Development of the RWTH Transcription System for Slovenian

Pavel Golik[1], *Zoltán Tüske*[1], *Ralf Schlüter*[1], *Hermann Ney*[1,2]

[1]Human Language Technology and Pattern Recognition, Computer Science Department,
RWTH Aachen University, 52056 Aachen, Germany
[2]Spoken Language Processing Group, LIMSI CNRS, Paris, France
{golik,tuske,schlueter,ney}@cs.rwth-aachen.de

Abstract

In this paper we describe the RWTH automatic speech recognition system for Slovenian developed within the *transLectures* project. The project aims at supporting the transcription and translation of video lectures freely available on the web.

Difficulties arise on all levels of modeling: Slovenian is a morphologically rich language with a high level of inflection (pronunciation model), and a large variety of dialects and recording conditions brings uncertainty into the audio signal (acoustic model). Moreover, the video lectures cover a wide spectrum of topics with a high share of spontaneous speech and technical terms (language model). These issues require application of robust and adaptive methods.

Besides the system description, this study mainly focuses on robust acoustic modeling. Building acoustic models from various resources, we also compare the influence of speaker adaptation to different neural network based acoustic features. Systematic application of these methods allows us to reduce the word error rate on the evaluation corpus from 59.2% to 43.4%. We also give a motivation for Slovenian open vocabulary recognition and perform some first steps.

Index Terms: Slovenian, multilayer perceptron, bottleneck features, Tandem, open vocabulary, transLectures

1. Introduction

While a lot of effort within the automatic speech recognition (ASR) community is spent on widely spoken languages like English, French, German, Mandarin Chinese or Arabic, only few groups focus on less spread languages. Slovenian language, or Slovene, spoken by over 2 million speakers, can be considered "low-resource", even in comparison with other Slavic languages. Since it is one of the official languages of the European Union, more and more audio and video documents become digitally available that need to be translated, accessed by disabled people or just effectively searched. This motivates the development of robust ASR systems for this language.

First of all, Slovenian, just like many other Slavic languages, differs e.g. from English by a high level of inflection. The grammatical categories like case, number, person or tense are mostly indicated by appending suffixes to the word root. Sometimes, prefixes are used to modify the semantics of a word. A native speaker can understand a new word by a simple morphosyntactical analysis, if he is familiar with the semantics of its parts. A typical English ASR system would usually rely on a fixed pronunciation lexicon and consider even slight modifications of any word not covered explicitly as an out-of-vocabulary (OOV) word which, in general, can not be recognized.

Second, spoken Slovenian is characterized by a less strict word order than written language. The semantic relationship between the sentence parts is covered by morphology and is therefore mostly unambiguous. This makes the language modeling with n-grams, commonly used in ASR, more difficult than in some other languages.

From the acoustic point of view, the challenge in building a Slovenian ASR system lies in a large variety of dialects. Although the number of native speakers is comparably low, the language can be divided into up to 46 dialects (different estimations exist). The dialects differ mostly by the phonology, leading to many valid pronunciation variants of the same word. A difficulty in modeling this by a pronunciation lexicon arises from a number of homophones in different dialects. Moreover, some words have completely different regional forms.

The *transLectures* project aims at transcribing and translating recorded lectures that are available on the web, e.g. on *videolectures.net*. Academic lectures usually contain one main speaker and some short replicas from the audience in the beginning and in the end. Most of the lectures can be qualified as spontaneous speech, which is prone to grammatical errors, mispronunciation, colloquialisms, fillers etc. As a project participant, RWTH developed ASR systems for different languages and this work describes the Slovenian system in detail.

This paper is structured as follows. We first describe the development of the acoustic models in Section 2. Then we provide details on the construction of the language models in Section 3. The experimental results are reported in Section 4, where we also discuss open vocabulary recognition. The conclusions are drawn in Section 5.

2. Acoustic modeling

Our initial systems were trained and evaluated on the data available within the *transLectures* project, denoted as *tl-train*, *dev* and *eval*. Later, we added data from two popular Slovenian speech corpora: GOS, which contains TV/radio shows, telephone speech and recorded conversations[1], and BNSI, which is a collection of broadcast news [1]. The corpora details are given in Table 1. The recordings were downsampled to 16 kHz mono. We first compare various short-term spectral features. Then we make use of features derived from a multilayer perceptron (MLP). Finally, speaker adaptation is applied to the best systems to improve the acoustic models (AM).

2.1. Pronunciation lexicon

Slovenian language has relatively straightforward pronunciation rules. Mostly there is a one-to-one correspondence between graphemes and phonemes, although some context dependencies exist. For the initial system we started with a manually created

[1] http://www.korpus-gos.net

Table 1: *Acoustic data.*

Corpus	Dur. [h]	Content	# words
tl-train	27.0	34 lectures	208,997
gos	39.0	TV shows, conversations	342,116
bnsi	24.0	broadcast news	216,384
dev	3.1	4 lectures	27,303
eval	3.3	4 lectures	22,937

pronunciation 100k lexicon available in the LC-STAR project. After dropping some redundant phonemes (e.g. `F` and `f`) we ended up with 39 phonemes. Two additional non-speech events and the silence were added to the phoneme set. Then we applied a statistical grapheme-to-phoneme (g2p) conversion tool [2] to create pronunciation lexicon for the training set. We included only the first-best hypothesis for each word, resulting in 26k lemmas. It is important to note that the lexicon distinguishes between long/short and stressed/unstressed variants of the vowels.

For the recognition lexicon, we selected the 400k most frequent words across all available text data in order to obtain a small OOV rate (see Section 3 for more details on the data sources). Table 2 shows the relationship between lexicon size, the perplexity on the development set and the OOV rate. After the first recognition pass, we used the top 50 confusion pairs to analyze the possible errors due to wrong pronunciation and fixed a few pronunciation variants to include some dialect versions.

Table 2: *Perplexity and the OOV rate on the development set for different lexicon sizes.*

Lexicon size	PPL	OOV rate [%]
100k	453	5.3
200k	525	3.2
300k	552	2.4
400k	576	2.2

2.2. Baseline acoustic model

All our systems are based on the common HMM structure: across-word position dependent 6-state left-to-right topology with skip and loop transitions. The emission probabilities are modeled with Gaussian mixture models (GMM) with a globally pooled diagonal covariance matrix. The number of triphones is reduced to 4500 by tying with a phonetic classification and regression tree (CART). The GMMs were trained according to the maximum likelihood (ML) criterion.

One system was trained using Mel frequency cepstral coefficients (MFCC) calculated every 10 ms from overlapping windows of 25 ms. The 16-dimensional MFCCs were derived from 20 Mel filters and normalized segment-wise w.r.t the mean and the variance. This normalization is able to reduce variability in the signal that comes from different recording conditions. The variability among speakers can be reduced by performing a vocal tract length normalization (VTLN). This piece-wise linear function can be estimated by performing a grid search over the warping factors and maximizing the acoustic likelihood under a single Gaussian model. The transformation can also be estimated on the test data with a classifier based *Fast-VTLN* [3]. Finally, a sliding window of 9 consecutive frames was projected with linear discriminant analysis (LDA) to a 45-dimensional subspace.

The next system differs by the way of processing the short-term power spectrum. Instead of MFCCs, perceptual linear prediction (PLP) coefficients were extracted from 20 ms windows as described in [4].

Finally, another system was trained on Gammatone (GT) features [5] derived from an audiological filter bank of 68 filters with an infinite impulse response.

2.3. Neural network features

Various methods of integrating neural networks into ASR systems are known. With the progressive research on deep learning and increasingly high computational power of graphical processing units (GPU) this topic becomes prevalent in the ASR community. While the *hybrid approach* replaces likelihood computation of GMM by a rescaled posterior estimate from an MLP [6], the class of *Tandem approaches* uses MLPs for extraction of probabilistic features and trains a GMM based acoustic model in a second stage [7][8].

For the Slovenian ASR system we decided to start with the Tandem approach and compare different ways of including MLP features in the feature space. By choosing Tandem over the hybrid approach, we keep the possibility to perform speaker adaptation in the same way as with the short-term features. The first method, sometimes referred to as *probabilistic Tandem*, concatenates the state posterior estimates of an MLP with the short-term features and re-estimates the GMM parameters. In a pre-processing step, the 42 phoneme posterior estimates are transformed by a logarithm and decorrelated with principal component analysis (PCA).

By taking the MLP output as acoustic feature, we constrain the MLP to learn the 1-of-C coding scheme of the reference labels. While this is a common representation where neural networks are used for classification, it does not fit naturally into the Gaussian modeling used in the final step. A method that allows the MLP to choose a different low-dimensional representation of the input data freely is known as the *bottleneck* approach [8]. Its core idea is to include a relatively small hidden layer in the neural network, thus forcing the MLP to learn the transformation to and from this low-dimensional space. The feature extraction then consists of forwarding the input vectors up to the bottleneck layer, taking its output without any non-linearity and decorrelating it with PCA.

The bottleneck approach has been shown to be more powerful with two extensions: first, the short-term input features to the network are replaced by multi-resolution RASTA (or MRASTA) filtered critical band energies [9]. Second, a hierarchical processing is performed by cascading two MLPs in the following way [10]: the first MLP is trained on the filters corresponding to slow modulation frequencies. Then, its bottleneck features are augmented by the fast modulation frequencies and fed into the second MLP. The bottleneck features from the second MLP are finally used in combination with short-term features to train a new GMM acoustic model.

2.4. Speaker adaptation

One of the advantages of the Tandem approach over the hybrid approach is the capability to apply to the MLP features all kind of processing that is known from training GMMs on short-term features. An important class of such transformations is the speaker adaptation, performed in a supervised or an unsupervised manner. The adaptation can be performed during the training (speaker adaptive training or SAT), during the recognition (by adapting to the recognition result of the first pass system) or, more typically, both. The transformation we applied to the Slovenian *transLectures* system is constrained maximum likelihood linear regression (CMLLR) [11].

In order to estimate similar transformation on training and testing data, we performed unsupervised clustering of the segments rather than relying on the speaker annotations, obtaining "speaker-like" clusters that correspond not only to different speakers, but also to different speaking manners or even recording conditions. This is a suitable approach for recognizing video lectures, that are given by different lecturers and recorded in heterogeneous environments.

Then we used the simple target model approach, which consists of estimating a single linear transform per cluster that maximizes the likelihood under a single Gaussian model [12].

3. Language modeling

We trained 4-gram language models (LM) smoothed by the modified Kneser-Ney method. The texts were pre-processed by converting to lower case, removing the punctuation marks and special characters, leaving one sentence per line. The initial LM was trained on the transcriptions of lectures used for acoustic training. Compared with other languages, the Slovenian LM resulted in very high perplexity. The reason for this is twofold: the number of words is very high due to inflectional forms of nouns, adjectives and verbs in Slovenian, but also the strong topic mismatch between the training and testing data leads to different word statistics.

Increasing the amount of data can compensate for these differences to some extent, so we crawled the web for openly available texts and open corpora. Table 3 shows details on the text sources we found. The sources include transcriptions of the European Parliament Plenary Sessions (EPPS), books from Wikisource, a collection of legislative texts (JRC-Acquis[2]) and transcriptions of TV shows of the Slovenian national public broadcasting organization (RTV[3]).

Table 3: *Language model training data sources with corresponding perplexities (PPL) on the dev set and the interpolation weights λ.*

Corpus	# words	PPL	λ	Content
tl-train	208.0K	978	0.30	Lectures
Web data:				
EPPS	7.1M	1767	0.04	2007-2011
Wikisource	14.6M	1529	0.03	Books
JRC-Acquis	38.6M	4469	0.03	Legal texts
RTV	14.3M	544	0.61	TV shows
Interpolated	74.7M	468		

Excluding the three LMs with the smallest interpolation weights increased the final perplexity by over 15 points. Although only *RTV* LM performed better than the one estimated on *tl-train*, a linear interpolation between all five LMs allowed to obtain the lowest perplexity on the development set.

4. Results

The evaluation of the methods described in this work is structured as follows. First we train several systems on short-term features and look at the different corpora for acoustic training. Then we compare different probabilistic features derived from MLP. Finally, we compare standard language modeling with the open vocabulary approach. In most experiments, we apply speaker adaptive training as described in Section 2.4 and report the first pass results for the speaker independent (SI) model and the second pass results for the speaker adapted (SA) model. Note that the term "speaker" is used for convenience, although the "speaker" labels are obtained by unsupervised clustering and do not necessarily correspond to real speakers.

Three systems were trained on MFCC, PLP and GT features as described in Section 2.2. The training data contained only lectures from the *tl-train* corpus. For the sake of comparability, no VTLN was applied to the MFCC system. As expected, Table 4 shows only minor differences between various short-term features, which is why we chose to stick with MFCC as the simplest of the methods. A simple system combination with ROVER [13] reveals the high potential of improving the acoustic model, as the acoustic features are the only difference between the three systems.

Relatively high error rates do not distribute uniformly among the lectures: the performance ranges between approx. 27% and over 40% WER for different lecturers. This indicates strongly heterogeneous recording conditions and topics.

Table 4: *Single systems trained on the tl-train corpus using short-term features.*

System	WER [%]	
	dev	eval
MFCC	38.8	59.2
PLP	38.7	57.3
GT	39.3	58.3
ROVER	35.2	50.7

Now we turn to the acoustic corpora. For these experiments, we extracted VTLN normalized MFCC features for each of the sets and trained systems on different combinations. Table 5 shows the recognition results along with the total durations. On the one hand, the corpora *gos* and *bnsi* do not outperform *tl-train*, which is presumably due to different recording conditions and speech types (both contain prepared speech to a large extent, whereas the evaluation is performed on spontaneous speech). On the other hand, while combining the corpora barely affects the performance on the development set, the WER on the evaluation set is reduced significantly. This holds true for both, speaker independent and speaker adapted models.

Having trained a system on the *bnsi* corpus, we were able to perform a sanity check by evaluating the ASR system on the *bnsi-dev* and *bnsi-eval* sets. The word error rates of 26.9% and 27.8% indicate that the system has been trained correctly and the high WER on *transLectures* test data is due to the acoustic condition mismatch.

Table 5: *Acoustic training on different corpora using VTLN normalized MFCC features.*

AM data			Dur. [h]	WER [%]			
				SI		SA	
tl-train	gos	bnsi		dev	eval	dev	eval
×			27	38.6	57.6	35.3	47.3
	×		39	43.2	56.8	40.9	49.1
		×	24	42.8	57.8	40.0	48.7
×	×		66	37.9	53.5	35.8	45.5
×	×	×	91	36.8	52.6	35.7	45.4

In the next step we look at different ways of extracting probabilistic features from MLPs as described in Section 2.3. To maintain the comparability with the previous results, we only

[2] http://ipsc.jrc.ec.europa.eu/index.php?id=198

[3] http://www.rtvslo.si

used acoustic data from *tl-train*. We trained neural networks for two systems (referred to as s1 and s2) on 9 consecutive frames consisting of 16-dimensional MFCCs, its first temporal derivative and the second derivative of the 0th coefficient, totaling in $9 \cdot (16 + 16 + 1) = 297$ dimensions. The output labels for both systems s1 and s2 correspond to 42 phonemes, but the network structure varies: while s1 has only one hidden layer with 2000 neurons, s2 has three hidden layers of size 1000, 60 and 1000. The small inner layer is referred to as the bottleneck (BN) layer. During training the activation function in all hidden layers is a sigmoid and the output activation function is a softmax. The training is performed according to the cross entropy criterion by the stochastic gradient descent on mini-batches of size 4096. The input features for the first two GMM Tandem systems are constructed by augmenting LDA transformed MFCCs either with the logarithm of the posterior estimates (s1) or the output of the bottleneck layer without the non-linearity (s2). In both cases the MLP features are decorrelated with a PCA transform retaining 95% of the total variance.

The system s3 follows the hierarchical approach by training two MLPs with the hidden layers of size 5000, 60 and 5000 and the output layers corresponding to 4500 tied triphone classes. The first network is trained on 248-dimensional slow part of MRASTA features. Its linear bottleneck features are then PCA reduced and augmented with the fast MRASTA features before being fed into the second network. The bottleneck features from the second MLP are processed in the same manner as in s2 and concatenated with MFCC to train the last Tandem system.

The results of the different MLP feature extraction methods are presented in Table 6. The systems show better results with increasing complexity, measured by the number of trainable parameters of both MLP and the final GMM acoustic model. An additional line (s4) shows how s3 can be improved by increasing the amount of training data to 91 hours. It reveals the much stronger benefit that MLP draws from additional data compared to the GMM based acoustic model (cf. last line in Table 5). The overall best result was obtained by applying speaker adaptive training to the system s4.

Table 6: *Tandem systems on different MLP based features.*

Tandem system	Dim.	WER [%]			
		SI		SA	
		dev	eval	dev	eval
MFCC baseline	45	38.6	57.6	35.3	47.3
s1: + PCA($log(p(s\|x))$)	45+20	36.8	55.1	32.9	44.4
s2: + PCA(BN(MFCC))	45+25	36.7	54.8	33.0	44.7
s3: + PCA(BN(HMRASTA))	45+33	34.3	49.6	30.9	43.4
s4: s3 + more data	45+39	30.9	43.7	29.5	38.5

4.1. Open vocabulary recognition

As we mentioned in Section 1, morphologically rich languages like Slovenian follow syntactic rules for constructing inflections of words. This is often done by appending prefixes and suffixes to word roots. From the language modeling point of view, it should be sufficient to know the co-occurrence probability of "concepts" encoded by the word roots rather than of their concrete inflectional forms. For example, if we could capture the probability of the English bigram *("green", "apple")*, it will remain the same in different contexts like "with the green apple" or "in the green apple". In Slovenian, however, different cases will cause the endings to change ("z zelen**im** jabolk**om**" vs. "v zelen**em** jabolk**u**").

An analysis of the substitution errors made by the system s4 (cf. Table 6) on the development set revealed that over 17% of the confusion pairs share a prefix of five or more characters. In a vast majority of cases (over 95%) the recognized word corresponds to the wrong inflection of the correct word, since the correct inflection is not present in the pronunciation lexicon.

The standard approach of estimating the probabilities from n-gram counts in the training text spreads the probability mass between different inflections and is highly sensitive to unseen forms. A possible approach to handle this source of uncertainty would be the *open vocabulary recognition* [14], that consists in estimating a hybrid recognition lexicon and LM from text sources where OOV words are split into fragments (or sub-word units), e.g. {`pod+, vrže+, nemu, nimi`} instead of {`podvrženemu, podvrženimi`}). Afterwards, the recognized fragment sequences need to be merged in a post-processing step before the word error rates are calculated.

In a preliminary experiment, we followed [15] and performed an unconstrained segmentation on the OOV words. The fragments were extracted from approx. 570k OOV words not included in the initial 400k lexicon. For simplicity, the pronunciation for the fragments was estimated with the same g2p model as the initial lexicon, although a better result can be expected by segmenting the original phoneme transcription in order to retain correct coarticulation. The final lexicon contained 400k entries, 55k of which were fragments. After replacing the OOV words by the corresponding fragments in the text data, the new LMs were estimated and interpolated in the same manner as described in Section 3. However, this led to an insignificant improvement on the development set of below 0.1% WER. While some OOV words could be recognized by merging the fragments, new errors have been made due to confusions between fragments and short words. More analysis is needed to find a segmentation method that helps the recognition of unseen inflections without harming the other hypotheses.

5. Conclusions

In this work we described the RWTH transcription system for Slovenian video lectures in detail. We compared available data sources for building both the acoustic and the language models. The evaluation included different acoustic short-term features and probabilistic features derived from neural networks. It is notable that the baseline AM trained on MFCC features did not benefit from the additional training data as much as the Tandem system. Most significant improvements were obtained by applying CMLLR based speaker adaptation to the hierarchical MRASTA bottleneck features. The overall best system trained on the *tl-train* corpus (s3) showed an improvement of over 26% relative to the baseline.

An error analysis revealed the strong necessity for better modeling of inflections, being one of the characteristic features of Slavic languages. We described our first attempt to handle this problem by performing open vocabulary recognition.

Our future work will include investigations on unsupervised acoustic training and language model adaptation, as well as training hybrid MLP-HMM systems. We will also study how better segmentation of words can lead to an improved open vocabulary recognition.

Acknowledgements

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 287755 (transLectures). H. Ney was partially supported by a senior chair award from DIGITEO, a French research cluster in Île-de-France.

6. References

[1] A. Žgank, V. Darinka, A. Zögling Markuš, and Z. Kačič, "BNSI Slovenian broadcast news database - speech and text corpus," in *Proc. of 9th European Conference on Speech Communication and Technology*. Lisbon, Portugal: ISCA, Sep. 2005, pp. 1537–1540.

[2] M. Bisani and H. Ney, "Multigram-based grapheme-to-phoneme conversation for LVCSR," in *European Conference on Speech Communication and Technology*, vol. 2, Geneva, Switzerland, Sep. 2003, pp. 933–936.

[3] L. Welling, S. Kanthak, and H. Ney, "Improved methods for vocal tract normalization," in *Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Phoenix, Arizona, USA, Mar. 1999, pp. 761–764.

[4] H. Hermansky, "Perceptual linear prediction (PLP) analysis for speech," *Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[5] R. Schlüter, I. Bezrukov, H. Wagner, and H. Ney, "Gammatone features and feature combination for large vocabulary speech recognition," in *Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Honululu, HI, USA, Apr. 2007.

[6] H. A. Bourlard and N. Morgan, *Connectionist Speech Recognition: A Hybrid Approach*. Norwell, MA, USA: Kluwer Academic Publishers, 1993.

[7] H. Hermansky, D. Ellis, and S. Sharma, "Tandem connectionist feature extraction for conventional HMM systems," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, vol. 3, 2000, pp. 1635–1638.

[8] F. Grézl, M. Karafiát, S. Kontár, and J. Černocký, "Probabilistic and bottle-neck features for LVCSR of meetings," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 2007, pp. 757–760.

[9] H. Hermansky and P. Fousek, "Multi-resolution RASTA filtering for TANDEM-based ASR," in *Proc. of 9th European Conference on Speech Communication and Technology*. Lisbon, Portugal: ISCA, Sep. 2005, pp. 361–364.

[10] F. Valente and H. Hermansky, "Hierarchical and parallel processing of modulation spectrum for ASR applications," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Las Vegas, USA, April 2008, pp. 4165–4168.

[11] M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, vol. 12, pp. 75–98, 1998.

[12] G. Stemmer, F. Brugnara, and D. Giuliani, "Adaptive training using simple target models," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, vol. 1, 2005, pp. 997–1000.

[13] J. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER)," in *IEEE Automatic Speech Recognition and Understanding (ASRU) Workshop*, Santa Barbara, CA, USA, Dec. 1997, pp. 347–354.

[14] M. A. Basha Shaik, A. El-Desoky Mousa, R. Schlüter, and H. Ney, "Investigation of maximum entropy hybrid language models for open vocabulary German and Polish LVCSR," in *Interspeech*, Portland, OR, USA, Sep. 2012.

[15] S. Hahn and D. Rybach, "Building an open vocabulary ASR system using open source software," Florence, Italy, Aug. 2011, Interspeech, Tutorial M3.