INFORMATICA, 2019, Vol. 30, No. 3, 573–593

DOI: http://dx.doi.org/10.15388/Informatica.2019.219

Comparison of Phonemic and Graphemic Word to Sub-Word Unit Mappings for Lithuanian Phone-Level Speech Transcription

Gailius RAŠKINIS[1,4] *, Gintarė PAŠKAUSKAITĖ[2],
Aušra SAUDARGIENĖ[1,3], Asta KAZLAUSKIENĖ[1],
Airenas VAIČIŪNAS[1]
[1]*Vytautas Magnus University, K. Donelaičio 58, LT-44248, Kaunas, Lithuania*
[2]*Kaunas University of Technology, K. Donelaičio 73, LT-44249, Kaunas, Lithuania*
[3]*Lithuanian University of Health Sciences, Eivenių 4, LT-50161, Kaunas Lithuania*
[4]*Recognisoft, Ltd., K. Donelaičio 79-1, LT-44249, Kaunas, Lithuania*
e-mail: gailius.raskinis@vdu.lt, gintare.paskauskaite@ktu.lt, ausra.saudargiene@lsmuni.lt, asta.kazlauskiene@vdu.lt, airenass@gmail.com

Received: June 2018; accepted: May 2019

Abstract. Conventional large vocabulary automatic speech recognition (ASR) systems require a mapping from words into sub-word units to generalize over the words that were absent in the training data and to enable the robust estimation of acoustic model parameters. This paper surveys the research done during the last 15 years on the topic of word to sub-word mappings for Lithuanian ASR systems. It also compares various phoneme and grapheme based mappings across a broad range of acoustic modelling techniques including monophone and triphone based Hidden Markov models (HMM), speaker adaptively trained HMMs, subspace gaussian mixture models (SGMM), feed-forward time delay neural network (TDNN), and state-of-the-art low frame rate bidirectional long short term memory (LFR BLSTM) recurrent deep neural network. Experimental comparisons are based on a 50-hour speech corpus. This paper shows that the best phone-based mapping significantly outperforms a grapheme-based mapping. It also shows that the lowest phone error rate of an ASR system is achieved by the phoneme-based lexicon that explicitly models syllable stress and represents diphthongs as single phonetic units.

Key words: speech recognition, grapheme, phoneme, G2P conversion, HMM, SGMM, TDNN, BLSTM, Lithuanian.

1. Introduction

Conventional large vocabulary automatic speech recognition (ASR) systems require a mapping from words into sub-word units to generalize over the words that were absent in the training data and to enable the robust estimation of acoustic model parameters. Mapping words into phones by constructing pronunciation dictionaries that take into account

*Corresponding author.

sound assimilation rules and coarticulation effects was the dominant approach for many years. This approach has the advantage of trying to match the process of speech production. Mapping words into graphemes (letters) is an alternative approach (Kanthak and Ney, 2002; Killer *et al.*, 2003) advocated by some recent studies (Collobert *et al.*, 2016). It has the advantage of skipping the process of dictionary build-up that is costly and requires an involvement of linguistic experts. Grapheme based ASR systems showed relatively good performance for Lithuanian ASR as well (Gales *et al.*, 2015; Lileikytė *et al.*, 2016; Alumae and Tilk, 2016; Salimbajevs and Kapočiūtė-Dzikienė, 2018). Finding the best lexicon of sub-word units for any particular language is a complex problem that can be answered only through an experimental investigation. ASR systems based on different word to sub-word unit mappings have to be built and their performance has to be compared. Much of the complexity originates from the fact that the optimum sub-word unit lexicon may depend on the size of the training corpus and on the setup of an ASR system, i.e. on selected acoustic modelling technique, amount of linguistic knowledge incorporated into the system, and performance comparison criteria. Experimental investigation is also costly in terms of computation time.

Multiple different word to sub-word unit mappings for the purposes of Lithuanian ASR were investigated and compared during the last 15 years. Studies addressing this topic often arrived to opposite conclusions or these conclusions were not supported by the tests of statistical significance. Thus, the practical question about which mapping should be chosen or tried first if someone has sizeable amounts of acoustic data (50 hours and more) and intends to build an ASR system remains open.

This study aims to obtain an additional insight into this question. The first distinction of this study is that we follow a "divide and conquer" approach to the ASR tuning. We eliminate lexical and syntactic-semantic layers of the ASR system and evaluate word to sub-word unit mappings on the basis of the performance of an acoustic model alone. Given that the language model (LM) and pronunciation dictionary are absent, we use Phone Error Rate (PER) rather than Word Error Rate (WER) as the ASR performance criterion. We believe that such approach makes our findings independent from the lexical content of the training/evaluation data. Second, we carefully prepare the data. Our investigations are based on a solid 50-hour speech corpus. Allophone-level annotations of the corpus have grapheme-to-phoneme (G2P) conversion ambiguities resolved by means of advanced G2P conversion tools. The third distinction of this study is that we compare word to sub-word mappings on the basis of a broad range of acoustic modelling techniques including state-of-the-art deep learning techniques. Finally, we dedicated lots of computational resources for the cross-validation experiments to verify the statistical significance of our findings.

The paper is organized as follows: Section 2 presents the background, describes the relationship between graphemes, phonemes and allophones of Lithuanian, and presents the prior work, Section 3 presents our methods, describes phonemic and graphemic mappings investigated in this paper, and presents the experimental setup, Section 4 presents the results, and finally the discussion and conclusions are presented in Section 5.

2. The Background

2.1. *Lithuanian Graphemes, Phonemes and Allophones*

Traditional Lithuanian spelling is based on the set of 32 graphemes: *a*, *ą*, *b*, *c*, *č*, *d*, *e*, *ę*, *ė*, *f*, *g*, *h*, *i*, *į*, *y*, *j*, *k*, *l*, *m*, *n*, *o*, *p*, *r*, *s*, *š*, *t*, *u*, *ū*, *ų*, *v*, *z*, *ž* that includes 9 diacritic symbols.[2] Lithuanian orthography is essentially phonological, i.e. standardized spelling reflects the essential phonological changes but also tolerates phonological inaccuracies. The definition of Lithuanian phoneme is subject to debate among linguists. Girdenis (2014) describes Lithuanian as having 58 phonemes (13 vowels and 45 consonants) whereas Pakerys (2003) talks about 49 phonemes (12 vowels and 37 consonants). This study is not concerned by different phoneme definitions, because it focuses on allophones and their sets. The following considerations summarize the essence of the relationship among graphemes, phonemes and allophones and illustrate the main difficulties of Lithuanian G2P conversion:

- Lithuanian consonants are either palatalized, or non-palatalized. Palatalization property of a consonant is not exposed by its grapheme symbol,[3] but can be inferred from its right context. One right standing grapheme is often enough, as consonants are always palatalized before graphemes *e*, *ę*, *ė*, *i*, *į*, *y*, *j*. However, in rare cases four right standing graphemes are required to infer this property correctly, e.g. *perskrido* [^{1}ˈpæːrʲsʲkrʲɪdoː] (flew over).
- Lithuanian vowels are either short (lax), or long (tense). Duration property of a vowel is not exposed by graphemes *a*, *e*, *o* (see Table 1).
- Grapheme pairs *ie*, *uo*, *ai*, *au*, *ei*, *ui* make up a diphtong (e.g. *paukštis* [^{2}ˈpɐu·kʃʲtʲɪs] (bird)) or hiatus (e.g. *paupys* [pɐ.ʊ^{2}ˈpʲiːs] (riverside)) if they are within the same syllable or span syllable boundaries respectively.
- Grapheme pairs *al, am, an, ar, el, em, en. er, il, im, in, ir, ul, um, un, ur* make up a mixed diphthong if they are within the same syllable.
- Syllable boundaries are not exposed by standard spelling.
- Lithuanian syllables are either stressed, or unstressed. Stress falls on a nucleus of the syllable, where nucleus may be a vowel, a diphthong or a mixed diphthong. Lithuanian phonetics distinguishes between two syllable accents: acute and circumflex. If a diphthong or a mixed diphthong is stressed, the acute and the circumflex make their respective first (vowel) and the second (vowel or consonant) components more prominent. Syllable accent is not exposed by standard spelling.
- Traditional Lithuanian spelling uses irregular affricate encoding. Affricates are encoded either by graphemes such as *c* ([t͡s, t͡sʲ]), *č* ([t͡ʃ, t͡ʃʲ]) or by digraphs: *dz* ([d͡z, d͡zʲ]), *dž* ([d͡ʒ, d͡ʒʲ]).

[2] Linguistic entity, like a grapheme or word written according to Lithuanian orthography is given in italics. International Phonetic Alphabet (IPA) based phonetic transcription is enclosed within square brackets. SAMPA-LT based allophonic transcription is given in plain text.

[3] In certain cases, palatalization is indicated by the grapheme *i* written after the palatalized consonant, e.g. *geriu* (drink), *gražios* (nice), i.e. palatalization is represented by a digraph.

Table 1
The relationship of Lithuanian graphemes and vowels. Graphemes *a*, *e*, *o* represent both short and long vowels.

Grapheme	*a*	*ą*	*e*	*ę*	*ė*	*i*	*į*, *y*	*o*	*u*	*ų*, *ū*
Phoneme	[ɐ], [ɑː]	[ɑː]	[ɛ], [æː]	[æː]	[eː]	[ɪ]	[iː]	[ɔ], [oː]	[ʊ]	[uː]

- Digraph *ch* encodes sounds [x] and [x^{j}].

The considerations above imply that G2P conversion of Lithuanian is quite complex. G2P converter that relies on a word spelling and grapheme rewrite rules (Greibus *et al.*, 2017; Lileikytė *et al.*, 2018), henceforth referred to as a shallow G2P converter, is incapable of resolving ambiguities related to vowel duration, syllable stress, and syllable boundaries and consequently is incapable of producing detailed and consistent allophone sequences. Only G2P converter making use of supplementary pronunciation dictionaries (Skripauskas and Telksnys, 2006) or of accentuation algorithms (Norkevičius *et al.*, 2005; Kazlauskienė *et al.*, 2010), henceforth referred to as a knowledge-rich G2P converter,[4] might be capable of disambiguating and modelling these phonological properties correctly.

2.2. *Related Work*

The problem of finding the best word to sub-word unit mapping for the applications of Lithuanian ASR was first addressed by Raškinis and Raškinienė (2003), followed by Šilingas (2005), Laurinčiukaitė and Lipeika (2007), Gales *et al.* (2015), Greibus *et al.* (2017), Lileikytė *et al.* (2018), and Ratkevicius *et al.* (2018).

All abovementioned studies have used very different ASR setups (see Table 2). First, different proprietary speech corpora were used for ASR system training and evaluation (Laurinčiukaitė *et al.*, 2006; Harper, 2016; Laurinčiukaitė *et al.*, 2018). Second, ASR setups were based on different acoustic modelling techniques, such as monophone HMM system (Šilingas, 2005; Ratkevicius *et al.*, 2018), triphone HMM system (Raškinis and Raškinienė, 2003; Šilingas, 2005; Laurinčiukaitė, 2008; Greibus *et al.*, 2017), or hybrid HMM – neural network models (Gales *et al.*, 2015; Lileikytė *et al.*, 2018). Third, different evaluation methodologies were used. Raškinis and Raškinienė (2003), Laurinčiukaitė and Lipeika (2007), Ratkevicius *et al.* (2018) and this study prefer accuracy estimation through cross-validation, whereas other studies estimate recognition accuracy on a held-out data, an approach that is less computation intensive. Fourth, different evaluation criteria were used. Studies differ by comparing PER (Šilingas, 2005), WER (Raškinis and Raškinienė, 2003; Šilingas, 2005; Laurinčiukaitė and Lipeika, 2007; Gales *et al.*, 2015; Lileikytė *et al.*, 2018; Ratkevicius *et al.*, 2018), and sentence error rate (Greibus *et al.*, 2017). Fifth, ASR setups incorporated different language models such as word loops (Raškinis and Raškinienė, 2003; Šilingas, 2005; Laurinčiukaitė and Lipeika, 2007; Ratkevicius *et al.*, 2018), word *n*-grams (Gales *et al.*, 2015; Lileikytė *et al.*, 2018), command lists (Greibus *et al.*, 2017), and phone *n*-grams (this study).

[4] Grapheme-to-allophone converter would be a more appropriate name.

Table 2
Comparison of experimental setups used to compare phonemic, graphemic and syllabic lexicons in various studies (WER – Word Error Rate; PER – Phone Error Rate; ATWV/MTWV – Actual/Maximum Term-Weighted Value[5]; SER – Sentence Error Rate).

Study	Corpus	Evaluation type	Comparison criteria	Language model	Acoustic modelling technique
Raškinis and Raškinienė, 2003	1 h of isolated words, 4 speakers	4-fold cross-validation, 15 min per round	WER	Word-loop	Triphone HMM
Šilingas, 2005	9 h of broadcast speech	Held out data, 14 min	WER, PER	Word-loop	Monophone HMM, Triphone HMM
Laurinčiukaitė and Lipeika 2007	23 speakers[6]	10-fold cross-validation, 1 h per round	WER	Word-loop	Triphone HMM
Gales et al., 2015	3–40 h of convers. telephone	Held out data, 10 hours	WER	Word n-gram	Triphone HMM, Hybrid HMM-DNN system
Lileikytė et al., 2018	speech		WER, ATWV/ MTWV	Word 3-gram	Triphone HMM, Hybrid HMM-DNN system
Greibus et al., 2017	46.5 h of read speech, 348 speakers	Held out data, 6.78 hours	SER	Command list	Triphone HMM
Ratkevičius et al., 2018	2.5 h of isolated words	5, 10-fold cross-validation	WER	Word-loop	Monophone HMM
This study	50 h of read speech, 50 speakers	10-fold cross-validation 1 hour per round	PER	Fully inter-connected triphones; phone 3-gram, 4-gram	Triphone HMM, LDA+MLLT Triphone HMM, SAT-HMM, SGMM, Hybrid HMM-TDNN, BLSTM (recurrent DNN)

Though word to grapheme mappings investigated by different studies are quite similar, word to phoneme mappings are different and mostly incompatible across studies. Each study makes its own choices about whether to and how to represent stress, duration, palatalization, affricates, diphthongs and mixed diphthongs in a phonemic lexicon (see Table 3). Laurinčiukaitė and Lipeika (2007) go beyond word to phoneme mappings and investigate word to sub-word unit mappings, where sub-words may be phonemes, syllables and pseudo-syllables.

[5] Actual/maximum term-weighted value is used to evaluate keyword spotting performance.

[6] Data of 10 speakers makes up 89% of the corpus. Every speaker is present in both training and test data.

Table 3
Comparison of phonemic lexicons that were investigated by various studies. Symbols in the table denote fine-grained (✚), partial (✓), and absent (**O**) modelling of some phonetic property.

Study	Phonemic lexicon as referenced by authors	Syllable stress[7] (vowels & diphth.)	Vowel[8] duration	Fronting of back[9] vowels	Syllable stress[10] (consonants)	Consonant[11] palatalization	Affricate[12] modelling	Diphthong[13] modelling	Mixed diphthong[14] modelling	Number of phonetic units
Raškinis et al. 2003	A	✚	✚	✚	✚	✚	✚	✚	**O**	115
	AB	✓	✚	✚	✚	✚	✚	✚	**O**	101
	ABC	**O**	✚	✚	**O**	✚	✚	✚	**O**	73
	ABD	✓	✚	✚	✚	**O**	✚	✚	**O**	76
	ABCD	**O**	✚	✚	**O**	**O**	✚	✚	**O**	50
Šilingas, 2005	BFR1	✚	✚	**O**	✚	✚	✚	✚	✚	229
	BFR2	✚	✚	**O**	✚	**O**	✚	✚	✚	140
	BFR3	**O**	✚	**O**	**O**	**O**	✚	✚	✚	86
	BFR4	**O**	✚	**O**	**O**	✚	✚	✚	✚	139
	BFR5	✚	✚	**O**	**O**	✚	**O**	✚	**O**	87
	BFR6	✚	✚	**O**	**O**	**O**	**O**	✚	**O**	71
	BFR7	**O**	✚	**O**	**O**	**O**	**O**	✚	**O**	41
3pt] Greibus et al., 2017	FZ1.3	**O**	✓	✚	**O**	**O**	✚	**O**	**O**	36
	FZ15.5	**O**	✓	**O**	**O**	✚	✚	**O**	**O**	61
	FPK1	✚	✚	✚	✚	✚	✚	✓	**O**	93
Lileikytė et al., 2016	FLP-32	**O**	✓	**O**	**O**	**O**	**O**	**O**	**O**	29
	FLP-36	**O**	✓	**O**	**O**	**O**	✚	**O**	**O**	33
	FLP-38	**O**	✓	**O**	**O**	**O**	**O**	✚	**O**	35
	FLP-48	**O**	✓	**O**	**O**	✚	**O**	**O**	**O**	45
This study	detailed	✚	✚	✚	✚	✚	✚	✚	✓	130
	no stress	**O**	✚	✚	**O**	✚	✚	✚	✓	79
	no palatalization	✚	✚	✚	✚	**O**	✚	✚	✓	98
	no mixed dipthongs	✚	✚	✚	✚	✚	✚	✚	**O**	122
	no diphthongs	✚	✚	✚	✚	✚	✚	✓	✓	112
	no affricates	✚	✚	✚	✚	✚	**O**	✚	✓	122

[7] Lexicon includes allophones to represent differently stressed variants of all vowels and diphthongs (✚), or only diphthongs *ai, au, ei, ui* (✓). Lexicon ignores the opposition of stressed vs. non-stressed sounds (**O**).

[8] Lexicon includes allophones to represent the opposition of short vs. long vowels and phone symbols in the actual transcription reflect this opposition consistently (✚), or to the extent that is possible with a shallow G2P converter (✓).

[9] Lexicon represents (✚) or ignores (**O**) the opposition of fronted vs. regular back vowels (e.g. [ɔ̣], [ʊ̣] vs. [ɔ], [ʊ]).

[10] Lexicon represents (✚) or ignores (**O**) the opposition of stressed vs. non-stressed consonants.

[11] Lexicon represents (✚) or ignores (**O**) the opposition of palatalized and non-palatalized consonants.

[12] Lexicon represents affricates by a single (✚) or two (**O**) consonants.

[13] Lexicon includes allophones to represent all diphthongs (✚), or only diphthongs *ie, uo* (✓) by a single phone. Lexicon encodes all diphthongs by a sequence of two phones (**O**).

[14] Lexicon represents mixed diphthongs by different dedicated allophones (✚). Lexicon models mixed diphthongs by the sequence of two constituent phones (**O**) but it also models sonorants which make part of a mixed diphthong as distinct allophones (✓).

Given such a variety of the experimental setups it is not surprising that different studies came to different and even opposite conclusions. For instance, Raškinis and Raškinienė (2003) achieved the best WER by the word to phoneme mapping that ignored stress and preserved palatalization (see Table 3, ABC phonemic lexicon), whereas (Šilingas, 2005) achieved best WER by preserving stress and ignoring palatalization (see Table 3, BFR6 phonemic lexicon). Greibus *et al.* (2017) achieved best SER by ignoring both stress and palatalization. Gales *et al.* (2015) found that grapheme-based system outperforms phoneme-based system, whereas Šilingas (2005), Greibus *et al.* (2017) and Lileikytė *et al.* (2018) came to an opposite result. Laurinčiukaitė and Lipeika (2007) found that mapping into a mixture of phonemes and syllable-like units improves WER.

Incompatible conclusions are partially due to the limitations of the experimental setups. Some findings are based on a small training corpus (Raškinis and Raškinienė, 2003; Ratkevicius *et al.*, 2018) or on a limited carefully selected held-out data (Šilingas, 2005). Other studies (Greibus *et al.*, 2017; Lileikytė *et al.*, 2018) are testing limited word-to-phoneme mappings due to the usage of a shallow G2P converter which is unable to produce allophone-rich phonemic transriptions. Conclusions of many studies are dependent on a single (though generally state-of-the-art at the time of investigation) acoustic modelling technique. Finally, recognition accuracies obtained by the majority of studies are not “pure” indicators of performance of different word to sub-word mappings as they are strongly influenced by different amounts of linguistic constraints embedded into ASR setups. For instance, Greibus *et al.* (2017) restrict their language model (LM) to a command list, where commands share 271 unique word types, and Ratkevicius *et al.* (2018) restrict their LM to a 10-digit word loop.

3. The Method

3.1. *Investigated Phonemic and Graphemic Lexicons*

In this study, we have adopted an experimental approach common to other similar studies (Raškinis and Raškinienė, 2003; Šilingas, 2005). It consists of defining some phonemic lexicon which serves as a reference point. Thereafter, reductions of this lexicon are derived by elimination of various phonological properties (e.g. stress, palatalization) or by splitting compound phonetic units (e.g. diphthongs, affricates) into sub-parts and measuring the performance of the ASR system for every reduced lexicon. Our reference phonemic lexicon consists of 130 allophones (henceforth referred as to “detailed” lexicon). It is presented in Table 4 using SAMPA-LT (Raškinis *et al.*, 2003) encoding.

We have compared the “detailed” lexicon against 5 reduced phonemic lexicons and one graphemic lexicon in order to answer the questions about what is the best approach to:

- Stress modelling (present vs. absent),

[15] N encodes velarized *n* ([ŋ]).

Table 4
Detailed list of Lithuanian allophones in SAMPA-LT encoding. Acute and circumflex are encoded by double quote (”) and caret (∧) respectively. Column (:) distinguishes long vowels from short ones. Palatalization is encoded by a single quote (’). Sonorants that make part of a mixed diphthong are labelled by period (.).

(a) Vowels and diphthongs

	Short		Long		
	Unstressed	Stressed	Unstressed	Stressed (acute)	Stressed (circumflex)
Vowels	a, e, i, o, u	"a, "e, "i, "o, "u	a:, e:, E:, i:, o:, u:	"a:, "e:, "E:, "i:, "o:, "u:	^a:, ^e:, ^E:, ^i:, ^o:, ^u:
Fronted vowels	io, iu	"io, "iu	io:, iu:	"io:, "iu:	^io:, ^iu:
Diphthongs			ie, uo, iuo ai, au, ei, eu, ui, iui	"ie, "uo, "iuo "ai, "au, "ei, "eu, "ui, "iui	^ie, ^uo, ^iuo ^ai, ^au, ^ei, ^eu, ^ui, ^iui

(b) Plosives, fricatives and affricates

	Non-palatalized		Palatalized	
	Voiced	Unvoiced	Voiced	Unvoiced
Plosives	b, d, g	p, t, k	b’, d’, g’	p’, t’, k’
Fricatives	z, Z, G, v, j	s, S, x, f	z’, Z’, G’, v’	s’, S’, x’, f’
Affricates	dz, dZ	ts, tS	dz’, dZ’	ts’, tS’

(c) Sonorants

	Non-palatalized			Palatalized		
	Standalone	Part of a mixed diphthong		Standalone	Part of a mixed diphthong	
		Unstressed	Stressed		Unstressed	Stressed
Sonorants	l, m, n, r	l., m., n., N.[15], r.	^l., ^m., ^n., ^N., ^r.	l’, m’, n’, r’	l.’, m.’, n.’, N.’, r.’	^l.’, ^m.’, ^n.’, ^N.’, ^r.’

- Palatalization modelling (present vs. absent),
- Diphthong modelling (one vs. two phones),
- Mixed diphthong modelling (distinguishing vs. not distinguishing constituent consonants),
- Affricate modelling (one vs. two phones),

and whether a phone-based ASR system outperforms a grapheme-based one.[16] Answers to those questions are important from the practical perspective. As mentioned previously, extracting stress and syllabification data from word spelling is costly in terms of human expertise.

[16] It may seem that a larger set of phonemes will always model pronunciation better with a sufficient corpus size. This may be true in case of monophone-based single-GMM (i.e. the simplest) acoustic models where model complexity directly depends on the number of phoneme symbols. Triphone acoustic models based on reduced lexicons may have more triphones than acoustic models based on detailed lexicons. Complexity of a triphone acoustic model, which can be expressed as a number of different acoustic states or a number of probability density functions (pdfs), isn’t directly related to the size of the symbol set. Pdf clustering procedure (to alleviate the data scarcity problem) usually makes all triphone models of approximately the same size/complexity for a given fixed corpus size.

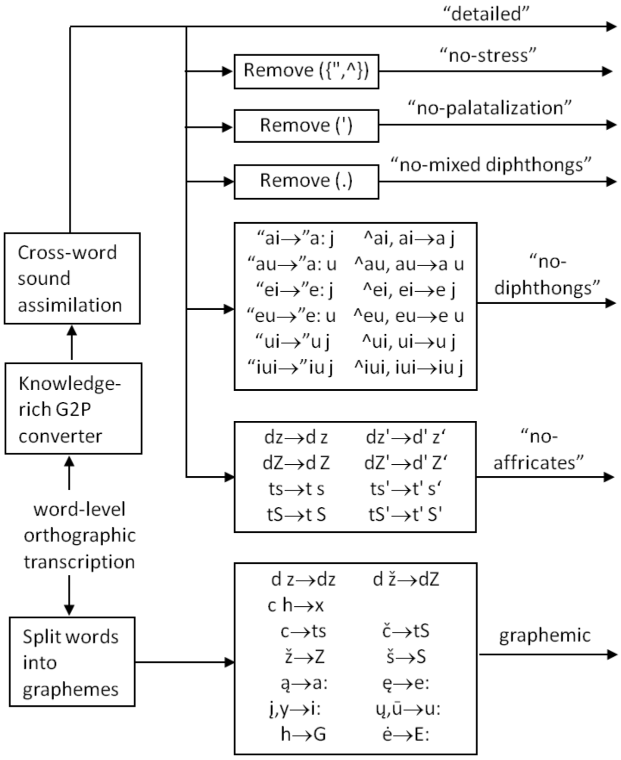


Fig. 1. The process of deriving phonemic and graphemic transcriptions from the input ortographic transcription. Arrow symbol denotes a broad "replace" operator (inclusive of "split" and "merge").

The process by which different phonemic and graphemic transcriptions were obtained is described in Fig. 1. First, word-level ortographic transcription was processed by the knowledge-rich G2P converter (Kazlauskienė *et al.*, 2010) resulting in allophone sequence that observes intra-word sound assimilation rules. Thereafter optional sound assimilation rules were applied at word boundaries in an automatic way on a basis of a maximum-likelihood criterion. This resulted in "detailed" phone-level transcription encoded by SAMPA-LT symbols that served as our reference word-to-phoneme mapping. Reduced phonemic transcriptions were derived from the "detailed" transcription by subjecting it to one or more editing operations (see Fig. 1).

Graphemic transcription was obtained from the word-level ortographic transcription by means of a few editing operations that encoded graphemes by SAMPA-LT symbols. This encoding was necessary in order to harmonize phonemic and graphemic transcriptions for their comparison at a later stage (see 3.6). Graphemic lexicon may look like a phonemic one, but this is a false impression. Graphemic transcription was not subjected to sound assimilation rules, and the changes in graphemic transcriptions are simple and mostly reversible transliterations.

3.2. *Speech Data and the ASR Cross-Fold Validation Setup*

Our experiments were based on a 50-hour speech corpus that was compiled at Vytautas Magnus University. The corpus consisted of 50 speakers (25 males and 25 females) each reading book excerpts for approximately 1 hour.[17] Word-level transcriptions of this corpus were manually adjusted to match misspellings and missacentuations present in audio recordings.

We built multiple ASR systems based on different phonemic/graphemic lexicons and tried to estimate their accuracies via a cross-validation technique. The cross-validation round consisted of training an ASR system (building acoustic and phone-level language models) on the speech and transcripts of 49 speakers and testing system accuracy on the speech of the held-out (or test) speaker. Full leave-one-out (or 50-fold) cross-validation was costly in terms of computational time. Instead, we approximated it with a "pessimistic" 10-fold cross-validation scheme. We call it "pessimistic" (with respect to the leave-one-out cross-validation) because of the inclusion of the most problematic speakers into the test set. The selection procedure clustered all speakers into 5 clusters of comparable size and selected 2 worst rated speakers per cluster for inclusion into the test set (2x5=10 in total).[18] Identifiers of selected speakers and their ratings are given in Table 6.

3.3. *Acoustic Models*

It is reasonable to expect that a certain phonemic lexicon performs better when coupled with some particular acoustic modelling technique. In order to investigate this relationship and to assess the possible limitations of our conclusions we have built and compared the ASR systems based on the acoustic models of 7 different types[19] including:

1. Monophone HMM system (henceforth referred to as "mono" system) was the simplest ASR system, where each phone was modelled by a single HMM. HMMs had from 2 to 5 states (number of states was related to the average phone duration) and shared left-to-right topology. The number of Gaussian probability density functions (pdfs) per state was estimated as an exponential function of the state occupation counts targeting 1000 pdfs in total. Speech data was parametrized by extracting 13 mel-frequency cepstral coefficients (MFCC) and their first-order and second-order derivatives from 25 ms speech frames at 10 ms intervals. Per speaker cepstral mean normalization was applied.
2. Triphone HMM system (henceforth referred to as "tri-mfcc" system) was trained on the same features as "mono" system, but each phone was modelled by multiple

[17] Silent segments make 15–20% of the corpus depending on the speaker.

[18] Speaker rating was determined on the basis of WER obtained in our earlier recognition and adaptation experiments (Rudžionis *et al.*, 2013). Speaker clustering was based on the feature set that included insertion, deletion and substitution errors.

[19] We have used an open-source Kaldi ASR toolkit (Povey *et al.*, 2011a) for training and evaluating all ASR systems. Some other techniques, mostly discriminative training approaches, have been tried but not described in this paper, because their accuracy estimates correlated with the results of non-discriminative training.

context-dependent HMMs (triphones). The system targeted 11000 Gaussian pdfs in total. Triphone state tying was performed using decision-tree clustering technique and resulted in approximately 2000 clusters (tree leaves).

3. Triphone HMM system (henceforth referred to as "tri-lda" system) was trained in the same way as "tri-mfcc" system, but was based on a different speech parametrization. It consisted of splicing 13-dimensional MFCC vectors across 7 frames (3 frames on each side of the current frame) resulting in 91-dimensional feature vectors, applying Linear Discriminant Analysis (LDA) to reduce vector dimensionality to 40, and finally estimating the Maximum Likelihood Linear Transform (MLLT) (Gales, 1999) over multiple iterations. MLLT represents a square feature transformation matrix with the objective function being the average per-frame log-likelihood of the transformed features given the model.
4. Speaker-adaptively trained (SAT) triphone HMM system (henceforth referred to as "tri-sat" system) differed from the previous one as speaker-specific feature-space maximum likelihood linear regression (fMLLR) adaptation was added on the top of LDA+MLLT speech parametrization. fMLLR is an affine feature transformation whose estimation techniques are detailed in Gales (1998).
5. System based on the Subspace Gaussian Mixture Models (SGMM) is a general HMM model where states share the same GMM structure (henceforth referred to as "sgmm" system). The acoustic model is defined by vectors associated with each state and by a global mapping from this vector space to the space of parameters of the GMM. Thus GMM means and mixture weights are constrained to vary in a subspace of the full parameter space (Povey *et al.*, 2011b). This system was trained on the top of fMLLR adapted speech features.
6. System based on a feed-forward deep neural network known as Time-Delay Neural Network (TDNN) henceforth referred to as "tdnn" system. This system was trained using procedure described in Zhang *et al.* (2014). First, "tri-sat" system was asked to produce frame-level state labelling for the training speech. Thereafter, state labels were used as targets to train the TDNN acoustic models. Speech data was parametrized by extracting 40 mel-frequency filterbank coefficients, splicing 40-dimensional vectors across 9 frames resulting in 360-dimensional feature vectors. Thus, TDNN had 360 inputs and aproximately 1750 outputs[20] corresponding to the context-dependent phone state labels. In between the input and the output layers TDNN had two hidden layers based on tangent non-linearity. TDNN was trained for 20 epochs by reducing learning rate during the first 15 epochs.
7. System based on a recurrent deep neural network known as Low Frame Rate Bidirectional Long Short Term Memory (LFR BLSTM). This system was trained using procedure described in Povey *et al.* (2016). Two additional speed perturbed copies of training data were used for 3-fold data augmentation (Ko *et al.*, 2015). 100-dimensional iVectors were extracted in online manner and were used as additional inputs to the BLSTM network to perform instantaneous adaptation of the

[20] The exact number is dependent on the held-out speaker identity in a particular cross-validation round.

Table 5
Average perplexities of the phone-level n-grams, measured on the held-out parts of the speech corpus.

Lexicon	Categorial 3-gram	1-gram	2-gram	3-gram	4-gram
Detailed	31.45	60.66	18.05	12.23	9.46
No stress	27.00	45.37	14.57	10.97	8.81
No palatalization	33.17	44.24	17.96	12.32	9.44
No mix. diphthongs	32.05	56.35	18.59	12.36	9.53
No diphthongs	30.54	52.91	16.76	11.50	8.94
No affricates	30.89	59.12	17.83	12.11	9.35
Graphemes	23.55	22.14	12.79	9.67	7.86

neural network (Soan *et al.*, 2013). LFR BLSTM architecture had 3 forward and 3 backward layers. The model was trained for 4 epochs by linearly reducing learning rate throughout the training process. This ASR system is referred to as "blstm" in the subsequent sections.

3.4. *Phone-Level Language Models*

We aim to build an experimental setup such that the ASR system is stripped from its lexical and grammatical knowledge (list of words of a language and probabilities associated to word sequences) that influences recognition accuracy, so that the accuracy of the ASR system reflects the performance of the word to sub-word unit mappings under investigation. It should be noted that phonotactic knowledge cannot be eliminated from our comparisons because it makes an integral part of an acoustic (starting from triphones) model. If we take a triphone acoustic model, extract the list of all triphones, and make a fully-connected triphone network that respects adjacency constraints (i.e. triphone a-x+y is connected to every triphone x-y+b in the list, where a, b, x, y denote any sub-word unit of the lexicon) we obtain a phone 3-gram with the uniform probability distribution over the outgoing links. It represents the set of categorial phonotactic constraints embedded into an acoustic model. Let's call it the categorial phone 3-gram. Table 5 compares perplexities of the categorial and probabilistic phone-level n-grams.

We have taken the categorial phone 3-gram as our baseline decoding setup. In addition, we performed decoding experiments with phone 3-grams and 4-grams to observe how additional probabilistic phonotactic knowledge affected the ASR performance.[21]

Decoding with categorial 3-grams, probabilistic 3-grams and 4-grams exploited phonotactic but not lexical or syntactic-semantic knowledge, so we believe that our comparisons were independent from the lexical content of the training/evaluation data.

To summarize, our experimental investigation consisted of building 7 (phonemic/graphemic lexicons) × 7 (acoustic model types) × 10 (speaker-specific cross-

[21] We did not perform decoding with phone 1-grams and 2-grams because their decoding accuracies are hard to interpret. On the one hand, phone 1-grams and 2-grams are under-constrained with respect to the categorial phonotactic constraints integral to the triphone acoustic model, and, consequently, the decoder is forced to synthesize triphones that violate phonotactic constraints of the language. On the other hand, 1-grams and 2-grams are more constrained by probabilistic knowledge than the categorial 3-gram by taking advantage of statistics of the training corpus.

validation rounds) = 490 different acoustic models and performing 490 (acoustic models) × 3 (phone-level language models) = 1470 decoding experiments in total.

3.5. *Scoring: Accuracy Estimation*

We have used Phone Error Rate (PER) criterion to compare the performances of different ASR setups. It was calculated according to:

$$PER = \frac{S+I+D}{N}100\%, \tag{1}$$

where S, I and D denote substitution, insertion and deletion errors respectively, and N is the total number of phones/graphemes in the test data. S, I and D estimates were extracted from automatic alignments of recognized and reference transcriptions.

3.6. *Scoring: Transcription Normalization*

Automatic alignment of recognized and reference transcriptions was preceded by the transcription normalization step. This step consisted of projecting every individual phoneme/grapheme onto a symbol or a sequence of symbols over the normalized alphabet. Without projecting lexicons of different sizes into the common lexicon the comparison would be biased against allophone-rich ASR setups as they naturally tend to result in more substitution errors than ASR setups based on reduced lexicons. Moreover, without normalization, substitution errors involving, e.g. stressed vs. unstressed or palatalized vs. non-palatalized allophones of the same phoneme, will look like equally important as phoneme substitutions.

Normalized alphabet contained 27 symbols (a b d e E: f g G x i i: j k l m n o p r s S t u u: v z Z). It represented the intersection of all investigated lexicons, i.e. it contained symbols that were common to all lexicons. Other allophone units were projected into this alphabet by eliminating their phonetic properties or by spliting compound units (affricates, diphthongs, fronted back vowels) into the sequences of 2 or 3 symbols.[22] As the only exception to this rule, we have eliminated symbols a: and e: from the normalized alphabet even if these symbols were present in all investigated lexicons. The effect of this exception was that a / a: and e / e: substitutions were no longer interpreted as errors. This was done to eliminate the bias against the graphemic lexicon so that it was not penalized for the failures to resolve duration ambiguities it was hardly able to resolve.[23]

The process of projecting phonemic and graphemic transcripts into scoring transcripts over the normalized alphabet was realized by 4 steps:

1. Remove double quote("), caret(^), single quote('), period(.) from SAMPA-LT phone descriptions;

[22] For instance, graphemic lexicon lacks the symbol o: (see Table 1 and bottom part of Fig. 1), so o: is not included into the normalized alphabet, and all phonemic lexicons are projecting o: → o. The "no affricates" lexicon lacks affricates ts, tS, dz, dZ, so other lexicons are projecting affricates into a sequence of two symbols.

[23] Acoustic models of graphemes *a* and *ą* are both trained on acoustic samples of [a:] (see Table 1).

2. Split multi-symbol SAMPA-LT phone descriptions (ai, au, ei, eu, ie, ui, uo, iui, iuo, dz, dZ, ts, tS) into forming symbols;
3. Split: iu → i u, iu: → i u:, io → i o, io: → i o:;
4. Replace: e: → e, a: → a, o: → o, N → n.

Though grapheme-based and phoneme-based reference transcriptions are mapped to transcriptions over the same normalized alphabet, they are not identical. For instance, Lithuanian word *džiaugsis* (will rejoice) is mapped to d Z e u k s i s (phonemic) and d Z i a u g s i s (graphemic) over the same normalized alphabet. The difference stems from the fact that phonemic transcriptions by definition are transcriptions subjected to sound assimilation rules.[24]

4. Experimental Results

Let $\mathrm{PER}_{\mathrm{LX,AM,LM,SPK}}$ denote a Phone Error Rate that is obtained by the ASR setup based on the lexicon LX, the acoustic modelling technique AM, the phone-level LM and corresponds to the cross validation round, in which the data of SPK speaker is decoded.[25] Values of $\mathrm{PER}_{\mathrm{detailed,*,categorial\ 3\text{-}gram,*}}$ and $\mathrm{PER}_{\mathrm{graphemic,*,categorial\ 3\text{-}gram,*}}$ are shown in Tables 6a and 6b respectively for illustration purposes. Each table corresponds to the PER values obtained by 70 different ASR setups (10 speaker specific cross validation rounds × 7 acoustic modelling techniques).

To compare different word to sub-word unit mappings, we are mainly interested not in the PER values themselves but in the differences between $\mathrm{PER}_{\mathrm{LX,AM,LM,SPK}}$ values for different choices of LX everything else being fixed (e.g. differences between corresponding cells of Tables 6a and 6b).

Let's define a discrete random variable:

$$X_{\mathrm{LX1,\,LX2,\,AM,\,LM}} = \left\{ \frac{PER_{\mathrm{LX2,\,AM,\,LM,\,i}} - PER_{\mathrm{LX1,\,AM,\,LM,\,i}}}{PER_{\mathrm{LX1,\,AM,\,LM,i}}} \right\}, \tag{2}$$

where index i ranges over speaker identities. This random variable represents a relative increase (if it is positive) or relative decrease (if it is negative) of PER as a consequence of replacement of the lexicon LX1 with a lexicon LX2 in the ASR setup that has acous-

[24] The original word spelling cannot be restituted neither from a phonemic, nor from a graphemic transcription expressed over the normalized alphabet due to the one-to-many mapping, e.g. a t S i u: could be restituted as *ačiū* (thanks), *ačių*, *atšiū*, *atšių*, *ąčiū*, *ąčių*, *ątšiū*, *ątšių* (nonsense words).

[25] LX ∈ {detailed, no-stress, no-palatalization, no-mixed diphthongs, no-diphthongs, no-affricates, graphemic}, AM ∈ {mono, tri-mfcc, tri-lda, tri-sat, sgmm, tdnn, blstm}, LM ∈ {categorial phone 3-gram, probabilistic phone 3-gram, probabilistic phone 4-gram}, SPK ∈ {ARM, BLA, CIZ, DEK, EID, JUK, LEO, MAL, RUP, SKA}.

[25] The best speaker has the rating of 1 and the worst speaker has the rating of 50.

Table 6
Phone error rates obtained by different ASR setups when decoding with categorial phone 3-gram. Columns represent different acoustic modelling techniques. Rows represent 10 speaker-specific cross validation rounds.

(a) ASR setups based on "detailed" phonemic lexicon ($\mathrm{PER}_{\text{detailed},*,\text{categorial 3-gram},*}$)

Speaker (rating[26])	mono	tri_mfcc	tri_lda	tri_sat	sgmm	tdnn	blstm
ARM (50)	53.89	48.90	45.61	33.36	31.46	28.82	32.14
BLA (39)	42.27	28.98	25.46	21.37	17.29	14.62	11.46
CIZ (37)	44.60	30.87	29.30	21.94	17.11	15.67	12.46
EID (43)	46.27	34.93	31.53	26.39	21.78	19.27	15.39
DEK (1)	30.59	17.02	14.61	12.38	9.84	9.06	6.89
JUK (46)	41.78	32.76	29.35	25.14	21.68	19.11	17.54
LEO (34)	40.27	27.38	22.97	18.31	14.72	12.85	6.50
MAL (49)	47.90	37.95	32.46	27.27	23.55	22.24	17.38
RUP (36)	35.39	24.30	21.59	17.46	13.45	11.65	8.78
SKA (47)	46.70	37.55	35.74	31.27	26.28	24.59	19.59
Average	**42.97**	**32.06**	**28.86**	**23.49**	**19.72**	**17.79**	**14.82**

(b) ASR setups based on graphemic lexicon ($\mathrm{PER}_{\text{graphemic},*,\text{categorial 3-gram},*}$)

Speaker (rating)	mono	tri_mfcc	tri_lda	tri_sat	sgmm	tdnn	blstm
ARM (50)	54.30	52.81	48.02	37.57	35.95	31.32	32.19
BLA (39)	42.37	33.29	29.80	25.57	21.36	17.14	12.94
CIZ (37)	46.98	34.82	32.93	26.05	21.47	18.60	12.03
EID (43)	48.40	38.47	34.96	30.16	25.79	22.41	16.56
DEK (1)	33.48	20.22	17.44	15.46	12.94	10.95	7.50
JUK (46)	45.08	35.70	32.72	28.66	25.72	21.38	18.46
LEO (34)	41.20	31.24	27.78	22.02	18.80	16.06	7.60
MAL (49)	48.35	40.08	35.44	30.59	26.94	24.23	18.48
RUP (36)	37.17	28.27	25.28	21.26	17.66	14.49	9.86
SKA (47)	49.89	40.55	38.55	34.64	29.94	26.91	20.28
Average	**44.72**	**35.54**	**32.29**	**27.20**	**23.66**	**20.35**	**15.59**

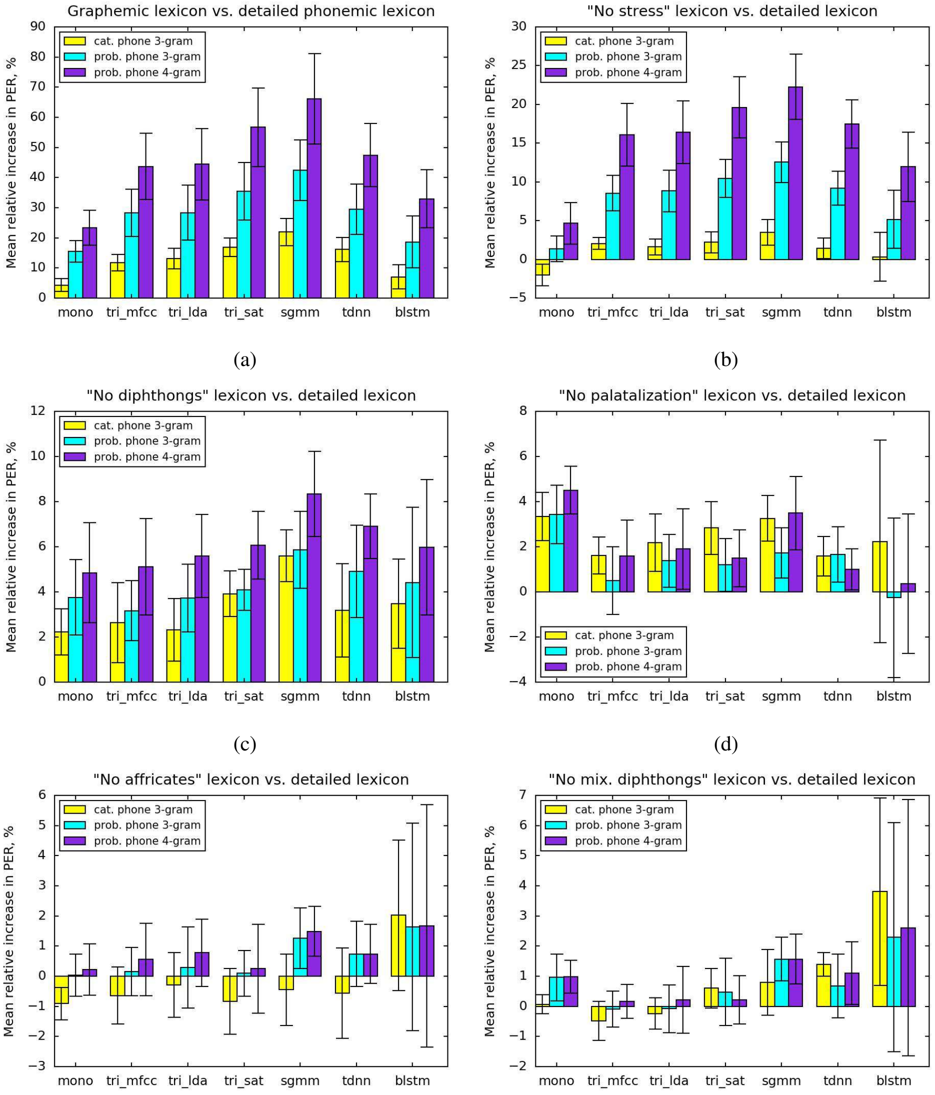
tic modelling technique AM and the phone-level language model LM fixed. Confidence intervals for this random variable could be computed by:

$$\overline{X}_{\mathrm{LX1,\,LX2,\,AM,\,LM}} \pm t \times \frac{S_{\mathrm{LX1,\,LX2,\,AM,\,LM}}}{\sqrt{n}}, \tag{3}$$

where $\overline{X}_{\mathrm{LX1,LX2,AM,LM}}$ and $\mathrm{S}_{\mathrm{LX1,LX2,AM,LM}}$ are the mean and standard deviation of the random variable $\mathrm{X}_{\mathrm{LX1,LX2,AM,LM}}$, $n = 10$ is the sample size and $t = 2.262$ is t-value for the 95% confidence level with $n - 1$ degrees of freedom. Fig. 2 compares detailed phonemic lexicon with reduced phonemic lexicons and a graphemic lexicon. It shows means and confidence intervals for random variables $\mathrm{X}_{\mathrm{detailed,LX,AM,LM}}$ given different LX, AM, and LM values.

Plots of Fig. 2 reveal the following tendencies:

- Detailed phonemic lexicon significantly outperforms graphemic lexicon across all investigated acoustic modelling techniques and all phone-level language models (Fig. 2(a)).



(a)

(b)

(c)

(d)

(e)

(f)

Fig. 2. Mean relative increase in phone error rate and 95% confidence intervals after substituting detailed phonemic transcription with a) graphemic transcription; b) "no-stress", c) "no diphthongs", d) "no palatalization", e) "no affricates", f) "no mixed diphthongs" phonemic transcriptions.

- Detailed phonemic lexicon that models diphthongs as a single unit significantly outperforms reduced phonemic lexicon that models diphthongs as a sequence of two units across all investigated acoustic modelling techniques and all phone-level language models (Fig. 2(c)).
- Detailed phonemic lexicon that preserves distinction of stressed vs. non-stressed vowels and the distinction of palatalized vs. non-palatalized consonants is perform-

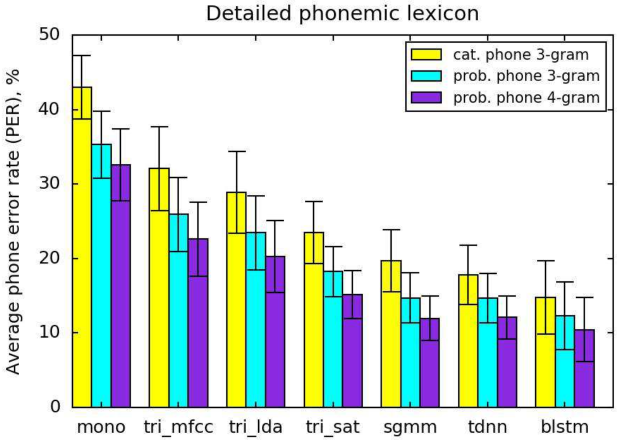


Fig. 3. Means and 95% confidence intervals of the phone error rate (PER).

ing significantly better with respect to the lexicons that ignore stress (Fig. 2(b)) or palatalization (Fig. 2(d)). PER obtained with the LFR-BLSTM acoustic model is the only, albeit statistically not significant, exception to this tendency. We hypothesize that bidirectional recurrent neural network is capable of capturing enough future context to model palatalization with a comparable accuracy to the lexicon that has distinct labels for palatalized and non-palatalized consonants.

- "No stress", "no diphthongs" and graphemic lexicons (Fig. 2(b), 2(c), 2(a)) become even less attractive if decoder is provided with more phonotactic knowledge.
- "No affricates" phonemic lexicon (modelling affricates by two sub-word units) slightly outperforms detailed lexicon if ASR setup consists of GMM-based or TDNN acoustic models and decoding is done with a categorial phone 3-gram (Fig. 2(e)). Giving more phonotactical knowledge to the decoder (probabilistic phone 3-gram or 4-gram) seems to reverse this tendency. BLSTM acoustic model is also in favour of detailed lexicon. However, all observed differences between detailed and "no affricates" lexicon are not statistically significant.
- It seems that distinguishing sonorants that make part of a mixed diphthong from the regular ones may be slightly preferred (Fig. 2(f)). Though such preference is not proven to be statistically significant.
- The LFR BLSTM acoustic model shows higher variability (Fig. 2(d)–2(f)) of the relative increase in PER in comparison to other acoustic models. Higher variability is due to the randomness of the BLSTM training procedure[27] and usually lower denominator values (variable $X_{\mathrm{LX1,\,LX2,\,AM,\,LM}}$ in expression (2)).

Absolute PER values for different acoustic modelling techniques are shown in Fig. 3.

[27] We have observed that PER on the test subset may differ by as much as 0.5–1.0% for two random initializations (training subset, validation subset, initial weights).

Though many different acoustic modelling techniques have been tried in this study, we do not make claims about their relative performance,[28] because we would need to prove that the optimum configuration was chosen for every acoustic modelling technique. Such an investigation was out of the scope of this paper. However, it seems that ASR setups based on the recursive deep neural network acoustic models compare well to the other acoustic modelling techniques. This result is in-line with the general tendency in ASR domain and represent the direction to go forward.

5. Discussion and Conclusions

This paper reviewed 15 years of research on the problem of the optimum word to sub-word unit mapping for the purposes of the Lithuanian ASR. It presented a common framework to compare different phonemic word to sub-word mappings. It also investigated and compared multiple phonemic and graphemic word to sub-word mappings across a broad range of acoustic modelling techniques.

Our investigation has shown that phonemic mappings outperform graphemic mappings by a large margin. We assume that other studies, that have found graphemic mappings better (Gales *et al.*, 2015) or comparable (Lileikytė *et al.*, 2018) in performance to phonemic ones, came up to this result by contrasting graphemic mappings to phonemic mappings lacking important features. For instance, phonemic lexicons investigated by Lileikytė *et al.* (2018) lack stressed allophones, whereas the importance of distinguishing stressed and non-stressed allophones is demonstrated in this study.

Though our investigation has not revealed which phonemic mapping is the best one, it gave insights about which mappings should not be used. Phonemic mappings that model diphthongs by two symbols and/or ignore stress were statistically significantly outperformed by the most detailed lexicon. What is the best approach to model palatalization, mixed diphthongs and affricates is still subject to the future investigations.

Our findings were obtained in the framework of separately tuning an acoustic model of the ASR system. Categorial phone 3-gram and PER criterion have helped us to eliminate lexical and syntactic-semantic layers of the ASR system and to evaluate word to sub-word unit mappings on the basis of the performance of an acoustic model alone. It is worth addressing the question of the best word to sub-word unit mapping in the framework of jointly tuning the complete ASR system (acoustic and word-level language models together) and checking if the gains in PER observed with an isolated acoustic model translate into the WER gains of the jointly optimized system.

Detailed lexicon was among the best performing lexicons investigated in this study. Thus, we believe that data scarcity played no major role in our investigations and our findings might be valid for corpora that are larger than 50 hours. It might be worth investigating even more detailed word to sub-word unit mappings including syllables, sylable-like units, consonant clusters, etc. following the suggestion of Laurinčiukaitė (2008).

[28] Decoding results obtained on the basis of LFR BLSTM acoustic models can not be directly compared to other decoding results because this ASR setup was trained on 3 copies of speed-perturbed data.

Acknowledgements. Part of this research has been supported by a grant from the Research Council of Lithuania under the National Lithuanian studies development programme for 2009–2015 through the project "A unified approach to Lithuanian prosody: the intonation, rhythm, and stress" (reg. no. LIT-5-4).

References

Alumäe, T., Tilk, O. (2016). Automatic speech recognition system for Lithuanian broadcast audio. In: *Human Language Technologies – The Baltic Perspective: Proceedings of the Seventh International Conference, Baltic HLT 2016*, Vol. 289, pp. 39–45.

Collobert, R., Puhrsch, C., Synnaeve, G. (2016). Wav2Letter: an end-to-end ConvNet-based speech recognition system. arXiv:1609.03193 [cs.LG].

Gales, M.J.F. (1998). Maximum likelihood linear transformations for HMM-based speech recognition. *Computer Speech and Language*, 12(2), 75–98.

Gales, M.J.F. (1999). Semi-tied covariance matrices for hidden Markov models. *IEEE Transactions on Speech and Audio Processing*, 7, 272–281.

Gales, M.J.F., Knill, K.M., Ragni, A. (2015). Unicode-based graphemic systems for limited resource languages. In: *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5186–5190.

Girdenis, A. (2014). *Theoretical Foundations of Lithuanian Phonology*. English translation by Steven Young., XVII, 413.

Greibus, M., Ringelienė, Ž., Telksnys, A.L. (2017). The phoneme set influence for Lithuanian speech commands recognition accuracy. In: *Proceedings of the Conference Electrical, Electronic and Information Sciences (eStream)*, pp. 1–4.

Harper, M. (2016). *Babel: US IARPA Project (2012–2016)*. https://www.iarpa.gov/index.php/research-programs/babel.

Kanthak, S., Ney, H. (2002). Context-dependent acoustic modeling using graphemes for large vocabulary speech recognition. In: *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vol. 2, pp. 845–848.

Kazlauskienė, A., Raškinis, G., Vaičiūnas, A. (2010). *Automatic Syllabification, Stress Assignment and Phonetic Transcription of Lithuanian Words* (in Lithuanian).

Killer, M., Stüker, S., Schultz, T. (2003). Grapheme based speech recognition. In: *Proceedings of Interspeech-2003*, pp. 3141–3144.

Ko, T., Peddinti, V., Povey, D., Khudanpur, S. (2015). Audio augmentation for speech recognition. In: *Proceedings of Interspeech-2015*, pp. 3586–3589.

Laurinčiukaitė, S., Šilingas, D., Skripkauskas, M., Telksnys, L. (2006). Lithuanian continuous speech corpus LRN 0.1: design and potential applications. *Information Technology and Control*, 35(4), 431–440.

Laurinčiukaitė, S., Lipeika, A. (2007). Framework for choosing a set of syllables and phonemes for Lithuanian speech recognition. *Informatica*, 18(3), 395–406.

Laurinčiukaitė, S. (2008). *Acoustic Modeling of Lithuanian Speech Recogniton*. PhD Thesis (in Lithuanian).

Laurinčiukaitė, S., Telksnys, L., Kasparaitis, P., Kliukienė, R., Paukštytė, V. (2018). Lithuanian speech corpus Liepa for development of human-computer interfaces working in voice recognition and synthesis mode. *Informatica*, 29(3), 487–498.

Lileikytė, R., Gorin, A., Lamel, L., Gauvain, J., Fraga-Silva, T. (2016). Lithuanian broadcast speech transcription using semi-supervised acoustic model training. *Proceedings of Computer Science*, 81, 107–113.

Lileikytė, R., Lamel, L., Gauvain, J., Gorin, A. (2018). Conversational telephone speech recognition for Lithuanian. *Computer Speech and Language*, 49, 71–92.

Norkevičius, G., Raškinis, G., Kazlauskienė, A. (2005). Knowledge-based grapheme-to-phoneme conversion of Lithuanian words. In: *SPECOM 2005, 10th International Conference Speech and Computer*, pp. 235–238.

Pakerys, A. (2003). *Lietuvių bendrinės kalbos fonetika* [Phonetics of standard Lithuanian]. Vilnius, Enciklopedija, 35, pp. 83–84.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, P., Schwarz, P., Silovsky, J., Stemmer, G., Vesely, K. (2011a). The Kaldi speech recognition toolkit. In: *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding (ASRU)*.

Povey, D., Burget, L., Agarwal, M., Akyazi, P., Kai, F., Ghoshal, A., Glembek, O., Goel, N., Karafiát, M., Rastrow, A., Rose, R. C., Schwarz, P., Thomas, S. (2011b). The subspace Gaussian mixture model – a structured model for speech recognition. *Computer Speech and Language*, 25(2), 404–439.

Povey, D., Peddinti, V., Galvez, D., Ghahremani, P., Manohar, V., Na, X., Wang, Y., Khudanpur, S. (2016). Purely sequence-trained neural networks for asr based on lattice-free MMI. In: *Proceedings of Interspeech-2016*, pp. 2751–2755.

Raškinis, G., Raškinienė, D. (2003). Parameter investigation and optimization for the Lithuanian HMM-based speech recognition system. In: *Proceedings of the Conference "Information Technologies 2003"*, pp. 41–48.

Raškinis, A., Raškinis, G., Kazlauskienė. A. (2003). Speech assessment methods phonetic alphabet (SAMPA) for encoding transcriptions of Lithuanian speech corpora. *Information Technology and Control*, 29(4), 52–55.

Ratkevicius, K., Paskauskaite, G., Bartisiute, G. (2018). Advanced recognition of Lithuanian digit names using hybrid approach. *Elektronika ir Elektrotechnika*, 24(2), 70–73.

Rudžionis, V., Ratkevičius, K., Rudžionis, A., Raškinis, G., Maskeliūnas, R. (2013). Recognition of voice commands using hybrid approach. In: *Information and Software Technologies. ICIST 2013. Communications in Computer and Information Science*, Vol. 403, pp. 249–260.

Salimbajevs, A., Kapočiūtė-Dzikienė, J. (2018). General-purpose Lithuanian automatic speech recognition system. In: *Human Language Technologies – The Baltic Perspective*, pp. 150–157.

Saon, G., Soltau, H., Nahamoo, D., Picheny, M. (2013). Speaker adaptation of neural network acoustic models using i-vectors. In: *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop*, pp. 55–59.

Skripkauskas, M., Telksnys, L. (2006). Automatic transcription of Lithuanian text using dictionary. *Informatica*, 17(4), 587–600.

Šilingas, D. (2005). *Choosing Acoustic Modeling Units for Lithuanian Continuous Speech Recogniton Based on Hidden Markov Models*. PhD Thesis (in Lithuanian).

Zhang, X., Trmal, J., Povey, D., Khudanpur, S. (2014). Improving deep neural network acoustic models using generalized maxout networks. In: *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 215–219.

G. Raškinis (born in 1972) received a PhD in the field of informatics in 2000. Presently, he works at the Center of Computational Linguistics and teaches at the Faculty of Informatics of Vytautas Magnus University. His research interests include application of machine learning techniques to music recognition, speech recognition and natural language processing.

G. Paškauskaitė (born in 1990) received BS and MS degrees from the Department of Automatics, Kaunas University of Technology. She is a PhD student in the Kaunas University of Technology from 2016. Her main research interests include automatic Lithuanian speech recognition.

A. Saudargienė (born in 1970) received a PhD degree in the field of informatics from the Institute of Mathematics and Informatics, Vilnius. Currently she works at the Department of Applied Informatics, Vytautas Magnus University, and Neuroscience Institute, Lithuanian University of Health Sciences. Her research field is learning and memory in artificial and biological neural systems.

A. Kazlauskienė (born in 1964) received a doctor's degree in the field of humanities (philology) in 1998. She teaches at the Department of Lithuanian Studies of Vytautas Magnus University. Her research interests are phonology, phonotactics, accentuation, rhythm, applied linguistics.

A. Vaičiūnas (born in 1976) received a PhD in the field of informatics in 2006. Since then he has worked as software engineer and researcher in various computational linguistics projects. His research interests are human language technologies.