Towards Automatic Learning in LVCSR: Rapid Development of a Persian Broadcast Transcription System

Christian Gollan and Hermann Ney

Human Language Technology and Pattern Recognition – Computer Science Department
RWTH Aachen University, D-52056 Aachen, Germany
{gollan,ney}@cs.rwth-aachen.de

Abstract

We present a new method for automatic learning and refining of pronunciations for large vocabulary continuous speech recognition which starts from a small amount of transcribed data and uses automatic transcription techniques for additional untranscribed speech data.

The recognition performance of speech recognition systems usually depends on the available amount and quality of the transcribed training data. The creation of such data is a costly and tedious process and the approach presented here allows training with small amounts of annotated data.

The model parameters of a statistical joint-multigram grapheme-to-phoneme converter are iteratively estimated using small amounts of manual and relatively larger amounts of automatic transcriptions and thus the system improves itself in an unsupervised manner.

Using the new approach, we create a Persian broadcast transcription system from less than five hours of transcribed speech and 52 hours of untranscribed audio data.

Index Terms: Automatic Learning, Unsupervised Training, Dictionary Learning, Automatic Transcription

1. Introduction

The recognition performance of a state-of-the-art large vocabulary continuous speech recognition (LVCSR) system depends on the amount and quality of task representative training data. Huge collections of language resources are necessary for robust estimation of the model parameters. The current version of the Arabic broadcast transcription system – described in [1] – is trained on more than 1 000 hours of speech data and more than 1 billion running words of text data. The data collection process is still ongoing to improve the systems transcription performance by increasing the amount of training data.

Often, only a relatively small amount of annotated speech data is available when an LVCSR system is build for a new language. In general, whenever a new system is designed, task representative speech data needs to be collected and manually transcribed which is a time-consuming and costly process. Commonly, an annotated speech corpus and a pronunciation dictionary are used for acoustic model training. For language model training, large text collections are used – most often online newspaper archives. The creation of the pronunciation dictionary and manual transcription of speech are by far the most expensive steps in setting up a new ASR system.

Using the automatic transcriptions of speech data to train an ASR system is commonly referred to as *unsupervised training* [2]. Automatic annotations are not only cheaper than manual annotations but can also be obtained much faster. On the downside, automatic annotations commonly contain more transcription errors than manually transcribed ones. Especially when the ASR system itself is trained only on a small amount of transcribed speech data. To avoid problems due to this drawback, confidence scores can be applied to remove the most likely errors from the automatic annotations. Here, we use the term *automatic learning*, to describe the ability of an ASR system to improve its parameters in general without a manually annotated speech corpus. Ideally, such an automatic transcription system trains and improves its model parameters without any human setup efforts. A broadcast news transcription system could improve its performance online by collecting broadcast audio streams and internet text news to automatically re-estimate its model parameters.

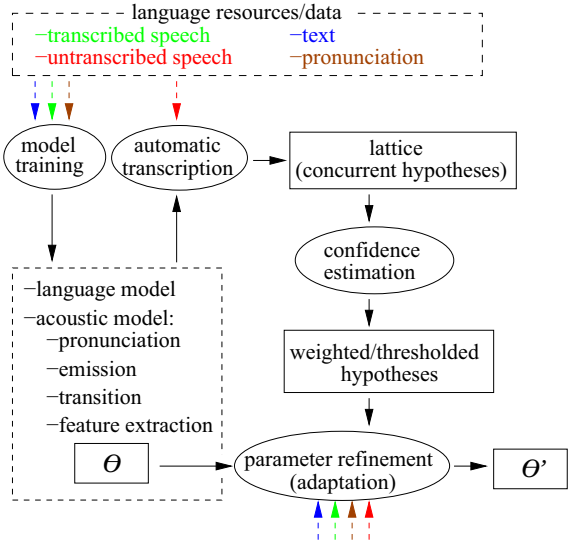

Figure 1: *Initial setup in the Automatic Learning framework: First iteration of unsupervised/automatic parameter refinement.*

Figure 1 shows an iteration of the automatic learning concept for an ASR system. Well known as acoustic model adaptation is the iterative improvement of the emission model parameters using the previous automatic transcriptions of the recognition task. If we use large amounts of additional speech data, we call the process unsupervised training. Depending on the amount and the quality of the automatic transcriptions, different acoustic refinement methods are used, e.g., linear transform based adaption, maximum a posteriori adaptation or full Baum-Welch re-estimation.

In the automatic learning framework also the parameters of the language or the acoustic feature generation model can be iteratively improved. E.g. the parameters of a multi-layer perceptron (MLP) network can be refined using additional automatically transcribed data.

Accepted after peer review of full paper

Table 1: *Persian words with automatically refined pronunciations.*

word rank	word	ASCII transliteration	first 2-best pronunciations	pron. probab.
1	و	w	v A o	0.65 0.32
19	ايران	AyrAn	Q i r a n	1.0
90	استفاده	AstfAdh	s d f a d e A f a d e	0.92 0.05
166	اجتماعي	AjtmAEy	A S t A m o Q i A s t A m o Q i	0.91 0.07
7655	اوباما	AwbAmA	Q o b a m a Q o b A m m a	0.76 0.09

In this paper, we describe the rapid development of an LVCSR system for Persian broadcast news where we apply the confidence based parameter refinement method for the emission model and the pronunciation model training. In [3] a similar experiment was presented to set up an American-English broadcast transcription system. The training data was artificially reduced to perform unsupervised acoustic model training as a proof of concept. In [4] we have presented the unsupervised emission model training on state/frame level. In this work, we present the automatic learning framework and we propose a novel method for automatic learning and refinement of a pronunciation dictionary. A related approach for dictionary learning is presented in [5], where the authors use an annotated speech corpus and a phoneme language model.

2. Task and System Description

2.1. Persian Language Details

In a few countries Persian is spoken as official language, i.e. Iran, Afghanistan and Tajikistan. The language is often referred to as Farsi which is the local name. Further languages are spoken in Iran, e.g. Luri and Bakhtiari. These languages are related to Persian but they have different words and grammar. Modern Iranian Persian is the official language in Iran and is therefore the main language in Iranian broadcasts and newspapers.

As in Arabic, short vowels are generally omitted in Persian documents and only the long vowels are written in the text. In Modern Iranian Persian, a modified variant of the Arabic alphabet is used for writing, i.e., some of the Arabic letters are modified and a few additional letters are used.

Table 1 shows frequent Persian words from our language model text data and their pronunciations. The appearance of Persian characters depends on their context and position within a word and for an untrained person it can be difficult to read them. Therefore, the ASCII transliteration is often used to represent Arabic or Persian text by the latin alphabet.

2.2. Persian Language Resources

Usually, huge data collections – in particular annotated speech corpora – are used to build LVCSR systems. Unfortunately, large collections are not available for all languages. For Persian, there are only a few annotated data resources publicly available.

Acoustic data: To train our initial acoustic model we use the the Farsdat speech corpus [6]. This database contains less than 5 hours of annotated read speech of Persian sentences read by 304 speakers. Clearly, this corpus is not the perfect fit to train acoustic models for a broadcast transcription system.

We collected audio recordings – representing our transcription task – from the news channel of the Islamic Republic of Iran News Network (IRINN). The first broadcast recording was collected in February 2008. It was annotated by Persian transcribers to be used as development and test set for our system development process. Since then, we have continuously collected more additional audio data from the IRINN satellite broadcasts. The data was decoded from MPEG-1 layer-2 audio and sampled down to 16kHz for feature extraction.

Table 2: *Text corpora statistics.*

	# running words	#uniq words	50k voc. OOV [%]	ppl.
Hamshahri corpus [7] Jun. 1996–Feb. 2003	64m	531k	2.7	141.1
Hamshahri web text Jun. 2006–Mar. 2008	26m	222k	2.1	114.5
IRIB web text Oct. 2005–Mar. 2008	30m	240k	3.1	45.0
IRINN web text Aug. 2006–Mar. 2008	10m	134k	2.5	61.6

Text data: For language modeling, we use different news text archives as presented in Table 2. In our first system setup we used the already preprocessed Hamshahri text corpus [7] compiled from the Hamshahri Iranian newspaper online archive. The Hamshahri corpus covers news articles from June 1996 until February 2003. In addition we download Hamshahri web texts as they are being made available. Furthermore, we collected news texts from the Islamic Republic of Iran Broadcasting (IRIB) web site and from the IRINN web site. Persian is a less morphologically complex language than Arabic. We estimate low out-of-vocabulary (OOV) rates using the top 50k words of the complete text compilation.

2.3. System Description

The acoustic model of our initial transcription system was trained on a relatively small amount of read speech. We used the manually produced phoneme alignment of the Farsdat corpus to extract a pronunciation dictionary. This pronunciation dictionary was used to train a data driven statistical grapheme-to-phoneme (G2P) model [8] to automatically generate pronunciations for the 50k recognition lexicon. Table 5 presents the statistics of the Farsdat pronunciation dictionary and the final recognition lexicon. We generate N-best pronunciation lists using the G2P approach to produce the pronunciation alternatives which are mostly caused by the short vowels.

In the following, we present the system details of the latest system we obtained due to iterative unsupervised training of the emission and pronunciation model parameters.

System details:

- Speaker and domain independent acoustic model
- Speaker adaptation: VTLN + SAT-CMLLR + MLLR
- 45 dimensional acoustic vectors after applying LDA on 9 time consecutive stacked 17 dimensional features (MFCC and voicing feature)
- 3-state left-to-right HMM topology
- 29 phonemes + 2 noise models + silence state
- 4 038 decision tree tied within-word triphone states
- 129k densities within Gaussian mixture models with globally pooled diagonal covariance
- Maximum likelihood training using Viterbi approximation
- 50k recognition vocabulary
- 4-gram language model with 28 million M-grams

The main difference to our initial system is the number of emission model parameters. Due to the relatively small Farsdat speech corpus, we estimated 3k densities for 1 517 states.

3. Automatic Learning

In the automatic learning framework the model parameters of a speech recognition system are iteratively improved for the long term using additional task representative data. Ideally, this data is collected during the application of the system.

3.1. Automatic Transcription

In the recognition/speech-decoding process the most likely word sequence $\hat{w}_1^N$ is determined by Eq. 1 using the Bayes decision rule. The decision depends on the language model probability $p(w_1^N)$ and the acoustic probability $p(x_1^T|w_1^N)$. x_1^T is the acoustic feature vector sequence. The true language probability is approximated by an M-gram language model (LM), see Eq. 2. We approximate and decompose the true acoustic probability $p(x_1^T|w_1^N)$ into the pronunciation probability $p_{\theta_{\text{pm}}}(v_1^N|w_1^N)$, the emission probability $p_{\theta_{\text{em}}}(x_t|s_t, v_1^N)$, and the transition probability $p_{\theta_{\text{tm}}}(s_t|s_{t-1}, v_1^N)$, as described in Eq. 4 and Eq. 5, where v_1^N denotes a pronunciation sequence and s_t is an HMM state at time t.

The pronunciation scale α and the acoustic scale β are used to cope with the model approximations and with different amounts of used training data to model the true probability distributions.

$$\hat{w}_1^N = \underset{w_1^N}{\operatorname{argmax}} \left\{ p(x_1^T|w_1^N) \cdot p(w_1^N) \right\} \tag{1}$$

$$p(w_1^N) := \prod_{n=1}^{N} p_{\theta_{\text{lm}}}(w_n|w_{n-M+1}^{n-1}) \tag{2}$$

$$p(x_1^T|w_1^N) := \max_{v_1^N} \left\{ p_{\theta_{\text{pm}}}(v_1^N|w_1^N)^{\alpha} p_{\theta_{\text{em},tm}}(x_1^T|v_1^N)^{\beta} \right\} \tag{3}$$

$$p_{\theta_{\text{pm}}}(v_1^N|w_1^N) := \prod_{n=1}^{N} p_{\theta_{\text{pm}}}(v_n|w_n) \tag{4}$$

$$p_{\theta_{\text{em},tm}}(x_1^T|v_1^N) := \max_{s_1^T|v_1^N} \prod_{t=1}^{T} p_{\theta_{\text{em}}}(x_t|s_t) p_{\theta_{\text{tm}}}(s_t|s_{t-1}) \tag{5}$$

3.2. Confidence Scores

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, we efficiently compute the relation between the competing hypotheses by estimating the lattice link posterior probabilities [9]. Depending on the lattice link labels and structure we can compute the confidence scores for different events, e.g. word, phoneme, state or pronunciation confidence scores.

A pronunciation lattice link $[w, v; \tau, t] \in L$ consists of a word w, a pronunciation v, a start time τ and end time t. The lattice link posterior probability is denoted as $p([w, v; \tau, t]|x_1^T)$. We now can calculate the pronunciation confidence score $C(v, n; L, \hat{w}_1^N)$ for our first best word sequence $\hat{w}_1^N$ using the maximum approximation analogously to the maximum word confidence scores as presented in [10]:

$$C(v, n; L, \hat{w}_1^N) := \max_{\substack{\hat{t}: \\ \tau_{\hat{w}_n} \le \hat{t} \le t_{\hat{w}_n}}} \sum_{\substack{[\hat{w}_n, v; \tau', t']: \\ \tau' \le \hat{t} \le t'}} p([\hat{w}_n, v; \tau', t']|x_1^T) \tag{6}$$

3.3. Pronunciation Model Refinement

We are using a statistical data driven pronunciation model derived from the G2P conversion method presented in [8]. There, the joint probability distribution $p(w, v)$ is reduced to the probability distribution $p(q_1^J)$ modeled by a standard M-gram:

$$p(q_1^J) := \prod_{j=1}^{J} p(q_j|q_{j-M+1}^{j-1}) \tag{7}$$

A graphone sequence q_1^J corresponds to a word pronunciation pair (w, v). In our experiments a graphone q is a pair of a word-symbol/grapheme and pronunciation-symbol/phoneme, where a grapheme or phoneme can be the empty grapheme or phoneme respectively. $S(w, v)$ defines the set of all possible graphone sequences to segment the pair (w, v) into graphones. We calculate the pronunciation probability $p_{\theta_{\text{G2P}}}(v|w)$ using the G2P model the following way:

$$p_{\theta_{\text{G2P}}}(v|w) := \frac{\max_{q_1^J \in S(w,v)} p(q_1^J)}{\sum_{v'} \max_{q'^{J'}_1 \in S(w,v')} p(q'^{J'}_1)} \tag{8}$$

The pronunciation counts $\#(v, w)$ and the word counts $\#(w)$ are estimated from the first best transcription alignment path and represent how often these events were observed. We refine the pronunciation probabilities in a maximum a posteriori fashion based on the normalized pronunciation counts and the balancing parameter λ:

$$p_{\theta'_{\text{pm}}}(v|w) := \frac{\#(v, w)}{\lambda + \#(w)} + \frac{\lambda}{\lambda + \#(w)} p_{\theta_{\text{pm}}}(v|w) \tag{9}$$

For the automatic transcriptions, we apply the pronunciation confidence score to estimate only counts for events with high confidence. This is done for the first best hypotheses $\hat{w}_1^N$ of the automatic transcriptions:

$$\#(v, w)_{\hat{w}_1^N} := \sum_{\substack{n: \\ C(v,n;L,\hat{w}_1^N) > thr.}}^{N} \delta(w, \hat{w}_n), \qquad \#(w) := \sum_{v} \#(v, w) \tag{10}$$

3.4. Emission Model Re-estimation

We perform unsupervised acoustic model training – more precisely, iterative acoustic emission model re-estimation – 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. This selection or filtering process is more precise than performing the thresholding on sentence or word level and was successfully applied for unsupervised training [4], as well as for unsupervised acoustic model adaptation [11].

Depending on the purpose we estimate confidence scores for other events. We perform the data selection for the Gaussian mixture training or the LDA estimation based on the confidence scores of tied HMM states. Whereas for the estimation of the state tying, we threshold the observations based on their allophone state confidence scores.

4. Experiments

Table 3 summarizes the rapid development of the Persian broadcast transcription system which was started in February 2008. The initial acoustic model was trained on the Farsdat speech corpus resulting in a WER of 75.8% on the IRINN development set. We use the manually produced transcriptions of the IRINN

Table 3: *Training steps with the corresponding emission model statistics and resulting system performance on the dev set.*

training step	audio data [h]	selected data [h]	#states	#densities	dev set WER [%]
Farsdat	4.7	4.7	1 517	3k	75.8
1. θ_{em}	11.4	8.1	1 517	6k	72.7
2. θ_{em}	11.4	8.6	1 517	12k	69.8
3. θ_{em}	23.6	15.8	1 517	48k	63.7
4. $\theta_{\mathrm{em}}/\theta_{\mathrm{lm}},\theta_{\mathrm{pm}}$	23.6	16.6	2 281	72k	61.1/57.3
5. θ_{em}	39.0	29.1	2 281	144k	51.7
6. θ_{em}	39.0	29.1	3 322	205k	51.4
7. $\theta_{\mathrm{em}}/\theta_{\mathrm{pm}}$	39.0	30.2	3 322	207k	50.7/50.2
8. θ_{em}	52.0	39.6	3 322	209k	48.0
9. θ_{em}	52.0	39.1	4 038	247k	46.5
10. $\theta_{\mathrm{em}}/\theta_{\mathrm{pm}}$	52.0	41.4	4 038	249k	45.9/45.1
11. $\theta_{\mathrm{em}}/\theta_{\mathrm{pm}}$	52.0	41.0	4 038	129k	44.1/43.4
12. θ_{em}	52.0	40.0	4 038	129k	42.3

Table 4: *Speech corpora statistics.*

	Farsdat	AT	dev	test
data [h]	4.7	52.0	0.3	2.5
# running words	41k	390k	3k	18k
# segments	6k	19k	116	890
# speaker cluster	304	1 111	11	36
OOV [%]	2.0	–	2.6	3.4
ppl.	225.0	156.9	141.1	201.5

broadcast recordings from February 2008 as evaluation sets. For unsupervised training, we have collected further IRINN recordings since March 2008 and the automatic transcription (AT) data set increased over time up to 52 hours. Table 4 gives an overview of the currently used speech corpora.

The initial G2P model was trained on the Farsdat pronunciation dictionary. With this initial G2P model, we generated the 4-best pronunciations for the words of our 50k recognition vocabulary. In the later unsupervised training iterations we estimated the pronunciation counts on the manual and the automatic transcriptions to refine the pronunciation probabilities of the training and the recognition lexicon. Furthermore, we re-estimated the G2P model based on the weighted pronunciation counts. Then, we updated the pronunciation entries of the recognition lexicon based on the 4-best pronunciations generated by the updated G2P model.

The improvements due to the model updates are presented in Table 3. Until the 3rd training step, we re-estimated the Gaussian mixture model using additional automatic transcriptions which allowed us to increase the number of densities and the WER dropped to 63.7%. In the 4th training step we have re-estimated the CART tying as well as the LDA matrix leading to a WER of 61.1%. Furthermore, using the same emission model in the 4th training step with an improved LM and a refined PM, we measured the WER of 57.3%. Currently, we are in the 12th training step leading to a WER of 42.3% with a speaker independent acoustic model. Table 6 presents the results of our 3-pass Persian transcription system.

As can be seen, most of the gain is due to the increased parameter set but the estimation of these is only robust using the additional automatic transcribed training data.

5. Conclusion and Outlook

We presented the rapid development of a Persian broadcast transcription system using a relatively small speech database. Within 6 weeks, we set up an LVCSR system using less than 5 hours of manual transcriptions and a manually created pronunciation dictionary covering only 1 000 words. We have successfully applied the automatic learning framework to re-estimate the emission model parameters and the pronunciation model parameters in an unsupervised training fashion.

Table 5: *Dictionary statistics.*

	Farsdat	recog.
# words	1 017	49 939
# pronunciations	4 853	190 168
# homophones	226	12 861

Table 6: *Results WER[%].*

	dev	test
spk. indep.	42.3	47.9
SAT-CMLLR	38.5	41.9
MLLR	37.5	39.9

In future work we will integrate acoustic MLP features into our transcription system. The MLP network can then also be iteratively re-estimated based on the automatic transcriptions and their phoneme confidence scores. Furthermore, we will try to perform language model adaptation using large amounts of automatic transcriptions to improve the systems performance for the long term.

Our future plans include the investigation of the amount of transcribed training data required to set up an ASR system. Ideally, it would be possible to bootstrap systems without any transcribed speech data at all.

Acknowledgement. This work was partly supported by Applications Technology, Inc. (AppTek).

6. References

[1] D. Rybach, S. Hahn, C. Gollan, R. Schlüter, and H. Ney, "Advances in Arabic Broadcast News Transcription at RWTH," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop*, Kyoto, Japan, Dec. 2007, pp. 449–454.

[2] G. Zavaliagkos and T. Colthurst, "Utilizing Untranscribed Training Data to Improve Performance," in *DARPA Broadcast News Transcription and Understanding Workshop*, Landsdowne, VA, USA, Feb. 1998, pp. 301 – 305.

[3] F. Wessel and H. Ney, "Unsupervised Training of Acoustic Models for Large Vocabulary Continuous Speech Recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 1, pp. 23 – 31, Jan. 2005.

[4] C. Gollan, S. Hahn, R. Schlüter, and H. Ney, "An Improved Method for Unsupervised Training of LVCSR Systems," in *Proc. European Conf. on Speech Communication and Technology*, Antwerp, Belgium, Aug. 2007, pp. 2101–2104.

[5] T. Sloboda and A. Waibel, "Dictionary Learning for Spontaneous Speech Recognition," in *Proc. Int. Conf. on Spoken Language Processing*, Philadelphia, PA, USA, Oct. 1996, pp. 2328–2331.

[6] M. Bijankhan, J. Sheikhzadegan, M. Roohani, Y. Samareh, K. Lucass, and M. Tabiani, "FARSDAT-The Speech Database of Farsi Spoken Language," in *Fifth Australian International Conference on Speech Science and Technology (SST-94)*, Perth, Australia, Dec. 1994, pp. 826–831.

[7] F. Oroumchian, E. Darrudi, and M. Hejazi, "Assessment of a Modern Farsi Corpus," in *2nd Workshop on Information Technology and its Disciplines (WITID)*, Kish Island, Iran, Feb. 2004.

[8] M. Bisani and H. Ney, "Investigations on Joint-Multigram Models for Grapheme-to-Phoneme Conversion," in *Proc. Int. Conf. on Spoken Language Processing*, Denver, CO, USA, Sep. 2002, pp. 105–108.

[9] G. Evermann and P. Woodland, "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.

[10] F. Wessel, K. Macherey, and H. Ney, "A Comparison of Word Graph and N-Best List Based Confidence Measures," in *Proc. European Conf. on Speech Communication and Technology*, Budapest, Hungary, Sep. 1999, pp. 315–318.

[11] C. Gollan and M. Bacchiani, "Confidence Scores for Acoustic Model Adaptation," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Las Vegas, NV, USA, Apr. 2008, pp. 4289–4292.