ISSN 1392 – 124X INFORMATION TECHNOLOGY AND CONTROL, 2006, Vol. 35, No. 4

LITHUANIAN CONTINUOUS SPEECH CORPUS LRN 0.1: DESIGN AND POTENTIAL APPLICATIONS

Sigita Laurinčiukaitė[1], Darius Šilingas[2], Mantas Skripkauskas[1], Laimutis Telksnys[1,2]

[1]*Institute of Mathematics and Informatics, Recognition Processes Department*
A.Goštauto Str. 12-204, LT-01108 Vilnius, Lithuania

[2]*Vytautas Magnus University, Department of Applied Informatics*
Vileikos Str. 8-409, LT-44404 Kaunas, Lithuania

Abstract. This paper presents design, development and contents of Lithuanian continuous speech corpus LRN 0.1 (Lithuanian Radio News, prototype-version 0.1). The corpus contains 17 hours 23 minutes of records from radio broadcast news read by 31 speakers. The recorded material is segmented into sentence-length records that are divided into training, development, and evaluation sets. Speech recordings are accompanied by word level transcriptions and automatically generated word-to-phone lexicon. The corpus is designed for the constructing and evaluating speaker-independent continuous speech recognition systems, and may also be used for linguistic research.

Keywords: speech corpus, speech annotation, phonetic units, text-to-phone transcription.

1. Introduction

Speech is a means for human communication. In the recent twenty years, the focus on developing language technologies accelerated the research of sound generation and perception processes. However, despite of a huge effort put into speech processing research and implementations, we still face many problems that limit practical applications of speech technologies. Specifics of phonological, syntactic and prosodic structure of the language, differences in vocal tracts of speakers, different environments and communication channels cause high variability of speech signals, e.g. signal-to-noise ratio, duration of phonetic units, formants and energy of acoustic segments. To investigate these aspects, one needs a large collection of speech records. Therefore specialized speech databases, more often called *speech corpora*, became a fundamental necessity for scientific research. Since most modern speech recognition methods are based on statistical models, construction of practical systems also requires large amounts of training data and considerable amount of test data for tuning parameters. The releases of different corpora – TI-DIGITS, TIMIT, WSJ, SWITCHBOARD, and others – mark the milestones of speech recognition technology evolution. They helped to accelerate the progress in speech technologies by enabling comparative evaluation of speech recognition algorithms and systems, which promoted cooperation of researchers, sharing and evolving ideas, and allowed common effort in achieving better performance of speech recognition systems [2].

2. A survey of existing speech corpora

In [12], authors indicate three trends in development of speech corpora according to their purpose: 1) for practical applications; 2) for preserving characteristics of national acoustics; 3) for profound investigation of speech acoustic.

Majority of speech corpora are designed to provide enough data of speech acoustics for the development and evaluation of automatic speech recognition systems. In order to develop an efficient speech recognition system, it's necessary to identify and capture in models the properties of a particular language. Languages differ in phonetic systems, grammar and syntax. Therefore, it is necessary to develop language-specific speech corpora that appropriately represent the important features of the language. Existence of such speech corpora is a crucial point in further development of speech recognition technologies.

Linguistic Data Consortium (LDC), an international organization founded in 1992, specializes in moderating efforts of collecting speech and text corpora, making them publicly available, and presenting specialized information about the corpora. The trends in scientific research determine the trends in design of speech corpora. Currently, the majority of available speech corpora contain telephone, microphone and broadcast speech recordings. According to LDC, the majority of corpora are developed for English. Other languages that are rarely used at international level have smaller representative speech corpora. Lithuanian speech corpus LRN 0.1 presented in this paper is

similar to Czech (Czech Broadcast News Speech), Spanish (Spanish Broadcast News Speech, Hub-4NE) and Chinese (Mandarin Broadcast News Speech, Hub-4NE) [7] broadcast news speech corpora. Each of the mentioned corpora contains from 30 to 50 hours of national radio news records.

There are a few speech corpora for Lithuanian. A concise survey of existing speech corpora for Lithuanian is given in [15]. We will try to highlight what these speech corpora have in common and what makes them exclusive. Different universities and scientific laboratories construct and use for research their own speech corpora. Major Lithuanian speech corpora were gathered at Kaunas University of Technology (KUT) – LTDIGITS, [13], and Vytautas Magnus University (VMU), [5]. These speech corpora consist of speech recordings from single or multiple speakers. They also include lexicons, verified speech transcriptions, and automatically generated phoneme and word alignments for speech signals. Some of them come with additional software for searching speech recordings in the corpus. These features are state-of-the-art in modern speech corpora construction. Lithuanian speech corpora differ in duration (0.5–21 hours), number of speakers (1–350), speech type (isolated words, connected words, continuous speech), and lexicon size (50–32 000 words). The more homogeneous speech signals are collected, the deeper investigation of speech acoustic can be made. On the other hand, homogeneity of speech corpus puts more restrictions on practical applications. Almost all mentioned speech corpora enable thorough investigation of speech acoustics. However, real world applications require systems to cope with large vocabulary continuous speech and multiple speakers. Before LRN 0.1, none of the existing corpora contained enough continuous speech data from multiple speakers that are necessary for constructing speaker-independent large vocabulary continuous speech recognition systems. The structure of LRN 0.1 corpus, conventions for textual alignment, chosen phone set was based on the best ideas found in the analyzed speech corpora. A tool for the automatic Lithuanian word-to-phone transcription is a new tool, which was created as a utility tool dedicated for developing LRN 0.1.

3. LRN 0.1 Corpus Design

In this section we present speech corpus LRN 0.1 developed at the Institute of Mathematics and Informatics (IMI) in cooperation with VMU. The corpus is being incrementally expanded by ~3–4 hours of speech data every year and at the moment contains 17 hours 23 minutes of speech. Next we describe the construction of LRN 0.1 corpus and present its main characteristics: information about recorded material, speaker characteristics, corpus structure, phoneme set, pronunciation dictionary, and other details.

3.1. Corpus Construction

The speech corpus was constructed in the following stages:

1. Preparation of speech waveform files;
2. Annotation of speech records;
3. Extraction of lexicon;
4. Construction of pronunciation dictionary;
5. Validation and experiments.

The corpus contains speech samples from the news broadcasts by Lithuanian Radio in 2003-2004. The broadcast news signals were received at the IMI and recorded digitally at the sampling rate of 44 kHz using 16 bits resolution and mono channel. Down sampling or compression of resolution may always be performed by the end users based on their needs. Each news session lasted about 10 minutes. Later, the session records were manually split into sentence-length Microsoft wave files. The majority of recorded speech was of high quality – no noise, clear and correct pronunciation.

Lithuanian Radio supplied corresponding texts that were used as facilitating means for manual annotation of records in word level. Records that didn't have corresponding texts were manually transcribed. All numerical data and abbreviations in annotations were written out in words, and syntax marks were removed.

After word-level annotations were prepared, we automatically retrieved the lexicon – list of unique words from all annotations.

We also prepared a pronunciation dictionary, which is necessary for constructing speech recognition system using acoustic models based on phonetic units. Every word from lexicon was manually transcribed to phones including lexical stress. Phone transcriptions were verified using a tool implemented by one of the paper authors (see subsection 3.7).

Validation of the speech corpus was performed when running speech recognition experiments using data from LRN0 corpus – the prototype of LRN 0.1 [16, 6, 17, 8]. When running experiments, multiple errors (various typos, missing elements) were found and fixed.

3.2. Speech Content

Recorded speech covers political, economical, cultural, and sport areas of local and foreign affairs. There are a lot of local and foreign names. This caused some problems for preparation of word-to-phone lexicon as transcription of foreign names requires using elements of other phonetic system then Lithuanian. Moreover, exact pronunciations were not known and pronunciation of foreign names differed in distinct recordings. News about sport events is also problematic issue since the speaking rate of sport newsreader is very high, which causes a lot of mispronunciations. Furthermore, sports news contain unusually many foreign names.

According to requirements for speech homogeneity in pronunciation and speaking rate, there might be a need to filter data leaving out foreign or sport events. To draw a boundary in speech corpus between reports of local and foreign events is rather difficult, but recordings of sports news could be easily excluded form speech corpus rather easily as they were read by two newsreaders, and newsreaders identities are encoded in file names.

3.3. Speaker Characteristics

The selection of the speakers was limited to the newsreaders only. There were 31 speakers: 17 females and 14 male. As shown in Figure 1, the distribution of speech records is not equally distributed among the speakers. Actually, speech recordings of 13 speakers make up 93% of entire speech corpus.

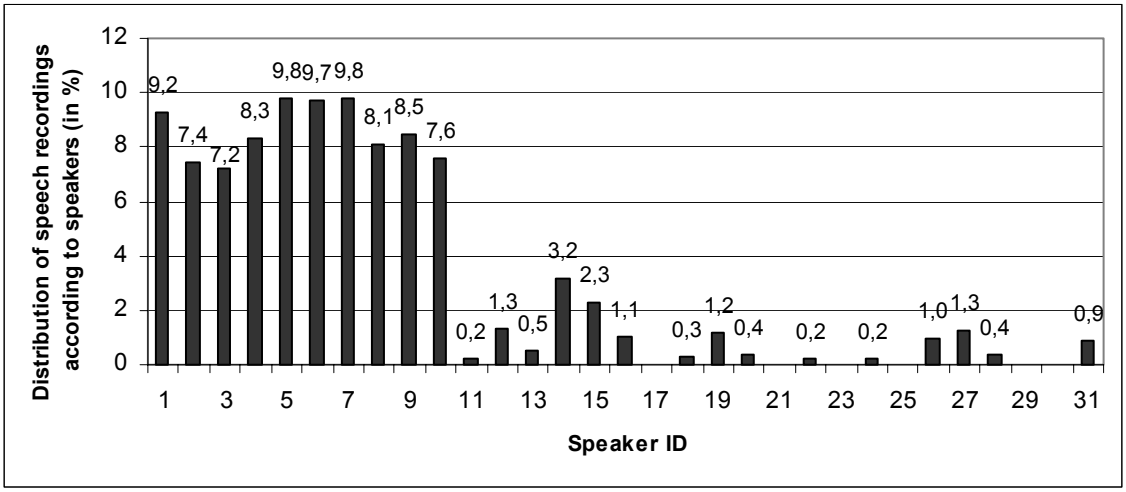


Figure 1. Distribution of records according to speakers in corpus (in %)

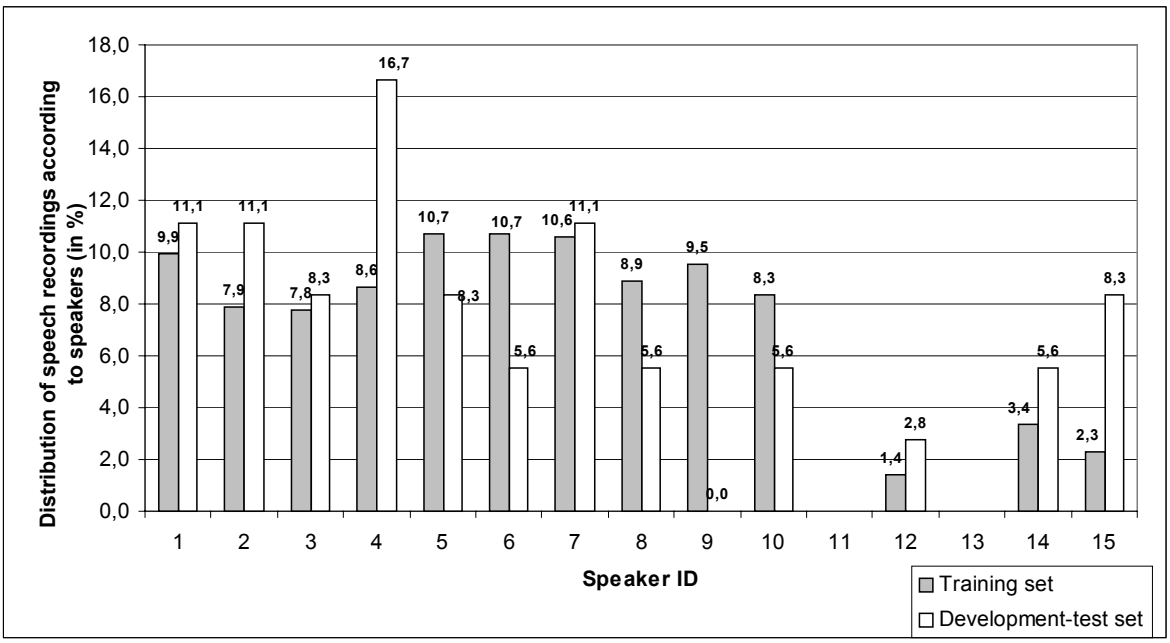


Figure 2. Distribution of speech recordings according to speakers in training, development sets. The percentage of training and development sets are counted separately

3.4. Structure of the Corpus

Speech corpus is partitioned into training, development and evaluation data sets. Speakers for training and development sets are the same. These speakers were selected for training task, as their recordings make up bigger part of speech corpus. Speakers of evaluation set are different. Distribution of speakers in training and development sets is shown in Figure 2, in evaluation set in Figure 3. Information about the speakers of these three sets is shown in Table 1.

Speech data in every set are grouped by speaker, but the structure of corpus can be easily changed as construction of data file names enables re-grouping of speech records and their annotation files. The name of every file in corpus encodes different information attributes (see subsection 3.5), which makes the structure of corpus to be flexible.

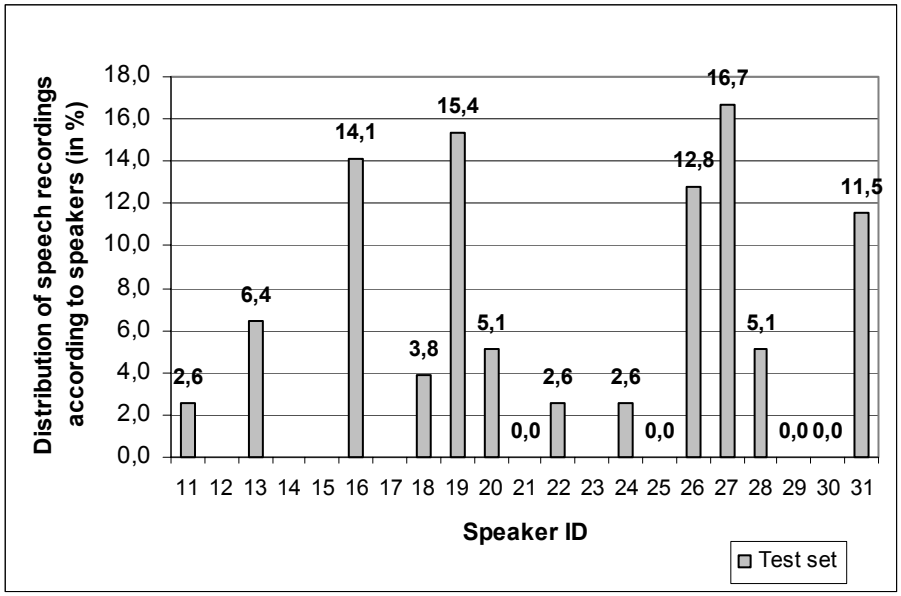


Figure 3. Distribution of speech recordings according to speakers in evaluation set

Table 1. Comparative characteristics of training, development and evaluation sets

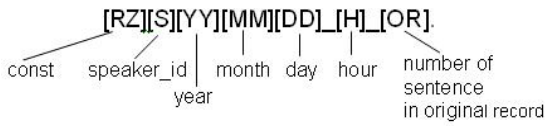
Name of the speech corpus set	Duration (h:min)[1]	Number of sentences	Number of speakers
Training set	15:26	11 127	13
Development set	0:37	736	13
Evaluation set	1:18	1 098	18
Total	**17:23**	**12 961**	**31**

3.5. Data Formats

Data types are differentiated by filename extensions. Extension **.wav* denotes waveform of sentence and extension **.txt* denotes word level annotation of sentence. All files associated with the same sentence have the same base name. The filename format is:

[sentence_id].[extension], where

[sentence_id] is composed of

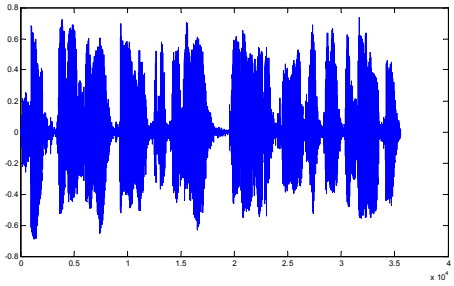


An example of waveform and text file pair is given in Figure 4.

As mentioned at the beginning of this section, the corpus is being expanded every year by ~3-4 hours. To secure interest of researchers to maintain earlier version of corpus, new data are easily distinguished as second element of *const* in *sentence_id* for every addition has further symbol of alphabet, starting with *a*. So far we have made two additions.

[1] Durations of the sets are approximate.

RZ10625_18_5.wav – sound record (data of waveform presented in figure)



RZ10625_18_5.txt – word-level transcription

s2iuos pareigu4nus seimo pritarimu skiria ir atleidz2ia prezidentas

Figure 4. An example of waveform and text file pair[2]

3.6. Phone Set

Most speech recognition systems are based on phonemes. This requires pronunciation for all words in test and training data. A common approach is to use lexicon with word-to-phone transcriptions. The fundamental issue is choosing a base phoneme set, which has to be defined before constructing pronunciation dictionary.

After investigation of existing phonetic systems [10, 19, 3, 11] that differ in few phonemes, phonetic system SAMPA-LT [11] was chosen as the most suitable for speech recognition purposes. The phoneme set with occurrence samples are listed in Appendix 1. The set includes 11 vowels (5 short and 6 long), 45 consonants (6 pairs of plosives, 4 pairs of affricates, 7

[2] Codification of Lithuanian symbols is used in the transcriptions of text files and the lexicon. All codified symbols are listed in Appendix 1.

pairs of fricatives, and 5 pairs plus 1 of liquids), 7 pure diphthongs and 16 mixed diphthongs. Softness of consonants and linguistic stress were also included as differentiating features.

3.7. Lexicon

There were 18 374 distinct words in the corpus LRN0 lexicon. Some examples are given in Figure 5.

phones.dic – word-to-phone pronunciation dictionary	
s2iuos	**s2' uo:1 s**
pareigu4nus	**p a r' ei g u:1 n u s**
seimo	**s' ei:1 m o:**
pritarimu	**p' r' i t a r' i m u0**
skiria	**s' k' i0 r' e**
ir ir1	
atleidz2ia	**a t' l' e:2i dz2' e**
prezidentas	**p' r' e z' i d' en1 t a s**

Figure 5. Some examples of word-to-phone transcriptions

Most of words were common Lithuanian words (e.g. “anksti”, “dukra”). There were also a lot of proper nouns (e.g. “Haityje”), international (e.g. “oficialiai”), mispronounced words (e.g. “sussunitu”), acronyms (e.g. “JAV”, “OFK”). All those words were transcribed manually. However, it was difficult to validate pronunciations. For this purpose we implemented a tool for automatic Lithuanian word-to-phone transcription. It also helped to transcribe the new words, which appeared during development of the corpus. The tool implementation combines formal pronunciation rules and exceptional word pronunciation dictionary. Formal pronunciation rules or just formal rules were used to describe relationships between phonemes and segments of text (e.g. word “lankas” is converted by pronunciation rules into the phoneme group “l aw k a s”). The most popular ones were related to subjects like:

1. assimilation,
   (e. g. “apdaila” → “a b d ai l a”)
2. diphthongs,
   (e. g. “laukas” → “l au k a s”)
3. vowel and consonant combination,
   (e. g. “rangas” → “r an g as”,
   “garas” → “g a r a s”)
4. palatalization vowels,
   (e. g. “liepti” → “l‘ ie p‘ t‘ i”)
5. long vowels,
   (e. g. “rūko” → “r u: k o”)
7. adjacent consonants,
   (e. g. “iššoko” → “i š o k o”,
   “išstumdavo” → “i s t um d a v o”,
   “apsčiai” → “a p’ š’ č’ ei”,
   “vabzdžiai” → “v a b’ ž’ dž’ ei”)
8. „n” before „k” or „g”,
   (e. g. “anga” → “aw’ g a”)
9. transformation of single or group of capital letters into phonemes,
   (e. g. “NKVD” → “en k a v’ ė d’ ė”)
10. transformation of numbers into phonemes, (e. g. “D-6” → “d’ ė š’ e š’ i”).

Formal rules were taken from Lithuanian acoustics theory [19, 1], and were implemented using regular expression technique.

Pronunciation dictionary contains linguistic stress and a lot of word pronunciation characteristics, which were exceptions to the formal pronunciation rules. It was derived from the other three dictionaries, [4, 9, 20], and is used to check for exceptions to the formal rules. Linguistic stress was assigned to transcribed word using only this dictionary. The main problem in checking and assignment procedures were how to find matching dictionary word for the transcribed one. Therefore, each line of dictionary consists of two words – headword and the second one – as a pronunciation correction of the headword, where we can also find a stress mark and position. The match between dictionary headword and transcribed word was implemented using partial morphological analysis. For this purpose, the set of 10 000 different Lithuanian roots and about 30 most common prefixes were used, which were taken from the composition analysis part of [4].

The transcription system was tested by comparing its results with manually obtained ones. It revealed that transcription of Lithuanian text with the exception dictionary in addition to formal pronunciation rules reduced errors for:

- Broadcast news text (from LRN0 corpus), from 2.2% to 0.9%,
- Nonfiction literature text wordbook, from 1.4% to 0.8% errors,
- Medical scientific text, from 5.2% to 2.1%.

At the moment lexicon of LRN 0.1 is processed.

3.8. Special Markings

Long and shorter pauses were marked by special conventional words *_tyla* and *_pauze*, respectively. Some speech recordings contain noise of aspirations that was marked by conventional word *_ikvepimas*.

Mispronunciations were marked by adding symbol “_” to the beginning of the word. Development set excludes records with such cases. Evaluation set is divided into data without and with mispronunciations.

4. Potential Applications

LRN 0.1 corpus may be used for various speech recognition system research and practical implementation purposes. Figure 6 shows the place of LRN 0.1 in a speech recognition system.

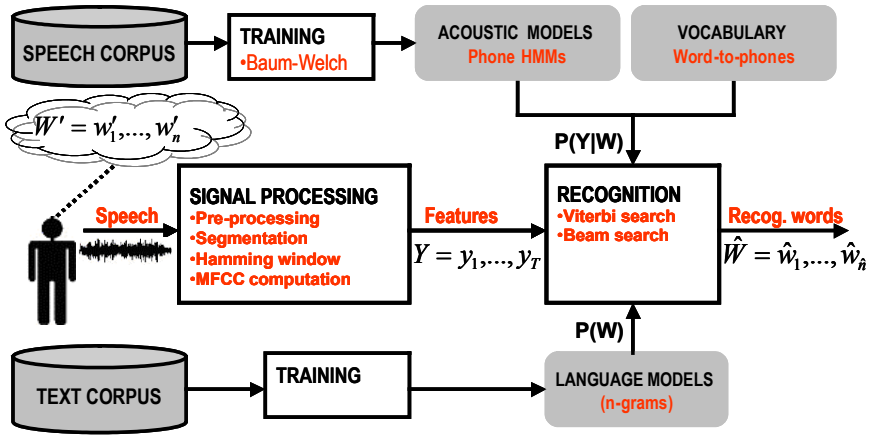


Figure 6. Continuous speech recognition framework

First of all, corpus defines reasonable amount of data for scientific research of algorithms and properties of Lithuanian speech recognition systems, such as modeling acoustic modeling units – graphemes, phonemes, triphones, and syllables [15, 6, 17, 14], choosing most effective acoustic feature vectors (energy, MFCC, LPC, PLP, combined, various lengths), comparing hypothesis search algorithms, evaluating language models [18].

Without having the data of size similar to LRN 0.1 such research makes little sense since its results would be dependent on data nature and not reliable. Although primary dedication of this corpus is for speech recognition research, it may be used for linguistic research as well. For example, one can investigate properties of specific phones such as energy, pitch, and formants, in different contexts. The corpus defines training, development, and evaluation data sets. Using these sets consistently allows comparing performance of different speech recognition engines that may be based on different methods such as HMM, neural networks, dynamic time warping using template recognition, and hybrid approaches, or differently tuned, e.g. using different acoustic feature vectors, acoustic modeling units, pruning value and other thresholds. It may also be used for analysis of influence of different language models to the recognition error rate.

For practical implementations, LRN 0.1 corpus enables training acoustic models that later can be used in the product. Also, it allows validation of the product performance such as word error rates and speed. Basically, building a practical product requires a lot of data for researching alternatives and definitely a large amount of speech for training models. Thus using a corpus is a must. However, it is necessary to mention that models are very sensitive to the nature of training data. Therefore, for speech recognition products dedicated to recognizing speech of different nature, e.g. spontaneous speech, or in different environments, e.g. noisy speech, low quality channel, different types of corpora should be collected and used. Despite the necessity of different corpora, the data of LRN 0.1 may still be used as a substitute after making transformations either to data or to the trained models.

Pilot speech recognition experiments using LRN0 data were performed using HTK system. These experiments are described in detail in [7]. The word error rate (WER) of 14.89% was achieved for 5500 word dictionary task. The phone error rate for the same task was 4.59%. These values were achieved without using language model. Adding a simple bigram Lithuanian language model enabled to decrease WER down to 9.75%, and the manual post-processing fixing dropped it down to 5.61%. Such an error rate is already sufficient for some practical applications, and it definitely can be improved using more sofisticated language models and post-processing.

There are also potential applications for which the size of corpus is still too small. For example, research of recognition of specific phones or triphones requires much larger development and evaluation sets that should also be linguistically representative, i.e. they should contain enough samples of even rare phonetic units. Therefore we are planning to continue increasing the size of the corpus in the future.

5. Conclusions

We have presented the new Lithuanian continuous speech corpus LRN 0.1. It has been built to satisfy the needs of speech recognition researchers in large quantity of research data. Major LRN 0.1 characteristics are summarized in Table 2.

We also described the corpus design and development: preparation of speech files, lexicon, pronunciation dictionary, data structure, validation of lexicon, and specifics of corpus speech samples.

The potential applications point out how the corpus can be used in scientific research and practical implementations of speech recognition systems.

Acknowledgments. The authors thank to all participants for their input to the construction of the speech corpus, especially to Mark Filipovič, Gintautas Tamulevičius and Tomas Lygutas.

Table 2. Major LRN 0.1 characteristics

Criterion	LRN 0.1 Characteristics
Speech type	Continuous
Speech content	Read broadcast news
Annotation	Orthogonal word-level transcriptions
Number of speakers	31
Sampling	44 kHz
Quantization	16 b
Channels	Mono
Training data	15 h 26 min. (11 127 sentences)
Development test data	37 min. (736 sentences)
Evaluation test data	1 h 18 min. (1 098 sentences)
Full vocabulary	18 374 words
Evaluation vocabulary	5 500 words
Phone set	74 simple phones, 156 diphthongs, 3 pseudo phones (including softness and lexical stress annotation)

References

[1] V. Ambrazas, (edit). Grammar of Modern Lithuanian language. *Vilnius: Science & Encyclopedia Publishing Institute,* 1994, (in Lithuanian).

[2] R.A. Cole, (edit.). Survey of the State of the Art in Human Language Technology, 1996. *Available from WWW: <http://cslu.cse.ogi.edu/HLTsurvey/ HLTsurvey.htm>* [*cited* 2006 07 03].

[3] A. Girdenis. Theoretical basics of Lithuanian Phonology. *Vilnius: Mokslo ir enciklopedijų leidybos institutas*, 2003, (*in Lithuanian*).

[4] S. Keinys, (edit.). Dictionary of Modern Lithuanian language, (4*th edition*). *Vilnius: Science & Encyclopedia Publishing Institute*, 2000, (*in Lithuanian*).

[5] D.Kuliešienė, G. Grigonytė. The Potential of the Lithuanian Speech Corpus. 2005. *Available from WWW:<http://www.speech.kth.se/~rolf/NGSLT/gslt_papers_2005/st_paper_version_2.pdf>*, [*cited* 2006 07 03].

[6] S. Laurinčiukaitė, A. Lipeika. Syllable-Phoneme based Continuous Speech Recognition. *Proceedings of Electronics and electrical engineering* 2006. ISSN 1392-1215, 2006, *No*. 6 (70), 91-94.

[7] Linguistic Data Consortium. University of Pennsylvania, 2005. *Available from WWW: <http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2004S01>*, [*cited* 2006 09 07].

[8] T. Lygutas. Investigations of Lithuanian Phone Sets. *Proceedings of Information Sciences*, 2005, 220-226, (*in Lithuanian*).

[9] National Committiee of Lithuanian language and Institute of Mathematics and Informatics. Terminology of Lithuanian language. *Vilnius, MII*, 2005. *Available from WWW: <http://www.terminynas. lt>*, [*cited* 2006 07 03].

[10] A. Pakerys. Phonetics of Common Lithuanian. *Vilnius: Enciklopedija*, 2003, (*in Lithuanian*).

[11] A. Raškinis, G. Raškinis, A. Kazlauskienė. Speech Assessment Methods Phonetic Alphabet (SAMPA) for Encoding Transcriptions of Lithuanian Speech Corpora. *Proceedings of Information technology and control. Kaunas: Technology*, 2003, *No*. 4(29), 52-55.

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

[13] A. Rudžionis, V. Rudžionis. Lithuanian speech database LTDIGITS. *Proceedings of LREC 2002*, *Las Palmas, Spain*, 2002, 877-882.

[14] D. Šilingas. Choosing Acoustic Modeling Units for Lithuanian Continuous Speech Recogniton Based on Hidden Markov Models. *PhD Thesis, Vytautas Magnus Unversity, Kaunas*, 2006.

[15] D. Šilingas, G. Raškinis, L. Telksnys. Review of Lithuanian Speech and Language Processing. *Proceedings of Human Language Technologies – the Baltic Perspective* 2004, *Riga*, 2004, 144-150.

[16] D. Šilingas, S. Laurinčiukaitė, L. Telksnys. Towards Acoustic Modeling of Lithuanian Speech. *Proceedings of SPECOM* 2004, *St. Petersburg: SPIIRAS*, 2004, 326-332.

[17] D. Šilingas, S. Laurinčiukaitė, L. Telksnys. A Technique for Choosing Efficient Acoustic Modeling Units for Lithuanian Continuous Speech Recognition. *Proceedings of SPECOM* 2006, *St. Petersburg: SPIIRAS*, 2006, 61-66.

[18] A. Vaičiūnas, V. Kaminskas, G. Raškinis. Statistical Language Models of Lithuanian Based on Word Clustering and Morphological Decomposition. *Vilnius: Informatica* 15(4), 2004, 565-580.

[19] V. Vaitkevičiūtė. Basics of Lithuanian Pronunciations and dictionary. *Vilnius: Pradai*, 2001, (*in Lithuanian*).

[20] V. Vaitkevičiūtė. Dictionary of the International Words. *Vilnius: Žodynas*, 2001, (*in Lithuanian*).

Received August 2006.

Appendix 1

Table 3. Vowels

Alphabetical symbol	Symbol in corpora	Types of vowel (simple, stressed)	Pattern off transcribed word (word)
Vowel (short) /5/			
a	a	a a0	**a** k' i0 s (**a**kìs) p **a0** s' m' er' k' e3: (p**à**smerkė)
e	e	e e0	d' **e** v' in1 t a: (d**e**viñtą) p a t' **e0** g d a v o: (pat**è**kdavo)
i	i	i i0	e0 t' **i** k o: s (èt**i**kos) a p' **i0** p' l' e3: s2' e3: (ap**ì**plėšė)
o	o	o o0	**o** p **o** z' i0 c' i j' a (**o**p**o**zìcija) a t **o0** m' i n' o: (at**ò**minio)
u	u	u u0	p a k' i0 l **u** s (pakìl**u**s) g r **u0** p' e: (gr**ù**pę)
Vowel (long) /6/			
a	a:	a: a:1	– **a**:1 k' c' i j' a (**ã**kcija)
ą	a:	a: a:1	a:1 k' c' i j' **a:** (ãkcij**ą**) g' ir t **a:1** v' i m o: (girt**ã**vimo)
e	e:	e: e:1	– **e:1** s a m a (**ẽ**sama)
ę	e:	e: e:1 e:2	d' i0 d' e l' **e:** (dìdel**ę**) t' **e:1** s' t' i (t**ę̃**sti) b' r' **e:2** s t an' c2' u s (br**ę́**stančius)
ė	e3:	e3: e3:1 e3:2	a0 p t a r' **e3:** (àptar**ė**) s u k' r' **e3:1** t u s' o: (sukr**ė̃**tusio) s' v' **e3:2** r' e3: (sv**ė́**rė)
į	i:	i: i:1 i:2	s' k' r' i:1 d' **i:** (skrỹd**į**) s u g' r' **i:1** z2 u s' i (sugr**į̃**žusi) **i:2** g u l a (**į́**gula)
y	i:	i: i:1 i:2	g' i:2 d' **i:** t o: j' a (gỳd**y**toja) s2 o: v' i n' **i:1** s (šovin**ỹ**s) p u s2' **i:2** n a s (pušy**́**nas)
o	o:	o: o:1 o:2	p a:1 j' a m **o:** s (pãjam**o**s) p r **o:1** t o: (pr**õ**to) p r **o:2** g a (pr**ó**ga)
ū	u:	u: u:1 u:2	n uo d u0 g n **u:** s (nuodugn**ū**s) p a r' ei g **u**:1 n ai (pareig**ū̃**nai) p r a d **u:2** r' e3: (prad**ū́**rė)
ų	u:	u: u:1 u:2	m' i n' e r a:1 l **u:** (minerãl**ų**) p a s' **u:1** s' t' i (pasi**ų̃**sti) –

Table 4. Consonants

Alphabetical symbol	Symbol in corpus (hard, soft)	Pattern off transcribed word (word)
Consonants (plosive) /6 pairs/		
p	p p’	a0 **p** t ar t o: s (à**p**tartos) a **p'** t' i0 k o: (a**p**tìko)
b	b b’	a r a:1 **b** u: (arã**b**ų) ar' **b'** i0 t r a s (ar**b**ìtras)
t	t t’	c' en1 **t** r o: (ceñ**t**ro) c' en' **t'** r' e0 (cen**t**rė)
d	d d’	d' e:1 **d** a (dẽ**d**a) **d'** e3:1 **d'** e3: (**d**ẽ**d**ė)
k	k k’	g r a:1 f' i **k** u s (grãfi**k**us) f r a:1 **k'** c' i j' a (frã**k**cija)
g	g g’	i:2 s t ai **g** a (įstai**g**a) i: **g'** i:2 t' i (į**g**ýti)
Consonants (affricate) /4 pair /		
c	c c’	p r an **c** u:1 z ai (pran**c**ũzai) i m' i t a:1 **c'** i j' a (imitã**c**ija)
č	c2 c2’	g' in1 **c2** o: (giñčo) p' r' i v a **c2'** uo s' e0 (priva**č**iuosė)
dz	dz dz’	– k u **dz'** i: s (Ku**dz**ys)
dž	dz2 dz2’	**dz2** o0 r **dz2** a s (**Dž**ò**rdž**as) a t' i **dz2'** u0 s (ati**dž**iùs)
Consonants (fricative) /7 pair /		
f	f f’	in **f** o r m a:1 c' i j' a (in**f**ormãcija) d' el' **f'** i0 n u: (del**f**inų)
s	s s’	in' t' e r' e0 **s** am **s** (interè**s**am**s**) in' t' en' **s'** i: v' i0 (inten**s**yvì)
š	s2 s2’	k a:2r **s2** t o: (kár**š**to) k ar'1 **s2'** t' i s (kar**š**tis)
z	z z’	l a **z** d o: m' i0 s (la**z**domìs) k a **z'** i n o0 (ka**z**inò)
ž	z2 z2’	m a **z2** a m' e0 (ma**ž**amè) m a **z2'** e3:2 j' a (ma**ž**ėja)
ch	ch ch’	t' e **ch** n o l o0 g' i j' o: s (te**ch**nològijos) p' s' i0 **ch'** i k o: s (psì**ch**ikos)
h	h h’	al k o **h** o0 l' o: (alko**h**òlio) b e0 k' **h'** e m a s (Bèk**h**emas)
Consonants (liquid) /5 pair + 1/		
v	v v’	b u0 **v** o: (bùvo) a0 p' **v'** er' t' e3: (àp**v**erte3)
m	m m’	c' e r' e **m** o0 n' i j' a (cere**m**ònija) ch' e0 **m'** i j' o: s (chè**m**ijos)
n	n n’ w w’	d ai **n** o: m' i0 s (dai**n**omìs) f' i n a:1 l' i **n'** e3: s (finãli**n**ės) v' ie **w** k ar' t' i0 n' i s (vie**n**kartìnis) k o **w'** k' r' e t' i0 (ko**n**kretì)
l	l l’	b a:1 **l** o: (bã**l**o) a p' **l'** iw1 k (ap**l**iñk)
r	r r’	c u0 k **r** au s (cùk**r**aus) b a **r'** e0 (ba**r**è)
j	j'	g al v o: **j'** e0 (galvo**j**è)

Table 5. Diphthongs

Alphabetical symbol	Symbol in corpus (simple, soft)	Patterns of diphthongs (simple, stressed)
Diphthongs /7/		
ai	ai	ai (ai), a:2i (ái), ai:1 (aĩ)
au	au	au (au), a:2u (áu), au:1 (aũ),
ei	ei	ei (ei), e:2i (éi), ei:1 (eĩ)
eu	eu	eu (eu), e0u (èu)
ui	ui	ui (ui), u0i (ùi), ui:1 (uĩ),
ie	ie	ie (ie), i:2e (íe), ie:1 (iẽ),
uo	uo	uo (uo), u:2o (úo), uo:1 (uõ),
Mixed diphthongs /16/		
al	al al’	al (al), a:2l (ál), al1 (alȉ) al’ (al’), a:2l’ (ál’), al’1 (alȉ’)
am	am am’	am (am), a:2m (ám), am1 (amȁ) am’ (am’), a:2m’ (ám’), am’1 (amȁ’)
an	an an’ aw aw’	an (an), a:2n (án), an1 (añ) an’ (an’), a:2n’ (án’), an’1 (añ’) aw (aw), a:2w (án), aw1 (añ) aw’ (aw’), a:2w’ (án’), aw’1 (añ’)
ar	ar ar’	ar (ar), a:2r (ár), ar1 (arȅ) ar’ (ar’), a:2r’ (ár’), ar’1 (arȅ’)
el	el el’	el (el), e:2l (él), el1 (elȉ) el’ (el’), e:2l’ (él’), el’1 (elȉ’)
em	em em’	em (em), e:2m (ém), em1 (emȁ) em’ (em’), e:2m’ (ém’), em’1 (emȁ’)
en	en en’ ew ew’	en (en), e:2n (én), en1 (eñ) en’ (en’), e:2n’ (én’), en’1 (eñ’) ew (ew), e:2w (én), ew1 (eñ) ew’ (ew’), e:2w’ (én’), ew’1 (eñ’)
er	er er’	er (er), e:2r (ér), er1 (erȅ) er’ (er’), e:2r’ (ér’), er’1 (erȅ’)
il	il il’	il (il), i0l (ìl), il1 (ilȉ) il’ (il’), i0l’ (ìl’), il’1 (ilȉ’)
im	im im’	im (im), i0m (ìm), im1 (imȁ) im’ (im’), i0m’ (ìm’), im’1 (imȁ’)
in	in in’ iw iw’	in (in), i0n (ìn), in1 (iñ) in’ (in’), i0n’ (ìn’), in’1 (iñ’) iw (iw), i0w (ìn), iw1 (iñ) iw’ (iw’), i0w’ (ìn’), iw’1 (iñ’)
ir	ir ir’	ir (ir), i0r (ìr), ir1 (irȅ) ir’ (ir’), i0r’ (ìr’), ir’1 (irȅ’)
ul	ul ul’	ul (ul), u0l (ùl), ul1 (ulȉ) ul’ (ul’), u0l’ (ùl’), ul’1 (ulȉ’)
um	um um’	um (um), u0m (ùm), um1 (umȁ) um’ (um’), u0m’ (ùm’), um’1 (umȁ’)
un	un un’ uw uw’	un (un), u0n (ùn), un1 (uñ) un’ (un’), u0n’ (ùn’), un’1 (uñ’) uw (uw), u0w (ùn), uw1 (uñ) uw’ (uw’), u0w’ (ùn’), uw’1 (uñ’)
ur	ur ur’	ur (ur), u0r (ùr), ur1 (urȅ) ur’ (ur’), u0r’ (ùr’), ur’1 (urȅ’)