RAPID DEVELOPMENT OF A LATVIAN SPEECH-TO-TEXT SYSTEM

Ilya Oparin[1], *Lori Lamel*[2], *Jean-Luc Gauvain*[2]

[1]LNE, National Metrology and Testing Laboratory, Trappes, France
[2]LIMSI-CNRS, Spoken Language Processing Group, Orsay, France
ilya.oparin@lne.fr, {lamel,gauvain}@limsi.fr

ABSTRACT

This paper describes the development of a Latvian speech-to-text (STT) system at LIMSI within the Quaero project. One of the aims of the speech processing activities in the Quaero project is to cover all official European languages. However, for some of the languages only very limited, if any, training resources are available via corpora agencies such as LDC and ELRA. The aim of this study was to show the way, taking Latvian as example, an STT system can be rapidly developed without any transcribed training data. Following the scheme proposed in this paper, the Latvian STT system was developed in about a month and obtained a word error rate of 20% on broadcast news and conversation data in the Quaero 2012 evaluation campaign.

Index Terms— Speech recognition, Latvian, under-resourced language

1. INTRODUCTION

The Quaero project (www.quaero.org) aims at developing technologies for automatic analysis and classification of multimedia and multilingual documents. Automatic speech processing is one of the main axes of this project and state-of-the-art speech-to-text (STT) systems are being developed for all official European languages [1]. For some languages, like English, French or German, there is no lack of transcribed and labeled speech corpora and text data that can be used for training. However this is not the case for many other languages as, for example, Hungarian, Bulgarian, Slovak or Latvian. For many of those languages no transcribed acoustic data exist and classical approaches for acoustic model (AM) training can not be applied. Latvian is one such language, for which no ready-to-use audio and text corpora exist and, to the best of our knowledge, no large vocabulary speech recognition system has previously been developed. Using Latvian as an example, this work shows how a state-of-the-art STT system can be developed for a low-resourced European language.

There are two presuppositions in the presented approach. First, it is assumed that it is possible to find some amounts of untranscribed raw audio (Internet radios, podcasts, etc.) and text data (news, blogs, etc.) on the Internet. This data can be used in an unsupervised manner for acoustic model training [2, 3]. Second, it is supposed that there already exist acoustic models for other languages, that cover phones similar to the language in question. These borrowed phone models are used to decode the untranscribed data (cross-language porting) and the one-best decoding hypotheses are taken as the ground truth to improve and refine the AMs for the language under development. Multiple decoding passes over untranscribed audio data are performed to enhance the models, to switch from phone to triphone based AMs, from gender independent to gender dependent ones, to add MLP acoustic features and neural network language models (LMs), etc. An important contribution of this paper is to show that such an approach, based on coherent application of existing tools, allowed quick development of an STT system for a language like Latvian, with a word error rate (WER) of 20% on broadcast news (*bn*) and broadcast conversation (*bc*) data.

Since the latter part of 1990's [4, 5], interest in the transcription of broadcast data (often called 'found data') has been growing. Some of the dimensions being addressed are wider content, that is covering more interactive broadcasts (generally called 'broadcast conversation'), or varied web data (podcasts, lectures), extended language coverage, and extended use of unsupervised training methods [2, 6].

Rapid development of STT systems for new languages has recently received a lot of attention [7, 8]. A number of approaches to vocabulary and dictionary generation, LM data collection and AM training were proposed (e.g. [9, 10, 11, 12]). One can argue that the last issue is often the most critical for rapid STT system development for a new language. As such, multilingual approaches to acoustic model training are currently a popular research direction [13, 14, 15, 16, 17].

In this paper it is shown that, even without any transcribed data available for the language in question, a rather straightforward approach that combines cross-lingual porting of initial seed models and uses existing acoustic and language model training tools, can be successfully used to build a state-of-the-art STT system for broadcast data in under a

This work has been partially supported by the Quaero program, funded by OSEO, French state agency for innovation.

Research done by the first author while he was with LIMSI-CNRS.

Text corpus	period	# sentences	# words
radio1	2008-2011	92k	1.6M
news1	2010-2011	34k	0.9M
news2	2000-2011	3.4M	85.3M
news3	2003-2011	2.1M	54.8M
total	2000-2011	5.6M	142.6M

Table 1. *Data size of available Latvian text corpora after normalization.*

LM	wgt	ppl	1gr	2gr	3gr	4gr
news2	0.44	907	16.2	48.6	26.4	8.8
news3	0.31	961	19.0	49.0	24.5	7.5
radio1	0.16	1995	49.2	40.3	8.7	1.9
news1	0.09	2583	54.0	37.9	6.9	1.2
int 4gr	-	721	13.8	46.8	28.8	10.6
int 3gr	-	741	13.8	46.8	39.4	-
int 2gr	-	914	14.1	85.9	-	-

Table 2. *Interpolation weights (wgt), perplexities with component LMs (ppl) and hit-rates on dev12 set for LMs trained on 4 subcorpora separately and after interpolation.*

2. LATVIAN DATA

month.

Latvian is a Baltic language, spoken by about 1.5-2 million people. It is a highly inflective language with many wordforms corresponding to one lemma. Latvian can be considered as an under-resourced language as there are very few available audio and text corpora. To our knowledge, there are no available Latvian audio corpora in the LDC catalog or other sources for large vocabulary continuous speech recognition. Raw (untranscribed) broadcast audio data was thus crawled from the Internet. Luckily, quite a lot of Latvian audio data (over 800 hours) were found.

Concerning text data, there exist corpora available online, such as the Balanced corpus of modern Latvian and Parliament sessions transcripts. As the STT task in the Quaero project addresses the automatic transcription of broadcast news and broadcast conversation data, these corpora do not seem to be of particular interest and, in addition, they are accessible only via a search interface. The LM training data were thus collected from the Internet with the focus on broadcast news and interview transcripts. These data were subsequently normalized and re-cased following [18].

The sizes of text subcorpora after normalization are given in Table 1. The *news3* corpus has some intersection with the *news2* as both Internet resources sometimes presented news provided by the same news agency. Instead of filtering text data, it was decided to keep all of the texts. As the final n-gram LM is trained as interpolation of n-gram LMs trained independently on all the four subcorpora, this issue is handled at the level of interpolation weights.

model	*projection size*	*hidden size*
NNLM 1	250	450
NNLM 2	200	500
NNLM 3	220	430
NNLM 4	300	500

Table 3. *Common parameters of NNLMs for different languages in the Quaero 2012 evaluation campaign.*

All data predate August 2011, which corresponds to the beginning of the development/test data period for the Quaero 2012 evaluation. No data after this date were collected.

3. LANGUAGE MODELS

3.1. Baseline language models

The recognition vocabulary is chosen as the one consisting of all words across 4 merged normalized subcorpora that occur at least 3 times. The size of this vocabulary is 560766 words and the out-of-vocabulary (OOV) rate on the Quaero 2012 development set (*dev12*) is 0.5%.

Interpolation weights (*wgt*), LM perplexities (*ppl*) and hit-rates (*1gr*, *2gr*, *3gr* and *4gr*) with this vocabulary are presented in Table 2. For the interpolated models results are also given for the lower-order n-gram LMs (in the last three rows for the interpolated 4-gram, 3-gram and 2-gram LMs).

It is worth noticing that, due to the inflective nature of the Latvian language and the relatively small amount of LM training data, the 4-gram hit rate is rather low, as well as the difference in perplexity between 4-gram and 3-gram LMs.

3.2. Neural network language models

Neural network language models (NNLMs) were also trained on Latvian data. These are four-gram feed-forward NNLMs that make use of a shortlist at the output layer [19, 20]. The shortlist size is 12k words. Four feed-forward shortlist NNLMs were built for each corpus and interpolated together. Each of four NNLMs differs in sizes of projection and hidden layers and uses slightly different resampling of training data (see Table 3). The perplexity results for individual NNLMs and in interpolation with the baseline 4-gram LM are presented in Table 4. A thorough description of NNLM configurations may be found in [21].

4. PRONUNCIATION MODELING

The Latvian orthography is based on Latin. The vowel letters A, E, I and U with a macron (Ā, Ē, Ī and Ū) are long versions of corresponding short counterparts. The letters C, S and Z, that in unmodified form are pronounced [ts], [s] and [z] respectively, with a caron (Č, Š and Ž) are pronounced as [tʃ],

model	perplexity	weight
4gr LM	**721**	0.29
NNLM1	675	0.17
NNLM2	671	0.18
NNLM3	677	0.17
NNLM4	668	0.19
4 NNLMs	**622**	-
4 NNLMs + 4gr LM	**577**	-

Table 4. *LM perplexities (stand-alone and in interpolation) on the Quaero dev12 data.*

[ʃ] and [ʒ]. The letters Ģ, Ķ, Ļ and Ņ that are written with a cedilla (or little 'comma' placed above the lowercase "g") are the palatalized versions of G, K, L and N and correspond to the sounds [ɟ], [c], [ʎ] and [ɲ].

Latvian spelling has almost perfect correspondence between graphemes and phonemes and every phoneme corresponds to its own letter. Latvian orthography has nine digraphs (seven vowels and two consonants), which are written as *ai, au, ei, ie, iu, ui, oi, dz* and *dž*. Corresponding phone units were explicitly modeled. Standard Latvian has fixed initial stress.

According to the rules mentioned above a simple grapheme-to-phoneme (g2p) script was developed for Latvian. The dictionary includes 42 different phonetic units, namely 26 consonants, 9 vowel phonemes and 7 diphthongs.

5. ACOUSTIC MODELING

As no labeled Latvian acoustic data was available for training, seed models from other languages, namely English, French and Russian were used for bootstrapping Latvian acoustic models. These models are used to decode untranscribed Latvian acoustic data, from which a first set of Latvian acoustic models were built. As it was mentioned above, the unsegmented acoustic data was collected from Latvian Internet radio sources. At each iteration, a larger set of audio data was decoded with the previous system, and the hypotheses were used as ground truth for training the next set of models [6, 18].

Table 5 summarizes the audio data used in successive acoustic model sets along with the model sizes. In the first 4 iterations decoding was done with models using PLP+F0 features, whereas the last decoding was done using models with MLP+PLP+F0 features.

As described in [22] the acoustic models are tied-state, left-to-right 3-state HMMs with Gaussian mixture observation densities (typically 32 components). The triphone-based phone models are word-independent, but position-dependent. The states are tied by means of a decision tree. Both gender-independent and gender-dependent models were tested as well as different sized silence models (96 or 1024 Gaussians).

Iteration	#audio files	duration (hours)	#contexts
1	1000	63	45
2	2633	156	14384
3	8770	521	20631
4	12983	769	23760
5	12983	783	26161

Table 5. *Acoustic training data and triphone coverage.*

6. STT SYSTEM DEVELOPMENT

STT system development initially focused on training acoustic models for Latvian, and made use of the interplated n-gram LMs, summarized in Table 2.

Context and gender independent PLP-based seed models borrowed from English, French and Russian were used for the first iteration over the training data. This PLP analysis has been used in all LIMSI STT systems since 1996 and is described in [22]. Using these acoustic models directly to decode the Latvian development data (*dev12*) resulted in a WER of about 74%.

Context-dependent, gender-independent models were trained during the second iteration on about 150 hours of audio data. With these updated AMs, covering about 14k phone contexts a WER of 53% was obtained.

Gender and context-dependent AMs were then trained during the third training iteration. Gender-independent models serve as priors for Maximum *a Posteriori* (MAP) estimation of gender-dependent models. The gender labels are produced automatically, by the audio partitioner [23]. These models were trained on about 500 hours of audio data as shown in Table 5. More data were used in the fourth iteration, and a larger number of phones in context were modeled.

Table 6 gives recognition results at different stages of acoustic model development (that is after several decoding iterations). In this table *CI WER* stands for the case-insensitive WER and *CS WER* corresponds to the case-sensitive WER. It can be seen that the size of the silence model does not have much influence on the STT performance. Gender-dependent (GD) models (models *3*, *4*, *7*, *8*) slightly improve over the gender-independent (GI) ones (models *1*, *2*, *5*, *6*).

Although the MLP features were not used during training at the fourth iteration, the addition of MLP features to the PLP ones at the testing phase significantly improved the results, by almost 10% absolute (models *5*, *6*, *7*, *8*). The MLP features are based on the Bottle-Neck architecture [24] and the MLP parameters that existed for another language (in this case Russian) were borrowed, as it was done with phone models at the first training iteration.

It should be noted that the recognition results depend on the decoder parameters. The parameters used with all the models in Table 6 were kept the same and were not tuned for each set of models individually. Thus, for example, a big gain

model	GD	MLP	silence	WER	
				CI	CS
1			96	40.4	42.5
2			1024	40.4	42.6
3	√		96	39.4	41.7
4	√		1024	38.8	41.2
5		√	96	28.4	31.5
6		√	1024	27.9	31.1
7	√	√	96	28.1	31.2
8	√	√	1024	27.4	30.6

Table 6. *Latvian STT results on dev12 set with context-dependent AMs from successive training interations.*

dev12	perplexity	CI WER	CS WER
4gr LM	721	27.4	30.6
4 NNLMs + 4gr	577	25.9	29.1

Table 7. *Recognition results with NNLMs (with the model 8).*

with MLP features may be due to a better fit of the decoding parameters with the MLP+PLP+F0 based AMs.

The best STT results are obtained with context and gender dependent AMs that use MLP+PLP+F0 features (model *8*).

The 4-gram lattices generated with this model and the baseline 4-gram LM for the test data were rescored with the neural network models. This resulted in an additional gain of about 1.5% absolute, as presented in Table 7. The NNLMs are used at the test phase and never during training due to high time costs and the simple 1-pass decoding of the training data runs in approximatively real-time.

As the last fifth iteration over the training data, a different weaker silence model was used in the dictionary, that kept strict separation between silence, filler word and breath events. The new AMs were trained after re-decoding of the training data with MLP+PLP+F0 based models. The results are presented in Table 8. The last row (*12+NNLM*) corresponds to the application of the neural network LMs on test data lattices, generated with the best model *12*.

model	GD	MLP	silence	WER	
				CI	CS
9		√	96	26.9	30.2
10		√	1024	26.1	29.5
11	√	√	96	26.0	29.3
12	√	√	1024	25.2	28.7
12+NNLM	√	√	1024	23.6	27.0

Table 8. *Latvian STT results on dev12 set with updated PLP+MLP+F0 AMs from iteration 5.*

dev12	CI WER	CS WER
1st pass	21.6	24.7
2nd pass	20.4	23.3
+ NNLMs	**18.8**	**21.7**

Table 9. *Latvian STT results on dev12 data with tuned decoder parameters and CMLLR/MLLR adaptation.*

Case-Insensitive		Case-Sensitive	
WER	NCE	WER	NCE
20.2	0.213	**23.4**	0.201

Table 10. *Latvian STT results on Quaero 2012 eval data.*

7. 2012 QUAERO EVALUATION

The aim of experiments reported above was to compare different AM configurations and to track progress. Thus, the decoding parameters were not optimized for each individual model and only one decoding pass was used for speed. The decoder parameters, such as LM score scale, word and silence insertion penalties, of a system for the submission to the evaluations must be tuned on development data.

In preparation for the Quaero 2012 evaluation, the decoder parameters were tuned on dev12 data. This system also makes use of two passes of decoding with unsupervised CMLLR/MLLR adaptation [25, 26], as compared to the simple one-pass strategy used during training. The results with the model *12* (see Table 8) with the parameters tuned at each pass are given in Table 9.

Finally, results with the STT system submitted to the Quaero 2012 evaluation campaign are presented in Table 10.

8. CONCLUSIONS

The approach presented in this paper was used to develop a Latvian STT system in a short time. The only supervision consists in mapping of Latvian phones to the ones existing in other languages to select cross-lingual seed models, text normalization and grapheme-to-phoneme conversion. The latter is very straightforward for Latvian and we expect it could easily be substituted with a graphemic model. The Latvian STT system was rapidly developed without use of an transcribed audio training data. Coherent iterative application of existing tools to untranscribed Latvian audio data led to a reduction from the initial word error rate of 74% at the first pass down to 19% at the fifth training interation. Only the Quaero 2012 development data, split evenly between broadcast news and broadcast conversation, had transcriptions. This LIMSI Latvian STT system, developed in about a month period, obtained a WER of 20.2% in the Quaero 2012 STT evaluation.

9. REFERENCES

[1] L. Lamel, "Multilingual Speech Processing Activities in Quaero: Application to Multimedia Search in Unstructured Data," in *HLT'12*, Tartu, Estonia, Oct 2012.

[2] T. Kemp and A. Waibel, "Unsupervised Training of a Speech Recognizer: Recent Experiments," in *ESCA Eurospeech'99*, Budapest, Hungary, 1999, pp. 2725–2728.

[3] L. Lamel, J.L. Gauvain, and G. Adda, "Lightly Supervised Acoustic Model Training," in *ISCA ITRW ASR 2000*, Paris, Sep 2000, pp. 150–154.

[4] J.L. Gauvain, L. Lamel, and G. Adda, "Transcribing Broadcast News for Audio and Video Indexing," *Communications of the ACM*, vol. 43, no. 2, pp. 64–70, Feb 2000.

[5] D. S. Pallett, "The role of the national institute of standards and technology in darpa's broadcast news continuous speech recognition research program," *Speech Communication*, vol. 37, no. 1-2, pp. 3–14, 2002.

[6] L. Lamel, J.L. Gauvain, and G. Adda, "Lightly supervised and unsupervised acoustic model training," *Computer Speech & Language*, vol. 16, no. 1, pp. 115–229, 2002.

[7] T. Schultz and K. Kirchhoff, Eds., *Multilingual Speech Processing*, Elsevier, 2006.

[8] D. Povey, N. Goel, L. Burget, M. Agarwal, P.Akyazi, F. Kai, A. Ghoshal, O. Glembek, M. Karafiát, A. Rastrow, R.C. Rose, P. Schwarz, and S. Thomas, "Low development cost, high quality speech recognition for new languages and domains," *Report from 2009 Johns Hopkins/CSLP Summer Workshop*, 2009.

[9] T. Slobada and A. Waibel, "Dictionary learning for spontaneous speech recognition," in *Proceedings of ICSLP'96*. IEEE, 1996.

[10] N. Goel, S. Thomas, M. Agarwal, P. Akyazi, L. Burget, K. Feng, A. Ghoshal, O. Glembek, M. Karafiát, D. Povey, A. Rastrow, R.C. Rose, and P. Schwarz, "Approaches to automatic lexicon learning with limited training examples," in *ICASSP'10*. IEEE, 2010, pp. 5094–5097.

[11] D. Povey, S.M. Chu, and B. Varadarajan, "Improving trigram language modeling with the World Wide Web," in *ICASSP'01*. IEEE, 2001, pp. 533–536.

[12] T. Schultz and A. Waibel, "Language-independent and language-adaptive acoustic modeling for speech recognition," *Speech Communication*, vol. 35, no. 1, pp. 31–52, 2001.

[13] D. Povey, S.M. Chu, and B. Varadarajan, "Universal background model based speech recognition," in *ICASSP'08*. IEEE, 2008, pp. 4561–4564.

[14] L. Burget, P. Schwarz, M. Agarwal, P. Akyazi, K. Feng, A. Ghoshal, O. Glembek, N. Goel, M. Karafiát, D. Povey, A. Rastrow, R.C. Rose, and S. Thomas, "Multilingual acoustic modeling for speech recognition based on subspace gaussian mixture models," in *ICASSP'10*. IEEE, 2010, pp. 4334–4337.

[15] N.T. Vu, F. Kraus, and T. Schultz, "Rapid building of an ASR system for under-resourced languages based on multilingual unsupervised training," in *Interspeech'11*, 2011, pp. 3145–3148.

[16] N.T. Vu, F. Kraus, and T. Schultz, "Cross-language bootstrapping based on completely unsupervised training using multilingual a-stabil," in *ICASSP'11*. IEEE, 2011, pp. 5000–5003.

[17] N.T. Vu, F. Metze, and T. Schultz, "Multilingual bottleneck features and its application for under-resourced languages," in *SLTU'12*, 2012.

[18] L. Lamel and B. Vieru, "Development of a speech-to-text transcription system for Finnish.," in *SLTU'10*, 2010, pp. 62–67.

[19] H. Schwenk and J.-L. Gauvain, "Connectionist language modeling for large vocabulary continuous speech recognition," in *ICASSP'02*. IEEE, 2002, pp. 765–768.

[20] H. Schwenk, "Continuous space language models," *Compututer, Speech & Language*, vol. 21, no. 3, pp. 492–518, 2007.

[21] H.-S. Le, I. Oparin, A.Allauzen, J.-L. Gauvain, and F.Yvon, "Structured output layer neural network language models for speech recognition," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 21, no. 1, pp. 197–206, 2013.

[22] J.L. Gauvain, L. Lamel, and G. Adda, "The LIMSI broadcast news transcription system," *Speech Communication*, vol. 37, no. 1-2, pp. 89–108, 2002.

[23] J.L. Gauvain, L. Lamel, and G. Adda, "Partitioning and transcription of broadcast news data," in *ICSLP*, Sydney, Australia, Dec 1998, vol. 4, pp. 1335–1338.

[24] F. Grézl and P. Fousek, "Bottle-neck features for LVCSR," in *ICASSP'08*. IEEE, 2008, pp. 4729–4732.

[25] C.J Legetter and P.C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density HMMs," *Computer, Speech & Language*, vol. 9, no. 2, pp. 171–185, 1995.

[26] M.J.F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer, Speech & Language*, vol. 12, no. 2, pp. 75–98, 1998.