INTERSPEECH 2009 BRIGHTON

Cross-language Bootstrapping for Unsupervised Acoustic Model Training: Rapid Development of a Polish Speech Recognition System

Jonas Lööf, Christian Gollan, and Hermann Ney

Lehrstuhl für Informatik 6 - Computer Science Dept.
RWTH Aachen University, Aachen, Germany
{loof,gollan,ney}@cs.rwth-aachen.de

Abstract

This paper describes the rapid development of a Polish language speech recognition system. The system development was performed without access to any transcribed acoustic training data. This was achieved through the combined use of cross-language bootstrapping and confidence based unsupervised acoustic model training. A Spanish acoustic model was ported to Polish, through the use of a manually constructed phoneme mapping. This initial model was refined through iterative recognition and retraining of the untranscribed audio data.

The system was trained and evaluated on recordings from the European Parliament, and included several state-of-the-art speech recognition techniques in addition to the use of unsupervised model training. Confidence based speaker adaptive training using features space transform adaptation, as well as vocal tract length normalization and maximum likelihood linear regression, was used to refine the acoustic model. Through the combination of the different techniques, good performance was achieved on the domain of parliamentary speeches.

Index Terms: speech recognition, unsupervised training, cross-language bootstrapping

10.21437/Interspeech.2009-20

1. Introduction

As described in [1, 2], rapid development of an automatic speech recognition (ASR) system can greatly profit from the use of acoustic model *unsupervised training*, i.e. the use of untranscribed acoustic data for training by utilizing an automatic transcription of the data, generated by a previous iteration of the automatic speech recognition system being trained.

Typically unsupervised training is used to improve an available ASR system through the use of additional acoustic data. This can range from the incremental improvement of a state-of-the-art system, as in [3, 4, 5], to the substantial improvement of a small bootstrap system trained using only a few hours of transcribed audio, using hundreds of hours of untranscribed data, as in [6, 2]. For the best performance, confidence measures [7, 8, 9] derived from the recognizer output are typically used to select or weight the contribution from the acoustic training data.

The use of acoustic models from one language as a starting point for training acoustic models for a different language, *cross-language bootstrapping*, have been described in [10]. In this approach a mapping between the phonemes in the source and target language is used to enable retraining of the source language acoustic model on new transcribed training audio data in the target language. This has the advantage of requiring a significantly smaller amount of acoustic training data than would be required for conventional training from scratch.

In this paper the combination of these two techniques is presented, and used to develop a Polish ASR system *without any* transcribed audio data for acoustic model training. The paper is organized as follows: First a short overview of ASR for Polish is presented. This is followed by the presentation of the European Parliament Polish ASR task; the data is discussed, and the development of the language model is described. Then cross language bootstrapping is presented, its use to the current task is discussed, and the pronunciation lexicon and phoneme mapping are discussed. The next section describes unsupervised training and discusses its use for the current task, and how it interacts with the cross-language bootstrapping. Finally the details of the system development and the ASR system are described, and results are presented from the different stages of system development.

2. Polish Speech Recognition

Although Polish is a language spoken by a large population in the European Union, published results on large vocabulary speech recognition for the Polish language are scarce. With few exceptions publications deal either with isolated word , or small vocabulary recognition. One exception is [11], where a large vocabulary recognizer for the read speech portion of the Polish SpeeCon data base has been developed – achieving word error rates of about 6%. Another exception is the large vocabulary Polish speech recognizer for telephone dialogs that has been developed by Loquendo [12] for use in the European Union project LUNA; for this system no public results are available at this time.

The Polish language poses some specific challenges due mainly to its rich morphology [13]. Nevertheless, as a first step it is important to establish a firm baseline, including methods expected in a current state-of-the-art large vocabulary ASR system. The aim of the present work was to establish such a baseline as quickly as possible – more advanced methods, and methods specific to Polish are subject of future work.

The present work was performed in the context of the JUMAS project [14], a European Union project aimed at information extraction and indexing in the judicial domain, specifically for the processing of court recordings. As part of the project a Polish ASR system for the domain of court proceedings is being developed. Due to the limited availability of in-domain data at the start of the project, the first efforts on a Polish ASR system was performed on the domain of political (parliamentary) speeches, using raw untranscribed speech data from the European Parliament which was already available.

3. Polish European Parliament ASR

As part of the (now ended) TC-STAR project, speech recognition in English and Spanish was performed on recordings from the European Parliament, as described in [15, 16]. In the European Parliament plenary sessions (EPPS) different languages of the EU are spoken, and simultaneously interpreted into every official language of the EU. Starting at the time of the TC-STAR project and continuing since then, the recordings of the parliamentary sessions, both the original politicians, as well as the interpreter audio, have been collected. In addition, the publicly available preliminary transcriptions (only available for the original language), as well as the final minutes of meeting (available in most official languages in parallel), have been collected. Further details on the EPPS language resources can be found in [17].

Several hundreds of hours of Polish parliament recordings are currently available, of which about 6% consist of original politician speeches, and the rest consist of interpreter speech. From this a black-out period was chosen, and a half hour tuning set as well as a three hour development set were extracted, see Table 1. The recognition sets were chosen to include only the politician portions. Since (approximate) official transcriptions are available for the politician portion, it was possible to develop a preliminary version of the tuning set without input from any native Polish speakers. This preliminary version was used for system development, but the final versions of both the tuning and development sets were corrected by native Polish speakers.

Table 1: *Acoustic corpora, statistics.*

	Tune	Dev	Train
Net Duration	0.45h	3.03h	127.8h
# Segments	195	1326	40995
# Speakers	9	37	–
# Running words	2944	21938	788098
Perplexity	368	404	–
OOV Rate	3.57%	3.89%	–

Table 1 also describes the acoustic recordings used for (unsupervised) acoustic training for the current system. This data was taken from outside of the blackout period, and included both original politician speeches as well as interpreter audio. Since this data is completely untranscribed, the word statistics are taken from the automatic transcription output, described in Section 6.

While for most languages the official minutes of meetings are available, with text data originating from several years, this is not the case for Polish – the translation into Polish only recently began, and was not available at the time the system was developed. This means that the only in-domain text data available was the preliminary transcriptions of the politician portions of the acoustic data, totaling about half a million running words. Since this is clearly inadequate for language model (LM) training, several additional sources of text data were used. The additional data consisted of official translations of European Union legal documents into Polish, as well as news articles, collected over the web from two Polish news sources. See Table 2 for details on the language model training data.

The language model was a four-gram LM using modified Kneser-Ney smoothing, and the vocabulary was chosen as the approximately 60k most common words of the European Parliament portion of the LM training data. Separate models were trained for each of the four portions and combined using interpolation. Interpolation weights were tuned by optimizing the perplexity on the text of the tuning corpus. The perplexity of the resulting model, as well as the out of vocabulary (OOV) rate, are shown in Table 1.

Table 2: *Text data used for language modeling.*

Source	Running Words
European Parliament	481 k
EU Legal Documents	29,425 k
Kurier Lubelski (News)	15,364 k
Nowosci (News)	27,720 k

4. Cross-language Bootstrapping

Cross-language bootstrapping is the technique of initializing acoustic model training using a acoustic model originally trained on a different language. For the present system it was decided to use an already available Spanish European Parliament acoustic model as a starting point. As described in [10], for cross-language bootstrapping, a mapping from the target language phoneme set (in our case Polish) to the source language phonemes (Spanish) is needed. In [10], two principle ways of arriving at a phoneme mapping are described. Either a mapping is constructed manually, or it is derived automatically from data, using a target language phoneme recognizer.

Although in [10] slightly better performance is presented using an automatically derived mapping, it was decided to use a manually constructed mapping for the present system, due to the simplicity of the method. It is expected that the slight difference disappears after several iterations of retraining. The Spanish acoustic model used the Spanish SAMPA phoneme set. For Polish the Polish SAMPA phonemes consisting of 37 phonemes were used. The pronunciations for the vocabulary were generated using letter to sound rules described in [18]. Due to the properties of the different SAMPA phoneme sets, the same phoneme symbol represents similar sounds in different languages. For Polish phonemes whose SAMPA symbols are also used for the Spanish phoneme set, the mapping was chosen to simply preserve the symbol. For the remaining 14 Polish phonemes, the Spanish phoneme with the most similar properties was manually chosen, see Table 3 for the mapping used. Once a mapping is available it is possible to use the Spanish acoustic model in combination with the Polish pronunciation lexicon for acoustic model retraining, and it can even be used for recognition (although with a high error rate) in combination with the Polish language model.

5. Unsupervised Training

The basic idea of unsupervised training is to improve an acoustic model by iterated recognition and retraining on training data for which no manual transcriptions are available. For the Polish system presented here, the acoustic model being improved started out as Spanish acoustic model, but once a phoneme mapping as described in the previous section is available, the principle steps remain the same.

For effective use of available acoustic data, it is important to use confidence measures to select or weight the contributions of the audio data in such a way that correctly recognized data is more likely to contribute to the modeling. For the present work,

Table 3: *Phoneme Mapping for non-identical SAMPA Symbols.*

Polish Phoneme	Spanish Phoneme
dz	tS
dz'	tS
dZ	tS
e~	e
I	i
n'	J
o~	o
s'	s
S	x
ts	tS
ts'	tS
v	B
z'	z
Z	x

the state posterior confidence method, as presented in [5, 9] was used.

This approach works as follows: In the speech decoding process, different dynamic pruning methods are applied to restrict the list or set of competing word sequences. This list can be efficiently represented using a lattice L. With the forward-backward algorithm, it is possible to efficiently compute the relation between the competing hypotheses by estimating the lattice link posterior probabilities [8]. Depending on the lattice link labels and structure, the confidence scores for different events, e.g. word, phoneme, state or pronunciation confidence scores can be computed.

Here, a lattice link $l = [w, s_1^{\tau-t}; \tau, t]$ represents the hypothesized word w with its start time t and end time τ as well as the corresponding state alignment $s_1^{\tau-t}$. The function $s(l_1^N, t)$ defines the state s of a lattice path l_1^N at time-frame t. Then, the first-best weight confidence score $C(t, \hat{s}_1^T, L)$ defines the time-frame confidence scores of the firs-best state hypothesis $\hat{s}_1^T$ and can be expressed as follows:

$$p(l_1^N|x_1^T) := \frac{p(x_1^T, l_1^N)}{\sum_{l'^{N'}_1 \in L} p(x_1^T, l'^{N'}_1)} \quad (1)$$

$$C(t, \hat{s}_1^T, L) := \sum_{\substack{l_1^N \in L: \\ \hat{s}_t = s(l_1^N, t)}} p(l_1^N|x_1^T) \quad (2)$$

Unsupervised acoustic model training – more precisely, iterative acoustic emission model re-estimation – is performed using confidence-thresholded automatic transcriptions. For Gaussian mixture training, the data filtering process is done on state/frame level to select the pairs of a state and a acoustic feature vector based on their confidence score. From previous work [5], this selection or filtering process is known to be more precise than performing the thresholding on sentence or word level.

Depending on the purpose confidence scores are estimated for other events. For the Gaussian mixture training or the LDA estimation the data selection is based on the confidence scores of tied HMM states, whereas for the estimation of the state tying, the thresholding is based on the allophone state confidence scores of the observations.

The word lattices needed to produce the confidence measures for each iteration of retraining, are produced using the acoustic model from the previous iteration. For the first two re-estimation iterations maximum a posteriori (MAP) adaptation was used instead of conventional acoustic model training. Generally MAP adaptation in an unsupervised framework performs rather poorly, but as was shown in [9], in combination with confidence measures good results can be achieved.

6. Experiments

On the original (Spanish) task, the Spanish bootstrap model achieves an error rate of approximately 10%. Using cross-language bootstrapping with no retraining, the error rate is obviously much higher, initially around 60% Several iterations of retraining are necessary to achieve adequate performance; the following paragraphs describe the procedure used. In Table 4 the recognition performance after the different retraining steps, as well as the amount of data selected by confidence thresholding, are summarized.

The bootstrap model was trained on vocal tract length normalized (VTLN) mel-frequency cepstral coefficient (MFCC) features, with the warping factors estimated using a Gaussian mixture model (GMM) classifier. The feature extraction front-end is completed by cepstral mean normalization and linear discriminant analysis (LDA) over a window of seven consecutive frames, resulting in a 45 dimensional feature vector.

The system uses a classification and regression tree (CART) state tying, grouping the possible triphones into 4500 generalized triphone states. The acoustic models used in the system consist of hidden Markov models with Gaussian mixture model (GMM) based emission probabilities. The GMMs use a single pooled variance vector, and a fully trained model consist of approximately 900k distributions in total.

The unsupervised retraining of the acoustic model was performed on about 130 hours of untranscribed recordings from the European Parliament (see Table 1; first iterations used reduced data sets) The first step needed for the unsupervised training is the segmentation of the training corpus, which consisted of uncut raw recordings, containing in addition to the actual speeches some music and other non-speech sections. For this purpose, the NIST tools acoustic segmentation software by CMU was used [19].

Since the starting point error rate was quite high, and also to achieve results faster, the first two iterations of acoustic model retraining was performed using MAP adaptation using confidence measures. Each of the iterations consist of one recognition of the training corpus, producing word lattices, followed by one MAP re-estimation of the acoustic model.

After yet another recognition producing lattices, the next retraining step consisted of confidence measures based estimation of state tying and LDA matrix. This was followed by several iterations of recognition and retraining. The final iterations were made using speaker adaptive training (SAT) with feature space maximum likelihood linear regression (fMLLR). The fMLLR matrices for SAT were estimated using confidence measures both in training and in recognition. The SAT was performed on automatically generated segment clusters, estimated using generalized likelihood ratio based clustering with a Bayesian information based stopping criterion.

As a final improvement, maximum likelihood linear regression (MLLR) adaptation was used in recognition. The final results for the two pass system on the tuning and development sets are presented in Table 5.

Table 4: *Performance in development – Tuning set.*

Training step	WER [%]	Data sel. [h]
Initial Spanish AM	63.4	–
First MAP iter.	49.6	1.7
Second MAP iter.	37.1	29.8
First training iter.	29.9	59.2
Second training iter.	26.9	53.9
First SAT iter.	24.1	66.1
First full data train.	23.2	103.4
SAT full data	20.7	106.0
SAT re-training	20.5	113.7
SAT re-training	20.0	111.0

Table 5: *Performance of final system – WER [%].*

System	EPPS Tune	EPPS Dev
1st Pass	21.8	21.2
+ SAT	18.1	18.5
+ MLLR	17.5	18.0

7. Summary

In this paper the fast development of a Polish speech recognition system has been described. The acoustic model of the system was trained in a completely unsupervised way, without using any transcribed Polish acoustic training data. This was made possible through the use of cross language bootstrapping, starting from a Spanish acoustic model.

The resulting system achieved performance well within what can be expected from a well tuned system for the task at hand, considering that no transcribed audio data was used, and also the limited amount of in-domain language model training data used.

8. Acknowledgements

This work was partly funded by the European Union under the FP6 project JUMAS, Contract No. 214306.

9. References

[1] Zavaliagkos, G. and Colthurst, T., "Utilizing untranscribed training data to improve performance," in *DARPA Broadcast News Transcription and Understanding Workshop*, Landsdowne, VA, USA, Feb. 1998, pp. 301 – 305.

[2] Gollan, C. and Ney, H., "Towards automatic learning in LVCSR: Rapid development of a Persian transcription system," in *Proc. Int. Conf. on Spoken Language Processing*, Sep. 2008, pp. 1441 – 1444.

[3] Ma, J., Matsoukas, S., Kimball, O., and Schwartz, R., "Unsupervised training on large amounts of broadcast news data," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, vol. 3, Toulouse, France, May 2006, pp. 1057 – 1059.

[4] Ramabhadran, B., "Exploiting large quantities of spontaneous speech for unsupervised training of acoustic models," in *Proc. Int. Conf. on Spoken Language Processing*, Lisbon, Portugal, 2005, pp. 1617–1620.

[5] Gollan, C., Hahn, S., Schlüter, R., and Ney, H., "An improved method for unsupervised training of LVCSR systems," in *Interspeech*, Antwerp, Belgium, Aug. 2007, pp. 2101–2104.

[6] Wessel, F. and Ney, H., "Unsupervised training of acoustic models for large vocabulary continuous speech recognition," *IEEE Trans. on Speech and Audio Processing*, vol. 13, no. 1, pp. 23 – 31, Jan. 2005.

[7] Wessel, F., Macherey, K., and Ney, H., "A comparison of word graph and N-best list based confidence measures," in *European Conference on Speech Communication and Technology*, Budapest, Hungary, Sep. 1999, pp. 315–318.

[8] Evermann, G. and Woodland, P., "Large vocabulary decoding and confidence estimation using word posterior probabilities," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Istanbul, Turkey, Jun. 2000, pp. 1655 – 1658.

[9] Gollan, C. and Bacchiani, M., "Confidence scores for acoustic model adaptation," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Las Vegas, NV, USA, Mar. 2008, pp. 4289 – 4292.

[10] Schultz, T. and Waibel, A., "Experiments on cross-language acoustic modeling," in *Proc. European Conf. on Speech Communication and Technology*, Aalborg, Denmark, Sep. 2001, pp. 2721 – 2724.

[11] Marasek, K., "Polish LVCSR in the Janus system. Preliminary results for the SpeeCon database," *Archives of Acoustics*, vol. 32, no. 1, pp. 119 – 126, 2007.

[12] Loquendo, "New Language for Loquendo ASR: Polish," 2006 http://www.loquendo.com/en/news/news_loquendo_ASR_polish.htm.

[13] Demenko, G., Grocholewski, S., Klessa, K., Ogórkiewicz, J., Wagner, A., Lange, M., Śledziński, D., and Cylwik, N., "JURISDIC – Polish speech database for taking dictation of legal texts," Atlanta, GA, USA, May 2008, pp. 1280 – 1287.

[14] "JUMAS, Judicial Management by Digital Libraries Semantics," http://www.jumasproject.eu.

[15] Lööf, J., Bisani, M., Gollan, C., Heigold, G., Hoffmeister, B., Plahl, C., Schlüter, R., and Ney, H., "The 2006 RWTH parliamentary speeches transcription system," in *Proc. Int. Conf. on Spoken Language Processing*, Pittsburgh, PA, USA, Sep. 2006, pp. 105 – 108.

[16] Lööf, J., Gollan, C., Hahn, S., Heigold, G., Hoffmeister, B., Plahl, C., Rybach, D., Schlüter, R., and Ney, H., "The RWTH 2007 TC-STAR evaluation system for European English and Spanish," in *Proc. Int. Conf. on Spoken Language Processing*, Antwerp, Belgium, Aug. 2007, pp. 2145 – 2148.

[17] Gollan, C., Bisani, M., Kanthak, S., Schlüter, R., and Ney, H., "Cross domain automatic transcription on the TC-STAR EPPS corpus," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, vol. 1, Philadelphia, PA, Mar. 2005, pp. 825–828.

[18] Oliver, D., "Polish text to speech synthesis," Master's thesis, Edinburgh University, Edinburgh, UK, 1998.

[19] Siegler, M. A., Jain, U., Raj, B., and Stern, R. M., "Automatic segmentation, classification and clustering of broadcast news audio," in *Proc. DARPA Speech Recognition Workshop*, 1997, pp. 97–99.