ISSN 1392 – 124X INFORMATION TECHNOLOGY AND CONTROL, 2009, Vol. 38, No. 3

LITHUANIAN CONTINUOUS SPEECH CORPUS LRN 1: AN IMPROVEMENT

Sigita Laurinčiukaitė[1], Mark Filipovič[2], Laimutis Telksnys[1]

[1]*Institute of Mathematics and Informatics, Recognition Processes Department*
A.Goštauto St. 12-203, LT-01108 Vilnius, Lithuania
e-mail: S.Laurinciukaite@vmi.lt

[2]*Information Technology and Communication Department under the Ministry of the Interior of the Republic of Lithuania*
Šventaragio St. 2, LT-01510 Vilnius, Lithuania

Abstract. This paper presents the development of Lithuanian continuous speech corpus LRN 1 (Lithuanian Radio News, version 1). The corpus was developed from speech corpus LRN 0.1 by increasing the duration of speech corpus (it lasts 20 hours 50 minutes). The major improvement of speech corpus LRN 1 was a development of time-aligned word level annotations of speech signals. Time-aligned word level annotations of speech signals were obtained after a two-stage process: automatic realignment of acoustic models of phonemes and subsequent manual correction of annotations. The improvement of the corpus is useful for constructing and evaluating speaker-independent continuous speech recognition systems and for linguistic research.

Keywords: speech corpus, speech annotation, phonetic labelling, speech realignment, hidden Markov model.

1. Introduction

Most modern speech recognition methods are based on statistical models that require large amounts of training and test data for evaluating and tuning of parameters. Therefore specialized speech databases, called speech corpora, become essential for scientific research. Different corpora are collected because of the high variability of speech signals in vocal tracts of speakers, environments, communication channels and style of speech. They help to investigate speech recognition in different aspects and serve for a better performance of speech recognition systems.

Larger corpora provide a better representation of language variability, but their use depends on how well they are constructed. Speech corpora differ not only in content of speech signals, but also in technical characteristics such as the number of speakers, annotation, sampling, structure of corpus, the type of subword units used, and others, that help to envisage how a corpus can be used in scientific research. The most valuable characteristic of a speech corpus for researchers is a comprehensive annotation of speech signals, i.e., time-aligned annotation of a speech signal on different levels: phoneme, word, and syllable. These facilities enable researchers to pursue a diverse and deep study of speech recognition. The process of manual producing of high quality linguistic annotations is time consuming and requires much effort and linguistic expertise. Therefore the work of human annotators and that of an automatic annotation system is combined.

2. Annotation methods of speech corpora

Currently, the majority of available speech corpora used worldwide such as TIMIT, TIDIGITS, WSJ, SWITCHBOARD, and Verbmobil are annotated on a few levels. Annotations, provided by a particular corpus, will vary depending on the purpose for which the corpus is intended. Typically annotations vary from phonetic labelling of segments (phonetic transcriptions as in TIMIT) to labelling of some semantic category [5]. Phonetic labelling of segments is of particular importance in the improvement of speech corpus LRN 1, as the previous version of the corpus contained just sentence level annotations of speech without time marks at the level of word boundaries or any other segment. So phonetic labelling is discussed in the sequel.

As mentioned before, the work of human annotators supplements the automatic annotation system, or vice versa. Automatic alignment is used for a number of corpora to create time-aligned annotation files on word or phoneme levels. A human annotator must verify these automatically generated annotations. On

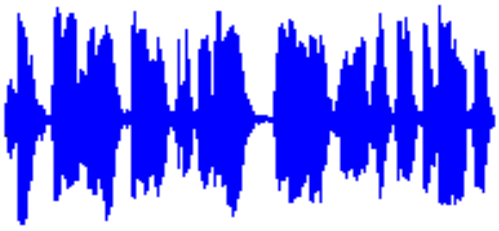
the other hand, manually produced multilevel annotations are used for annotations of new utterance, using a generalized finite-state transducer constructed from the training corpus [5]. The use of an automatic approach in these cases diminishes human efforts a great deal; tedious annotation marking and labelling can be potentially performed in an automatic way.

Lithuanian speech corpora collected at the Kaunas University of Technology, Vytautas Magnus University, and the Institute of Mathematics and Informatics differ in duration (from 0,5 to 21 hour), the number of speakers (from 1 to 350), type of speech (isolated words, connected words, continuous speech), size of lexicon (from 50 to 32 000 words), and the annotation level of speech data. Speech corpora LTDIGITS [2] and the Universal Annotated VDU Lithuanian Speech Corpus [1] have phoneme and word level annotations. These speech corpora are of medium duration, 1-6 hours. As it has been stressed earlier, larger corpora provide a better representation of language variability, so it is essential to have large, well annotated Lithuanian speech corpora. We hope that annotation of LRN 1 on a few annotation levels, i.e., inserting time marks at the level of word boundaries and afterwards inserting time marks at the level of phoneme boundaries, are some steps forward to obtain a comprehensive Lithuanian speech corpus.

3. Design and improvements of LRN 1 Corpus

Speech corpus LRN 1 was developed from speech corpus LRN 0.1 [3]. The characteristics of speech corpus LRN 1 are the same as those of speech corpus LRN 0.1 and are given in Table 1. A difference occurs in the duration of speech corpora; it increased from 17 hours 23 minutes to 20 hours 50 minutes. The duration of training, development, and evaluation of test sets have changed with an increase in the total duration of a speech corpus.

The major improvement of speech corpus LRN 1 was the development of one more annotation level of speech. The previous versions of speech corpora LRN 0 and LRN 0.1 had annotations of speech on a sentence level, i.e., word boundaries were not known, only a sequence of words that compose a sentence (Figure 1). These initial annotations were created employing trained transcribers. The work, described in this article, covers addition of annotations of speech on a word level. The boundary of each word is known in these annotations.



šiuos pareigūnus seimo pritarimu skiria ir atleidžia prezidentas

Figure 1. An example of a waveform and text file pair

The development of a word-level annotation of speech signals was a process of two stages. It included automatic annotation of speech signals and manual correction of results of the first stage. Each stage used a different method and tools. Subsequent sections describe the annotation method and tools for each of the two stages used for the development of word-level annotations of speech.

3.1. Stage 1

HTK tools [7] and model (speech data) realignment were major methods for automatic annotation of speech on the word level.

Realignment is usually applied to training data of the speech recognition system to create transcriptions (annotations) that match acoustic data best. Alignment of data can be performed if phoneme models are developed. Later on, these new phoneme-level transcriptions can be used to re-train phoneme models.

We used the approach described above in automatic annotation of a speech corpus for setting a boundary of words in transcriptions, i.e., for automatic annotation of a speech corpus on the word level. A short description of the realignment process follows.

We got through two main steps in our realignment of a speech corpus:

1) training of hidden Markov models (HMMs) based on a selected phoneme set;
2) realignment of data in the recognition process.

A phoneme set used for acoustic modelling was the same as that chosen for speech corpus LRN 0 (earlier version of LRN 1) and for the current speech corpus LRN 1. The phoneme set defines phonemes taking into account the properties proposed by linguists: softness of consonants, accent information, diphthongs (vowel-vowel), mixed diphthongs (vowel-consonant), and affricates. This phoneme set provides the standard mono-phoneme list. The effectiveness of this phoneme set was proved in different scientific researches [8, 9].

Training of HMMs was performed on data of the whole speech corpus, using phoneme models with 1-component Gaussian mixture, then increasing the number of Gaussian components by 1 until 4-component Gaussian mixtures have been obtained.

The realignment of training data that followed after training HMM's resulted in the creation of transcriptions of speech data. An example of the realigned phrase "tekstilės ir medienos" is given in Figure 2.

The whole realigned sentence is placed in an annotation file of the corresponding speech signal. Figure 2 shows the boundaries of separate phonemes and words present in the speech signal. The main object of the research was setting the boundaries of words and development of word-level annotations, while the boundaries of phonemes were not investigated. Two-level annotations were available: phoneme-level and

word-level, at the end of the first stage, but only word-level annotations were processed in the following stage 2. Phoneme-level annotations could be processed in the next step of development and improvement of LRN 1.

```
0 400000 t' tekstile3s
400000 900000 e
900000 1400000 k'
1400000 2300000 s'
2300000 2900000 t'
2900000 3300000 i0
3300000 3900000 l'
3900000 4400000 e3:
4400000 5600000 s
5600000 6300000 ir1 ir
6300000 7600000 m' medienos
7600000 7900000 e
7900000 8600000 d'
8600000 9800000 i:2e
9800000 11100000 n
11100000 11400000 o:
11400000 13400000 s
```

Figure 2. An example of the realigned phrase "tekstilės ir medienos" from the HTK annotation file

After the speech data of the whole speech corpus have been realigned and word-level annotations developed automatically, the following stage required to ensure that annotations of speech data had correct boundaries of words. Since the tools of HTK have limited options for a comprehensive investigation of speech data in manual setting of boundaries, another tool – Praat [6] – was used. Hence, the second stage included development of tools for converting formats of annotation files from HTK to Praat, and, vice versa, as well as examination and correction of annotations by hand. These processes are described in the next section.

3.2. Stage 2

In Praat, the annotation data are in a TextGrid object, which is stored later as a TextGrid annotation file. An example of the Praat TextGrid annotation file corresponding to the HTK annotation file, shown in Figure 2, is given in Figure 3. The TextGrid object consists of a number of interval tiers, where each interval tier represents a connected sequence of labelled intervals with boundaries in-between. Two interval tiers, "phones" and "words", are shown in Figure 3, which represent the phoneme-level transcription and word-level transcription of the phrase "tekstilės ir medienos", respectively. The time boundaries of labelled intervals in the Praat annotation file are measured in seconds, while in HTK – in 100ns units. The conversion of time units was therefore performed by the following relation

$$t_P = 0.1^7 t_H ,$$

where t_H denotes the time in HTK time units, and t_P – the time in Praat time units.

In order to perform a conversion from the HTK format of annotation files to the Praat format of annotation files and vice versa, two tools – Perl scripts "lab2tg.pl" and "tg2lab.pl" – have been developed. Various options for batch conversion and manipulation of annotation files are included in the developed tools. The tools are free and can be freely downloaded from the Comprehensive Perl Archive Network [10].

```
File type = "ooTextFile"
Object class = "TextGrid"

xmin = 0
xmax = 1.34
tiers? <exists>
size = 2
item []:
    item [1]:
        class = "IntervalTier"
        name = "phones"
        xmin = 0
        xmax = 1.34
        intervals: size = 17
        intervals [1]:
            xmin = 0
            xmax = 0.04
            text = "t'"
        intervals [2]:
            xmin = 0.04
            xmax = 0.09
            text = "e"
        ...
        intervals [17]:
            xmin = 1.14
            xmax = 1.34
            text = "s"
    item [2]:
        class = "IntervalTier"
        name = "words"
        xmin = 0
        xmax = 1.34
        intervals: size = 3
        intervals [1]:
            xmin = 0
            xmax = 0.56
            text = "tekstile3s"
        intervals [2]:
            xmin = 0.56
            xmax = 0.63
            text = "ir"
        intervals [3]:
            xmin = 0.63
            xmax = 1.34
            text = "medienos"
```

Figure 3. An example of the phrase "tekstilės ir medienos" from the Praat annotation file

The correction of annotation files by hand became possible after developing the above-mentioned tools for converting the formats of annotation data. A pair of annotation and the respective speech data files was used to compare automatically set boundaries of words (from the annotation file) with acoustic charac-

teristics of the boundaries (produced by the speech data file). Praat was used to this end. Figure 4 shows the main window of Praat where corrections have been made.

Helpful options of Praat that enable the comprehensive investigation of match between annotations and speech data are a spectrogram and a pitch contour (blue line) that could be configured, if necessary. Since the phoneme-level annotations were not investigated, the analysis of formant contour has not been made.

Annotation files were corrected in the following manner. The investigation of boundaries of each word started with listening to words and visual inspection of boundaries. The pitch contour was examined if there was some suspicion of an inexact boundary.

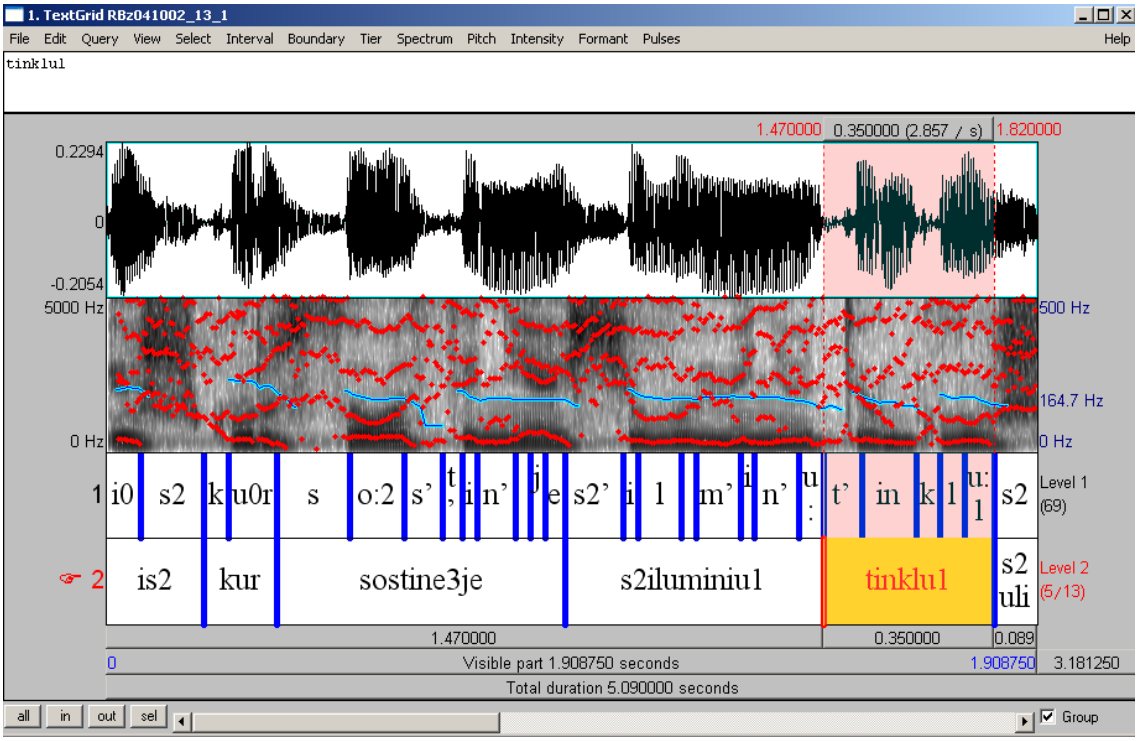


Figure 4. The stage of correction of annotation files. There are two pictures of the same sentence: speech waveform and spectrogram (with a pitch contour). Two annotation levels are shown below: phoneme-level and word-level. The analyzed word is marked and highlighted

The correction of annotation files was applied to a small segment of the speech corpus for two reasons:

1. Time consumption, and
2. Small amount of the corrections made.

The latter reason should be stressed as it indicates that Stage 1 or the process of realignment of speech data resulted well and Stage 2 is not necessary (this applies only to word-level annotations). However, the corrections of annotation by hand have been made. The next section describes some patterns of corrections.

3.3.The main mistakes of automatic transcriptions

The main patterns of correction of the word-level annotations of speech data were as follows:

1. Inserting of shorter pauses marked by special conventional words *_pauze* between two vowels.
2. Inexact boundaries of words starting or ending with the consonant r. The corrections were made broadening the segment of the phoneme *r*. The examples of two-word phrases are listed below (symbol ← or → denotes the direction of broadening):

aprūpinti ←rinką
stiprinti ← ryšius
pro ← rusiškos
dar → trys
atmeta ← Rusijos
audito ← rūmai
pernai ← rudenį

3. Inexact boundary of a word beginning with *im* or *ju*. The corrections were made by broadening the segment of the above-mentioned phoneme combinations. The examples of two-word phrases are listed below (symbol ← or → denotes direction of broadening):

Austrijos ← imperatorius
socialdemokratas ← Julius

4. Disappearance of the consonant *s* in collision with consonants *š* and *ž*. The examples of two word phrases are listed below (symbol ← or → denotes the direction of broadening and ~~s~~ – disappearance of *s*):

taršo~~s~~ ←židinio

aukojamo~~s~~ ← šventos

išgaunančia~~s~~ ← šalis

Baltijo~~s~~ ← šalių

5. Inexact boundary of the word that follows the word ending with *s* or *š*. The segments of phonemes *s*, *š* are too broad and capture a segment of the following word. The examples of two word phrases are listed below (symbol ← denotes the direction of broadening):

teismas ← nutarė

jis ← nevykdys

kaltinamas ← netikra

iš ← jo

6. Inexact boundary of words that collide with two vowels. The examples of two word phrases are listed below (symbol ← denotes the direction of broadening):

antiteroristinę → operaciją

sąjunga ←ir

4. Conclusions

Lithuanian continuous speech corpus LRN 1 and the process of annotation of the corpus have been introduced in this article. The major LRN 1 characteristics are summarized in Table 1.

The corpus presented is of ~21 hour duration with two-level annotations. New time-aligned word-level annotations of speech signals have been obtained after automatic model realignment and subsequent manual correction of annotations. The two-stage process, methods, and tools for developing these annotations, and the main patterns of correction are described here as well.

The different level of annotations provided by a corpus is a valuable linguistic resource and can broaden the scope of scientific research. We hope that the improvement of LRN 1 will be helpful for a more comprehensive research of Lithuanian speech recognition.

Table 1. Major LRN 1 characteristics

Criterion	LRN 1 Characteristics
Speech type	Continuous
Speech content	Read broadcast news
Annotation	Sentence-level and time-aligned word-level transcriptions
Number of speakers	31
Sampling	44 kHz
Quantization	16 b
Channels	Mono
Training data	18 h 09 min. (13 341 sentences)
Development test data	37 min. (736 sentences)
Evaluation test data	2 h 04 min. (1 755 sentences)
Full vocabulary	28 386 words
Phone set	74 simple phones, 156 diphthongs, 3 pseudo phones (including softness and lexical stress annotation)

References

[1] **A. Raškinis, G. Raškinis, A. Kazlauskienė**. Universal annotated VDU Lithuanian speech corpus. *Proceedings of Information Technologies* 2003, *KTU*, *Kaunas*, 2003, 28-34 (*in Lithuanian*).

[2] **A. Rudžionis, V. Rudžionis**. Lithuanian speech database LTDIGITS. *Proceedings of* 3*th International Conference on Language Resources and Evaluation* 2002, *Las Palmas, Spain*, 2002, 877-882.

[3] **S. Laurinčiukaitė, D. Šilingas, M. Skripkauskas, L. Telksnys**. Lithuanian Continuous Speech Corpus LRN 0.1: Design and Potential Applications. *Information Technology and Control,* 2006, *Vol*. 35, 431-440.

[4] **A. Geumann A.** Towards a new level of annotation detail of multilingual speech corpora. *Proceedings of* 8*th International Conference on Spoken Language Processing / Interspeech,* 2004, 1096-1099. *Available from WWW: http://muster.ucd.ie/pubs/Geumann ICSLP2004.pdf*, [cited 2008 08 21].

[5] **R. Kelly, J. Carson-Berndsen**. Semi-Automatic Phonological Annotations of Speech by Grammatical Inference. *Proceedings of the Workshop on Annotation Science, 5th International Conference on Language Resources and Evaluation,* Genoa, Italy, 2006, 1-8.

[6] **Boersma P., D. Weenik.** Praat, a system for doing phonetics by computer (version 3.4). *Technical Report 132*, Institute of Phonetic Sciences of the University of Amsterdam. *Available from WWW*: *www.praat.org*, [*cited* 2008 09 06].

[7] **HTK toolkit**. *Available from WWW*: *http://htk.eng.cam.ac.uk/*, [*cited* 2003 10 10].

[8] **D. Šilingas, S. Laurinčiukaitė, L. Telksnys.** Towards Acoustic Modelling of Lithuanian Speech. *Proceedings of 9th International Conference on Speech and Computer „SPECOM 2004“*, *Sankt Petersburg: Anatolya,* 2004, 326-333.

[9] **S. Laurinčiukaitė.** Acoustic Modelling of Lithuanian Speech Recognition. *PhD, Vilnius*, 2008, 108.

[10] **Comprehensive Perl Archive Network.** *http://www.cpan.org*, [*cited* 2008 08 21].

Received January 2009.

DOI: 10.5755/j01.itc.38.3.12122