INFORMATICA, 2003, Vol. 14, No. 1, 75–84

Building Medium-Vocabulary Isolated-Word Lithuanian HMM Speech Recognition System

Gailius RAŠKINIS
Faculty of Computer Science, Vytautas Magnus University
Vileikos 8, LT-3035 Kaunas, Lithuania
e-mail: idgara@vdu.lt

Danutė RAŠKINIENĖ
Institute of Information Technologies, Lithuanian University of Agriculture
Studentų 11, LT-4324 Akademija, Kauno r., Lithuania
e-mail: danra@info.lzua.lt

Received: September 2002

Abstract. In this paper, the opening work on the development of a Lithuanian HMM speech recognition system is described. The triphone single-Gaussian HMM speech recognition system based on Mel Frequency Cepstral Coefficients (MFCC) was developed using HTK toolkit. Hidden Markov model's parameters were estimated from phone-level hand-annotated Lithuanian speech corpus. The system was evaluated on a speaker-independent ~ 750 distinct isolated-word recognition task. Though the speaker adaptation and language modeling techniques were not used, the system was performing at 20% word error rate.

Key words: speech recognition, hidden Markov modeling (HMM), HTK toolkit, hand-annotated speech corpora.

1. Introduction

Automatic Speech Recognition is the process of converting an acoustic signal into a sequence of symbols. To recognize the underlying symbol sequence given a spoken utterance, the continuous speech waveform is first converted into a sequence of equally spaced feature vectors. The task of the recognizer is to map between sequences of feature vectors and the wanted underlying symbol sequences.

The dominant speech recognition paradigm of recent years is known as *Hidden Markov Models* (HMM). This approach takes a sequence of feature vectors as a realization of a concatenation of elementary processes described by hidden Markov models. HMM can be seen as a doubly embedded stochastic process with a hidden process that is not observable and can only be observed through another stochastic process that produces the time set of observations. Let the symbols be words composed of phonemes and let each word in a vocabulary be specified in terms of its component phonemes. HMM-based speech recognition system identifies speech by estimating the likelihood of each

phoneme at contiguous, small frames of the speech signal. A search procedure is used to determine the sequence of phonemes with the highest likelihood. This search is constrained to only look for phoneme sequences that correspond to words in the vocabulary, and the phoneme sequences with the highest total likelihood is identified with the spoken word. The detailed formal description of HMM technology can be found in (Young *et al.*, 2000).

2. Related Work

The past decade has witnessed significant progress in speech recognition technology. Word error rates are dropping twice every two years. One of the main factors that has contributed to this progress was the widespread use of HMM. At present, several commercial very large vocabulary HMM-based dictation systems (Dragon Naturally Speaking Solutions; IBM ViaVoice) and development environments for speech recognition applications (ProNexus VBVoice) are available for the English language.

Lithuanian speech recognizers are at the very first stage of their development. Lipeika *et al.* (2002) describe an isolated-word speech recognition system based on a dynamic time warping (DTW) technique. The system extracts linear predictive coding (LPC) features from a speech signal and uses speaker adaptation techniques. The system is trained to recognize Lithuanian spoken digits and has the vocabulary of 12 words. A. Rudžionis and V. Rudžionis (1999) describe the phoneme recognition system based on regularized discriminant analysis. This system discriminates a subset of Lithuanian phonemes in a fixed context. Some research was done investigating sub-problems of the general Lithuanian speech recognition problem. Feature extraction and normalization of a speech transmission channel are discussed in (Kandratavičius, 2001). The work (Žintelytė, 2001) aims at gaining deeper understanding about phonetic characteristics of certain Lithuanian consonants.

Lithuanian speech recognizers are still solving very constrained recognition tasks or trying to solve the sub-problems of the general Lithuanian speech recognition problem. There were no attempts to build large-vocabulary high-perplexity speech recognition systems. To our knowledge, even medium-vocabulary (> 100 words) recognition tasks have never been tackled. The HMM technology is still unpopular among Lithuanian speech recognizers, the preference being given to dynamic programming techniques (Lipeika, 2000). As a consequence, the aim of this research was to build a medium-vocabulary isolated-word Lithuanian HMM speech recognition system. The research presented in this paper is new and original because of the vocabulary size and the HMM recognition paradigm it is based on.

3. Building HMM-Based Speech Recognizer

Lithuanian HMM speech recognizer was built using HTK toolkit (Young *et al.*, 2000). Firstly, HTK training tools were used to learn HMMs using training utterances and their

associated transcriptions taken from a speech corpus. Secondly, HTK recognition tools were used to transcribe unknown utterances and to evaluate speech recognizer's performance by comparing them to reference transcriptions.

3.1. *Speech Corpus*

The experiments described were based on the VDU isolated-word speech corpus[1]. VDU isolated-word speech corpus contains broadband recordings of 4 speakers (2 males and 2 females) of Lithuanian, each reading 275 utterances that contain 2-5 phonetically rich words. At present, the corpus contains 60.6 minutes of speech[2].

The corpus includes time-aligned phone-level and word-level transcriptions as well as a 16-bit, 44.1 kHz speech waveform files for each utterance. Each recorded waveform is transcribed and verified at the phone level and word-level by a human labeler. The phone-level transcriptions use the original 116 monophone alphabet that covers a variety of Lithuanian speech sounds (Raškinis, 2001a; 2001b) [3]. The alphabet includes accented and unaccented, short and long vowels (*a, à, á, ã, ā, e, ...)*, accented and unaccented diphthongs (*ai, ái, aĩ, au*, ...), soft and hard consonants ($b, b', d, d', \ldots$), and accented consonants (l, m, n, r)[4]. The word-level transcriptions are also phonetically-based and include accent information. For instance, though the word *kasà* meaning *book office* and the word *kãsa* meaning *dig* share the same orthographic form *kasa*, they are considered to be two distinct words. The corpus is based on a 744-word vocabulary, with almost every word being uttered once by all the speakers.

3.2. *Pronunciation Dictionary*

We have manually constructed the pronunciation dictionary for every word in our vocabulary. The pronunciation dictionary consists of records containing words and associated monophone sequences. Some words have multiple pronunciations. The short passage of this pronunciation dictionary is shown in Fig. 1.

3.3. *Language Model*

In the statistical framework the task of the language model or grammar is to predict the next word given the previous word history. The short passage of manually constructed grammar in an extended Backus-Naur form used for our application is shown in Fig. 2.

This elementary language model allows an unconstrained, free word order and results in a recognition task of high perplexity.

---

[1]VDU isolated-word speech corpus is compiled by the Center of Computational Linguistics at Vytautas Magnus University, Kaunas, Lithuania.

[2]The corpus contains 50.9 minutes of pure speech (if pauses between words are not taken into account). The corpus is still growing.

[3]This monophone alphabet is developed at VMU for the purposes of speech recognition research.

[4]Accented consonants are taken from the second half of accented mixed diphthongs (*al, al, an, ar*, ...).

```
ãdresas    ã d’ r’ e s a s
ãpavu      ã p a v u
ãsilas     ã s’ i l a s
ãtvaizdu   ã t v ai z d u
àkti       à k t’ i
àprišu     à p’ r’ i š u
àtvedu     à t’ v’ e d u
áiškus     ái š k u s
álkanas    á l k a n a s
ámžinas    á m’ ž’ i n a s
ámpulė     á m p u l’ ė
ántika     á n’ t’ i k a
ánglas     á nq g l a s
árklį      á r’ k l’ ī
árka       á r k a
áukle      áu k l’ e
áuksas     áu k s a s
ážuolas    á ž uo l a s
āžuolaĩ    ā ž uo l aĩ
apdérgė    a b’ d’ é r’ g’ ė
atgal      a d g a l3
advokãtas  a d v oq k ã t a s
agurkas    a g u r3 k a s
```

Fig. 1. The initial passage of pronunciation dictionary used in building HMM speech recognition system.

```
$word = ãdresas | ãpavu | ãsilas | ãtvaizdu | àkti | àprišu |
        àtvedu | áiškus | álkanas | ámžinas | ámpulė | ántika |
        ánglas | árklį | árka | áukle | áuksas | ážuolas |
        āžuolaĩ | apdérgė | atgal | advokãtas | agurkas |...;
( sent-start < silence $word silence > sent-end )
```

Fig. 2. The passage of grammar used for building HMM speech recognition system. Here, the symbol *$word* denotes a string variable, the vertical bars denote alternatives and the angle braces denote one or more repetitions.

3.4. *HMM Learning*

Lithuanian HMM speech recognizer was built following the sample sequence of processing steps provided in HTK book (Young *et al.*, 2000). The processing scheme is summarized in Fig. 3.

Firstly, sequences of feature vectors were extracted from the raw speech waveforms using HCopy tool. We have investigated two different feature extraction schemes that were both based on Mel Frequency Cepstral Coefficients (MFCC). The first scheme used 8 MFCC coefficients and resulted in a 24-dimensional feature vector (the number, 24, is computed from the length of the static vector (MFCC_0=8), plus 8 delta coefficients, plus 8 acceleration coefficients). The second extraction scheme used 12 MFCC coefficients with energy normalization. It resulted in 36-dimensional feature vector.

Thereafter, each of the required 116 monophone HMMs was generated individually using HInit tool. HInit iteratively computed an initial set of parameter values. On the first iteration, the training data was uniformly segmented. On the second and successive itera-

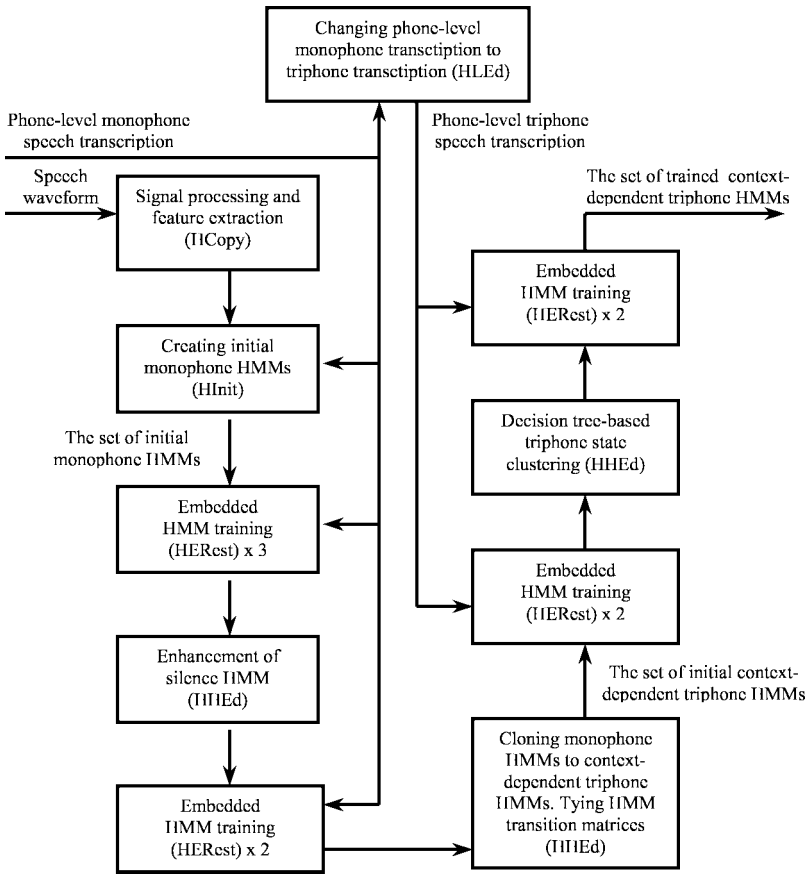

Fig. 3. The data processing scheme used to build Lithuanian HMM speech recognizer. Boxes denote processing steps. Arrows denote data transfer. Parenthesized names indicate the name of the HTK tool that performs necessary processing.

tions Viterbi alignment was used. HInit has been configured to generate single-Gaussian 3-state left-to-right HMMs with no skips.

Once an initial set of models has been created, the tool HERest was used for the first time to perform embedded training. HERest represents the core HTK training tool. It was repeatedly invoked several times (Fig. 3). Each time HERest was invoked it performed a single Baum-Welch (Young *et al.,* 2000) re-estimation of the whole set of monophone (triphone) HMM models simultaneously.

Silence HMM needed to be enhanced. Extra transitions from states 2 to 4 and from 4 to 2 were added in order to make the silence model more robust by allowing individual states to absorb various impulsive noises in the training data.

The next processing step consisted of constructing context-dependent triphone HMMs. HHed was used tool to clone monophone HMMs into context-dependent sets and to force all triphones in a monophone set to share the same transition matrix. In or-

der to be able to make more robust HMM parameter estimates, the decision-tree based clustering was applied to HMM states and then each cluster was tied using HHEd tool.

After the models have been re-estimated twice with HERest, the final set of trained context-dependent triphone HMMs was obtained.

4. Evaluation

In order to assess the accuracy of this Lithuanian HMM speech recognizer several experiments have been conducted:

1. Both 24-dimensional and 36-dimensional MFCC feature extraction schemes have been evaluated, to see how feature extraction affects speech recognizer's performance.

2. Four speaker-independent speech recognition experiments have been performed. Each of these experiments consisted of training speech recognizer on the speech data taken from 3 speakers and testing it on the remaining speech data of 1 speaker who did not take part in the training phase.

3. Additional speaker-dependent speech recognition experiment has been performed. This experiment consisted of training speech recognizer on the speech data of all 4 speakers and testing it on the same data set.

Speech recognition accuracy was computed using the HResults tool. The tool HResults found the optimal alignment comparing machine transcriptions with the original reference transcriptions. Comparison was made on an utterance-by-utterance and a word-by-word basis. The number of substitution errors (S), deletion errors (D) and insertion errors (I) was calculated.

The performance of speech recognition system can be described in terms of word error rate (WER) E defined as:

$$E = \frac{S + I + D}{N} \cdot 100\%.$$

The results of all speech recognition experiments are summarized in the Table 1.

Table 1 reveals encouraging results. The word error rate is as low as 20% for 3 out of 4 speaker-independent recognition experiments. This WER is achieved regardless of the high perplexity of the speech recognition task, without constraining word order and without making use of speaker adaptation techniques. Higher 30-50% WER was observed while recognizing the speech produced by the speaker #1. This can be explained by the speaker's #1 slower than average speech production rate and its voice particularities. The number of MFCC coefficients used to train HMMs appeared to have no decisive impact upon recognition accuracy.

Table 1

Speech recognition accuracy using 24-dimensional (above) and 36-dimensional (below) MFCC-based feature vectors. TRN denotes speakers in the training set, TST denotes speakers in the test set, H, D, S, I, N denote matches, deletions, substitutions, insertions and the total number of recognition entities (utterances, words) respectively. E denote word error rate

TRN	TST	Utterances				Words					
		H	*S*	*N*	***E*, %**	*H*	*D*	*S*	*I*	*N*	***E*, %**
1,2,3	4	181	94	275	**34.18**	619	0	112	20	731	**18.05**
1,2,4	3	176	99	275	**36.00**	609	0	125	11	733	**18.55**
1,3,4	2	178	97	275	**35.27**	616	0	113	56	729	**23.18**
2,3,4	1	146	129	275	**46.91**	564	0	167	78	731	**33.52**
1–4	1–4	1007	93	1100	**8.45**	2830	0	94	7	2924	**3.45**

TRN	TST	Utterances				Words					
		H	*S*	*N*	***E*, %**	*H*	*D*	*S*	*I*	*N*	***E*, %**
1,2,3	4	185	90	275	**32.73**	629	0	103	33	731	**18.60**
1,2,4	3	180	95	275	**34.55**	610	0	123	19	733	**19.37**
1,3,4	2	184	91	275	**33.09**	629	0	100	49	729	**20.40**
2,3,4	1	131	144	275	**52.36**	514	0	216	129	731	**47.20**
1–4[5]	1–4	1044	56	1100	**5.09**	2865	0	59	2	2924	**2.09**

[5] The list of all substitution and insertion errors for the experiment is presented in the Appendix A.

5. Future Work

There are many different ways to improve HMM speech recognizer's accuracy. First of all, we expect that recognition accuracy will improve if HMMs are trained on more speech data. Secondly, there is a set of possibilities offered by HTK toolkit that still remained unexplored. Among them there is a possibility to use alternative features such as LPC, LPCC, PLP, the possibility to build Multiple-Gaussian HMMs, the possibility to build HMMs having non-uniform number of states, the possibility of using data driven HMM state clustering instead of decision tree used in this research.

It is planned to build a large-vocabulary statistical Lithuanian language model in a more distant future. This model should reduce both the perplexity of the task constraining word order and the search. Alternative to monophones sub-word units such as syllables should be investigated. This would enable to cope with highly flexional nature of Lithuanian that results in a very large voacabulary if straight-forward approach is used.

6. Conclusions

In this paper the process of building Lithuanian HMM speech recognizer for the medium vocabulary isolated-word speaker-independent speech recognition task has been described. The triphone single-Gaussian HMM speech recognition system based on Mel

Frequency Cepstral Coefficients (MFCC) was developed using HTK toolkit. Hidden Markov model's parameters were estimated from phone-level hand-annotated Lithuanian speech corpus. Speech recognizer was tested experimenting with different feature sets and performing speaker-independent tests. The system was evaluated on the speaker-independent ~ 750 distinct isolated-word recognition task and performed at 20% word error rate on 3 out of 4 speaker-independent recognition tests. The results are encouraging taking into account that they were obtained without constraining word order and without making use of speaker adaptation techniques. Speech recognizer's accuracy is expected to increase if the size of the training corpus is increased or if the appropriate language model (grammar) is provided.

Acknowledgments

We would like to thank the Speech Vision and Robotic Group at Cambridge, the original developers of the HTK toolkit, and Paul Boersma (University of Amsterdam) the developer of the Praat analysis tool.

Appendix A

An example list (complete) of substitution and insertion errors obtained for the speaker-dependent recognition experiment

Reference	Recognized as	Reference	Recognized as
ãpavu	kãla válgo	apsisiaũsk	apsisiaũs
artóju	artójo	baĩdo	aĩtrų
bejè	biblija	kóšė	kõšė
béldė	belgę	mažù	mažumà
biržẽlio	dvejì žẽlia	obelìs	gobelènas
blogýbė	blogýbe	penkì	pintìnė
dėdė	geĩdė	rõgės	rogès
dárbu	dárbo	rýkšte	rýkštė
dègdavo	berdavo	rýkštė	rýkšte
dvejì	geĩdė	ritaũ	ràsdavo
èpas	rẽtas	suñkiai	sunkì
ėda	gėda	takaĩ	pirkaĩ
fãzė	fãzės	tamsù	tamsiũ
gìrdė	birbė	tempas	kampas
gõdų	guldo	tykiù	sykiù
ìrti	nepìlti	tostù	pū̃stù
japònas	impòrtas	úodas	púodas
judù	hipiùs	ùgnį	tankmė
kačiũ	karčiũ	válgė	béldė
káukė	kaũkia	vidùs	hipiùs
kels	žìlas	vidùs	medùs
kempìnė	pintìnė	žiũri	šiùkšlė

References

Charniak, E. (1993). *Statistical Language Learning*. Massachusetts Institute of Technology, Cambridge.

Cole, R.A., J. Mariani, H. Uszkoreit, A. Zaenen, V. Zue, G.B. Varile and A. Zampolli (1996). *Survey of the State of the Art in Human Language Technology*. `http://cslu.cse.ogi.edu/HLTsurvey/HLTsurvey.html`

Dragon Naturally Speaking Solutions. `http://www.scansoft.com/naturallyspeaking/`

IBM ViaVoice. `http://www-3.ibm.com/software/speech/`

Jelinek, F. (1997). *Statistical Methods for Speech Recognition*. Massachusetts Institute of Technology, Cambridge.

Jurafsky, D., J.H. Martin (2000). *Speech and Language Processing*. Prentice-Hall, New Jersey.

Kandratavičius, I. (2001). Telephone channel influence compensation using dynamic shift cms in speech recognition. *Information Technology and Control*, **20**(3), 51–55.

Lipeika, A. (2000). Optimal segmentation of random sequences. *Informatica*, **11**(3), 243–256.

Lipeika, A., J. Lipeikienė and L. Telksnys (2002). Development of isolated word speech recognition system. *Informatica*, **13**(1), 37–46.

ProNexus VBVoice. `http://www.pronexus.com/`

Raškinis, A. (2001a). Automatic segmentation and transcription of Lithuanian speech. Research program: *Lithuanian Language and Information Society 2000–2006*, 2001 year report. VDU, Kaunas (in Lithuanian).

Raškinis, A. (2001b). *The Atlass of Phonetic Units*. Manuscript, VDU, Kaunas (in Lithuanian).

Rudžionis, A., and V. Rudžionis (1999). Phoneme recognition in fixed context using regularized discriminant analysis. In *Proceedings of 6th European Conference on Speech Communication and Technology Eurospeech'99*. pp. 2745–2748.

Young, S., D. Kershaw, J. Odell, D. Ollason, V. Valtchev and P. Woodland (2000). *The HTK Book*. `http://htk.eng.cam.ac.uk/docs/docs.shtml`

Žintelytė, M. (2001). Methods for analysis of phonetic characteristics of lithuanian fricatives. *Information Technology and Control*, **20**(3), 56–60.

G. Raškinis (born in 1972) received his M.Sc. degree in artificial intelligence and pattern recognition from the University of Pierre et Marie Curie in Paris in 1995. He is a doctor of Physical Sciences (09P) since 2000. Presently, he works at the Center of Computational Linguistics and teaches at the Faculty of Informatics of VMU. His research interests include application of machine learning techniques to human language processing.

D. Raškinienė (born in 1945) studied mathematics at the Lomonosov Moscow State University. Her master's degree (1968) and her PhD (1973) are on numerical methods. She worked at the Vilnius University. At the moment she works at the Lithuanian University of Agriculture as associated professor in the field of mathematical modeling.

Vidutinio žodyno atskirai tariamų žodžių lietuvių šnekos atpažinimo sistema pagrįsta paslėptais Markovo modeliais

Gailius RAŠKINIS, Danutė RAŠKINIENĖ

Šiame straipsnyje aprašomas darbas, kuriant ir tiriant lietuvių šnekos atapžinimo sistemą, grindžiamą paslėptais Markovo modeliais (HMM). Panaudojant HTK programinę įrangą, sukurta HMM pagrįsta lietuvių šnekos atpažinimo sistema. Šnekos atpažinimo sistemą charakterizuoja kontekstinių fonemų paslėpti Markovo modeliai, kurių būsenas aprašo viena Gauso f-ja. Atpažinimo sistema kaip požymius naudoja melų dažnio kepstro koeficientus (MFCC). HMM buvo apmokyti naudojant rankiniu būdu fonemų lygmenyje anotuotą lietuvių šnekos garsyną. Atpažinimo sistemos tikslumas buvo įvertintas sprendžiant nuo kalbėtojo nepriklausomą $\sim$750 skirtingų atskirai tariamų žodžių atpažinimo uždavinį. Nenaudojant jokių uždavinio sudėtingumą sumažinančių priemonių, tokių kaip kalbos modelis (gramatika) ar prisitaikymas prie konkretaus kalbėtojo, pasiekta 20% žodžių atpažinimo klaida.