{trg}. The same procedure could be applied if no confusions between source and target phonemes were observed.

The advantage of described data-driven approach based on a confusion matrix is that it is fully data-driven method and theoretically no subjective expert knowledge required (in practice expert knowledge is necessary to solve situation when same or similar confusions were observed).

3 Multiple Transcriptions Based Recognition

This paper will deal with the task of adapting Microsoft speech recognizer for Lithuanian speech recognition using two different vocabularies for 100 Lithuanian names (first names and family names): it is expected that this vocabulary is less complicated since names are longer and task is related with the choice of possibilities;

Selection of vocabulary was determined by the practical potential of applications that could be developed from this vocabulary. The main characteristic of used approach is that multiple transcriptions were used for single command. The number of transcriptions used per word or command wasn't constant and was the case of some rough optimization (optimization was done by single speaker and developer trying to find some optimal performance level for that person and later those transcriptions were used for other speakers as well). And one more difference from that study was bigger number of speakers and utterances used in the experiments.

For Lithuanian first and family names database 33 different speakers uttered each name and surname once. The total number of combinations first name+family name was equal to 100 in this case (3300 utterances total).

4 Lithuanian Proper Names Recognition

Microsoft Speech Recognition engines were used as a basis for adaptation. Two speech databases were applied in these experiments: initial corpora and corrected corpora. Corrected database was freed from various inadequacies and mistakes that were present in the initial database. Between inadequacies and mistakes were pronunciation errors. Most of such pronunciation errors were situations when speaker used other phoneme than the phoneme present in the original family name, but still getting grammatically correct and often really existing name (in example, speaker said *Dalinkevicius* instead of *Danilkevicius*). Other errors were related with such problems as stammer near family name or some technical spoilage such as a truncated part of the word (most often at the end of the name).

Experiments were performed for male and female speakers separately and also for all speakers together. Table 1 show the results obtained in those experiments. Last row in the table presents results obtained using corrected speech corpora.

One of the most interesting observations was that recognition accuracy for initial and corrected databases was almost equal which show, that quite robust recognition system for small pronunciation errors may be developed (situation rather probable in applications, where we need to recognize people names). Detailed analysis showed that one speaker made even 19 errors when reading names and confusing at least one phoneme in the first name or family name, but all utterances were recognized correctly.

Table 1. The recognition accuracy of 100 Lithuanian names using the adapted transcriptions for the Microsoft Speech Recognition engine

Speakers	Correct, %	Insertions, %	Indeterminacies, %	Omissions, %
16 females	89.8	6.5	3.5	0.2
17 males	92.5	3.9	3.4	0.2
33 both genders	91.2	5.2	3.3	0.4
corrected	91.4	5.6	2.9	0.1

Another group of experiments using 100 Lithuanian names was carried out using clean and noisy or clipped speech. The aim was to evaluate robustness of recognizer for speech signal distortions. Only 5 speakers participated in this experiment and several SNR levels were selected and clipping coefficients were used. Table 2 shows results obtained in those experiments.

Table 2. Recognition accuracy of 100 Lithuanian names for different quality of speech signals (5 speakers)

Distortion	Correct, %	Insertions, %	Indeterminacies, %	Omissions, %
Clean	96.6	1.2	2.2	0.2
SNR 40dB	96.2	1.6	2.2	0.0
SNR 30 dB	91.6	1.4	7.0	0.0
SNR 20 dB	43.8	10.0	29.2	17.0
Clipped 0.3	93.0	4.8	2.2	0.0
Clipped 0.1	82.0	13.6	4.0	0.4

Looking at the table we see that the performance of the recognizer began to deteriorate significantly when the SNR level dropped below 30 dB and was in principle unacceptable at the SNR 20 dB. So the performance can't be treated as robust one looking from the SNR variations point of view. Using clipping coefficient 0.3 recognizer performance' dropped relatively insignificantly while clipping coefficient 0.1 resulted in much bigger loss in accuracy.

Table 3. Five names that result in the largest number of recognition errors in the 100 Lithuanian names recognition experiment

Indeterminacy errors		Insertion errors	
name	number of errors	name	number of errors
Gudas_audrius	17	Biaigo_sandra	16
Baublys_algis	12	Dolgij_andrej	16
Biaigo_sandra	6	Grigonis_audrius	11
Balcius_ernestas	6	Baublys_algis	10
Gailiunas_rytis	6	Giedra_nerijus	8

We've performed some recognition errors analysis trying to find the ways to improve the recognizer performance and find the ways to optimize adaptation procedure. There were 290 substitution and insertion errors (120 substitution and 170 insertion) in our first group of experiments and it was natural to expect that not all names will produce the equal number of errors. Here we've need to state that as an insertion error we treated situation, when recognition system produced some phonetic unit at the output of the system, which resulted in the name that wasn't present in the list of names (typical recognition system output in this situation was "*I don't understand you*"). Table 3 shows the 5 names that produced the largest number of errors in these experiments.

Looking to those results we see that 5 most confusing names produced almost 40% of all substitution errors and slightly more than 35% of all insertion errors. So the “concentration” of errors is big and more attention to the names that resulted in larger amounts of errors is necessary.

Detailed view to the most confusing names in these experiments showed that most of those names don’t have difficult phonetic structure. The bigger number of errors obtained by the name *Gailiunas_rytis* may be explained by the presence of the name *Gailiunas_vytautas* in the same list. But most of the errors can’t be explained straightforwardly. For example, name *Gudas* often was confused with the name *Butkus.*

5 Conclusions

This paper presented some of our activities trying to adapt the foreign language based speech recognition engines for the recognition of the Lithuanian speech commands. The speakers of less popular languages (such as Lithuanian) have several choices: to develop own speech recognition engines or to try adapting the speech recognition models developed and trained for the foreign languages to the task of recognition of their native spoken language. First approach is expensive in time, financial and human resources. Second approach can lead to a faster implementation of the Lithuanian speech recognition into some practical tasks, but the proper adaptation and optimization procedures should be found and investigated.

For 100 Lithuanian names recognition accuracy of more than 90% was achieved. These results show that the implementation of longer commands and transcription generation methodic proposed in [2] study were confirmed.

References

1. Lindberg, B., et al.: Noise Robust Multilingual Reference recognizer Based on Speech Dat. In: Proc. of ICSLP 2000, Beijing, vol. 3, pp. 370–373 (2000)
2. Kasparaitis, P.: Lithuanian Speech Recognition Using the English Recognizer. INFORMATICA 19(4), 505–516 (2008)
3. Zgank, A., et al.: The COST278 MASPER initiative – crosslingual speech recognition with large telephone databases. In: Proc. of 4th International Conference on Language Resources and Evaluation LREC 2004, Lisbon, pp. 2107–2110 (2004)
4. Zgank, A.: Data driven method for the transfer of source multilingual acoustic models to a new language. Ph.D. thesis, University of Maribor (2003)
5. Phoneme Table for English (United States), `http://msdn.microsoft.com/enus/library/bb813894.aspx` (retrieved December 19, 2008)