Improving LVCSR with Hidden Conditional Random Fields for Grapheme-to-Phoneme Conversion

Stefan Hahn[1,3]*,Patrick Lehnen*[1]*, Simon Wiesler*[1]*, Ralf Schlüter*[1]*, Hermann Ney*[1,2]

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

Abstract

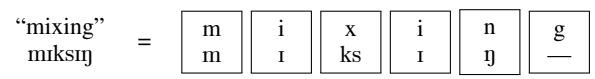
In virtually every state-of-the-art large vocabulary continuous speech recognition (LVCSR) system, grapheme-to-phoneme (G2P) conversion is applied to generalize beyond a fixed set of words given by a background lexicon. The overall performance of the G2P system has a strong effect on the recognition quality. Typically, generative models based on joint-n-grams are used, although some discriminative models have a competitive performance but the training time may be quite large.
In this work, the effect of using discriminative G2P modeling based on hidden conditional random fields (HCRFs) is analyzed. Besides measuring and comparing the G2P qualities on a textual level, one focus is the performance of LVCSR systems. Although the HCRF model does not outperform the generative one on text data, we could improve our English QUAERO ASR system by 1-3% relative on a couple of test corpora over a strong baseline by only replacing the G2P strategy.
Index Terms: grapheme-to-phoneme conversion, G2P, LVCSR, HCRF, hidden conditional random fields

1. Introduction

Grapheme-to-Phoneme conversion (G2P) is an important part of virtually every state-of-the-art large vocabulary continuous speech recognition (LVCSR) system. This task is usually defined as finding the most likely pronunciation, denoted as phoneme sequence $\boldsymbol{\varphi}$, given an orthographic form of a word, denoted as grapheme sequence $\boldsymbol{g}$. A grapheme $g \in \boldsymbol{g}$ is defined as a symbol used for writing language (e.g. a letter) and a phoneme $\varphi \in \boldsymbol{\varphi}$ as the smallest contrastive unit in the sound system of a language. The resulting optimization problem is then given as follows:

$$\boldsymbol{\varphi}(\boldsymbol{g}) = \underset{\boldsymbol{\varphi}' \in \Phi^*}{\operatorname{argmax}}\, p(\boldsymbol{g}, \boldsymbol{\varphi}')$$

Here, Φ^* denotes the set of all possible phoneme sequences. A number of different methods have been proposed over the years to tackle this task. On the one side, there are generative approaches like the ones based on joint-n-grams or graphones as introduced in [1]. Here, graphones are defined as joint or aligned units of graphemes and phonemes, on which classical n-gram models are trained, e.g.:



The authors of [2, 3, 4, 5] build upon these units, whereas the details for alignment of graphemes and phonemes, training and decoding differ. The last two methods are available as open source tools [6, 7]. A comparison of generative models for G2P is presented in [8].

On the other side, there are discriminative approaches, which have been proposed rather recently, e.g. online discriminative training [9], which is also available as an open source tool, or methods based on conditional random fields (CRFs) [10, 11]. While discriminative models usually lead to very good results, the training might be quite demanding w.r.t. computational time and memory consumption. Since typical (generative) G2P systems usually already have a very good performance ($< 10\%$ phoneme error rate), the effort of using discriminative models is usually not spent; at least not for larger tasks. Additionally, these G2P models are usually only evaluated on a textual level and thus without ASR experiments. In this paper, we investigate the effect of using discriminatively trained G2P models based on HCRFs instead of generative joint-n-gram models for English LVCSR experiments as well as their combination. We have chosen this method, since CRFs lead to good results on NLU and G2P tasks (see e.g. [10, 12, 11]). Additionally, we analyze the effect of varying the number of pronunciation variants per word as well as the pronunciation scores on speech recognition performance. This work directly builds upon our HCRF publication [12] where the focus was on evaluating HCRFs on text data only. Now, we want to analyze the effect on LVCSR performance as well as compare this method with the often used generative joint-n-gram approach on state-of-the-art tasks.

The remainder of the paper is structured as follows: The following section introduces the theoretical background of the two G2P methods, whereas in Sec. 3 we present the experimental setup which is used for the experiments presented in Sec. 4. Our findings are summarized in Sec. 5.

2. G2P Methods

As already presented in Sec. 1, there exist a number of methods to tackle the G2P task. For our experimental comparison, we have chosen a generative and a discriminative approach which are presented in this section. The joint-n-gram approach is available as an open-source toolkit, whereas the HCRF software is an in-house realization.

2.1. Generative Approach: Joint-n-Gram Model (Seq)

Models based on joint-n-grams usually rely on graphone sequences $\boldsymbol{q}$, which are defined as aligned units of graphemes and phonemes, resulting in the following probability decom-

[3]Stefan Hahn is now with Nuance Communications, Inc.,
e-mail: stefan.hahn@nuance.com

position:

$$Pr(\boldsymbol{g}, \boldsymbol{\varphi}) = Pr(\boldsymbol{q}) = \prod_{i=1}^{N} Pr(q_i | q_{i-1}, \ldots, q_1)$$

The resulting n-gram model as proposed in [4] is defined as

$$\begin{aligned} p(\boldsymbol{g}, \boldsymbol{\varphi}) &= \sum_{\boldsymbol{q} \in S(\boldsymbol{g}, \boldsymbol{\varphi})} p(\boldsymbol{q}) \\ &= \sum_{\boldsymbol{q} \in S(\boldsymbol{g}, \boldsymbol{\varphi})} \prod_{i=1}^{|\boldsymbol{q}|} p(q_i | q_{i-1}, \ldots, q_{i-M+1}) \end{aligned}$$

Here, $S(\boldsymbol{g}, \boldsymbol{\varphi})$ denotes the set of all co-segmentations of $\boldsymbol{g}$ and $\boldsymbol{\varphi}$ whereas M denotes the LM model order. Note that we use $Pr(\cdot)$ to indicate true probabilities while $p(\cdot)$ indicates model assumptions. Training of the model is performed using maximum likelihood EM training. For decoding, a maximum approximation is applied for the possibly non-unique segmentation into graphones.

The joint-n-gram model has been trained using the open-source toolkit Sequitur [6]. There are basically two parameters which control the quality of the model: a length restriction on the graphones (graphemes and phonemes can be restricted separately) and the n-gram order. For the graphones, we use the setting which has been reported to work best for English tasks, namely to allow the use of graphones of length one, whereas the grapheme or phoneme may be empty. The performance on the development set converges at $M = 8$.

2.2. Discriminative Approach: HCRFs

Compared to Linear Chain Conditional Random Fields as introduced in [13], Hidden Conditional Random Fields (HCRFs) additionally model an alignment between a source sequence (graphemes $\boldsymbol{g} = g_1, \ldots, g_K$) and a target sequence (phonemes $\boldsymbol{\varphi} = \varphi_1, \ldots, \varphi_N$), which is needed for G2P tasks. The alignment is integrated via a hidden variable. Hidden Conditional Random Fields (HCRFs), e.g. [14, 15], and Hidden Dynamic Conditional Random Fields (HDCRFs) [16] have been proposed in the literature. Our approach is similar to the latter one, where a sum over all possible alignments a_1^N is additionally introduced in training:

$$p(\boldsymbol{\varphi}|\boldsymbol{g}) = p_{\lambda_1^L}(\varphi_1^N | g_1^K) = \frac{\sum_{a_1^N} \exp H(\varphi_1^N, a_1^N, g_1^K)}{\sum_{\tilde{a}_1^N} \sum_{\tilde{\varphi}_1^N} \exp H(\tilde{\varphi}_1^N, \tilde{a}_1^N, g_1^K)}$$

$$H(\varphi_1^N, a_1^N, g_1^K) = \left(\sum_{n=1}^{N} \sum_{l=1}^{L} \lambda_l \cdot h_l(\varphi_{n-1}, \varphi_n, a_1^N, g_1^K) \right)$$

$H(\varphi_1^N, a_1^N, g_1^K)$ represents position dependent, binary feature functions $h_l(\varphi_{n-1}, \varphi_n, a_1^N, g_1^K)$. The maximization of the conditional log-likelihood is used as training criterion for the feature weights λ_1^L over a given training dataset. The decision criterion is given by the maximization of the sentence wise probability $p(\varphi_1^N | g_1^K)$, i.e. a maximum approximation is applied as for the joint-n-gram approach. To cope with the high computational complexity, certain restrictions are applied. Details about our implementation are given in [12]. It should be noted that within the HCRF and the Sequitur approach, an alignment respectively a segmentation of the data is implicitly and additionally learned.

For the HCRF model, we have applied lexical and source-n-gram features in a windows of $-5 \ldots 5$ around the current grapheme as well as the bigram features on phoneme side. Additionally, a prior has been introduced for smoothing. The model has been trained until convergence after 50 RProp iterations. Due to the large amount of features, feature selection methods have been applied (e.g. elastic net [17], feature count cut-off) resulting in 28M active features, i.e. features with non-zero weight.

Table 1: Statistics of the used QUAERO English corpora.

data set	duration[h]	# running words
train11	234.3	2.8M
dev10	3.3	40K
eval10	3.8	45K
eval11	3.3	35K
eval12	3.4	40K

Table 2: Statistics of the used background lexicon (Beep) and thus also the training data for the G2P models.

# symbols		⌀ word length		⌀ prons	# unique
source	target	source	target	per word	words
28	44	9.03	7.60	1.08	237k

3. Experimental Setup

In this section, the training schedule of the ASR system is presented as well as the strategy to integrate G2P into the ASR system. The various data sources used for training and testing are also introduced, whereas the experimental results will be presented in the following section.

3.1. Corpora

For the reported experiments, a state-of-the-art English LVCSR task based on the QUAERO 2011 data has been chosen [18]. The data for training comprises roughly 234h of audio data and mainly consists of broadcast news and podcasts. There are four datasets for development and evaluation of ASR systems provided which have a duration between three and four hours and are comprised of 35K to 45K words. An overview of the used ASR data is given in Tab. 1. For the training of the acoustic model as well as the G2P models, we have chosen the British English Example Pronunciation (BEEP) dictionary [19] as a background lexicon. The statistics are presented in Tab. 2. For the training of G2P models, the data have been split into a training and a development set, whereas the latter contains roughly 10K words (4% of the total data). The split has been done randomly and all pronunciation variants of a certain word are either in the training set or in the development set, but never spread across both. To avoid any encoding issues, all the data has been converted to UTF-8 in a preprocessing step.

3.2. ASR System

The used two-pass ASR system is based upon the QUAERO EN system as described in [18]. As features, Mel-Frequency Cepstral Coefficients (MFCC) were appended by a voicedness feature and phone-posterior-based features estimated using a multi-layer perceptron (MLP). More precisely, hmrasta bottleneck features have been utilized. The acoustic model itself is based on across-word triphone states represented by left-to-right three-state Hidden Markov Models. For speaker normalization, vocal tract length normalization (VTLN) on the feature vectors has been applied. Constrained Maximum Likelihood Linear Regression has been used as speaker adaptation

Table 3: Statistics about the overlap between the pronunciations of the two G2P models and the background lexicon. Both methods can more or less replicate the Beep pronunciations, but only 69.9% of the HCRF pronunciations are also in Seq; the other way around it is 70.9%.

	vocabulary	# pron vars	# pron vars[%] in HCRF	# pron vars[%] in Seq
Beep	69,956	76,318	98.4	96.6
Seq	145,385	179,656	70.9	100.0
HCRF	145,385	177,986	100.0	69.9

technique in training and recognition. For all presented results, the training criterion has been the Minimum Phone Error Rate (MPE). A pruned four-gram language model smoothed by modified Kneser-Ney discounting has been applied. This LM has been trained on approximately 3B words in various corpora, which have been linearly interpolated to optimize perplexity on a holdout data set.

3.3. Integrating G2P and ASR

For the ASR experiments, we did first fix a recognition vocabulary of 150K words, as usual based on count statistics from text data. Pronunciations for 5K regular abbreviations have been added via a rule-based approach (spelling of single letters). For the remaining 145K words, both G2P methods have been applied with the following setting: for each word, up to four pronunciation variants are generated. A variant is added to the lexicon, if it has a posterior confidence score ≥ 0.2. In several evaluations, we have found that this recipe leads to good performance on ASR tasks. A comparison of the overlap between the two resulting lexicons and the background lexicon is given in Tab. 3. Both G2P methods can more or less replicate the pronunciation variants from the background lexicon, which was to be expected since the models have been trained on that data. Interestingly, the pronunciations generated by the G2P models for words which are not part of the Beep lexicon do differ by roughly 30% (cf. Sec. 4.1 for a discussion of theses differences). Thus, there is an effect on ASR performance to be expected.

4. Experimental Results

Various experiments have been performed to measure the effect of G2P modeling on LVCSR performance. First, we will analyze and compare the two G2P systems which have been used within the ASR experiments. As error measure for these systems, we use the phoneme error rate (PER) and word error rate (WER). The PER is defined as the number of insertions, deletions and substitutions of a Levenshtein alignment between a hypothesis and a reference phoneme sequence. If there are multiple references, the alignment is done w.r.t. all references and the one with the least errors is chosen. The WER is defined as the number of wrongly recognized pronunciations w.r.t. the total number of reference pronunciations. Second, we will analyze the effect of three factors which influence the modeling of the lexicon: 1.) the G2P strategy itself, 2.) the number of pronunciation variants and 3.) the kind of pronunciation scores. For these ASR experiments, we use the well-known word error rate (WER) as measurement.

4.1. G2P

The results of the two tested G2P systems on the development set are presented in Tab. 4. With a PER of 1.7%, the Sequitur approach leads to better results than the HCRF approach with 2.1% PER. The overall performance of both methods on the Beep lexicon is quite good compared to other English G2P tasks. A comparison of errors of both methods revealed basically two sources for the performance differences. On the one hand, the HCRF approach tends to confuse the monophthongs "ae" (like in "*a*t", "f*a*st"), "ah" ("b*u*t", "s*u*n") and "ax" ("disc*u*s", "*a*bout") as well as the monophthongs "iy" (like in "b*ee*", "sh*e*") and "ih" ("b*i*g", "w*i*n") more often than the Seq approach. According to the reference pronunciations, they are often interchangeable across pronunciation variants, but not always. On the other hand, there are more deletions within the HCRF approach caused by e.g. omitting "r" like the final "r" in "bearer" than in the Seq approach. Here, the references are not consistent, since e.g. for "bearer" both pronunciation variants are within the BEEP lexicon (with and without the final "r" phoneme), while for "talebearer", only the variant with the final "r" phoneme is considered as correct. In general, the HCRF approach tends to generate shorter pronunciations than the Seq approach.

Table 4: Phoneme Error Rates (PER) and Word Error Rates (WER) on the development set for the Sequitur and the HCRF G2P system.

approach	PER[%] sub	del	ins	total	WER[%]
Seq	1.0	0.4	0.4	1.7	9.0
HCRF	1.2	0.5	0.3	2.1	11.6

4.2. LVCSR - Varying G2P Strategy

Concerning the ASR experiments, the following procedure has been applied: the vocabulary for training and recognition as well as the acoustic and language modeling data has been fixed and is the same for all experiments. Only the way of generating pronunciations has been interchanged. Since the G2P model is also needed in training the AM, we did a complete training from scratch for various ASR systems. Additionally, we always use pronunciation weights calculated on the training alignment via a forced alignment as presented in [20] for recognition. We have also tried to use pronunciation weights for training to better guide the alignment process, but they did not help. Thus, only words which have been observed in training will get pronunciation scores. The results are presented in Tab. 5 for three pairs of data sets, whereas the left set has been used for parameter optimization and the right one for testing. As baseline system, we use the Beep lexicon for pronunciation lookup and only if the respective word is not within the lexicon, we use the Sequitur G2P strategy (system 1 in the table). System 2 uses just the Sequitur G2P strategy without lookup in the Beep lexicon. For system 3, the pronunciation lookup has been performed with the Beep lexicon and the Sequitur G2P system and both outputs have been merged, denoted by "$\cup$". All following systems include the HCRF system. A combination of all available knowledge sources is the basis for system 6. All together, systems relying on HCRFs as G2P method outperform systems based on Sequitur, although the HCRF G2P model performs worse on text data (cf. Tab. 4). The best systems 5 and 7 (tuned on the respective dev set; denoted in bold) lead to a gain in performance between 1-3% relative over the baseline system.

4.3. LVCSR - Varying Number of Pronunciation Variants

Within another set of experiments, we wanted to analyze the effect of varying the number of pronunciation variants and

Table 5: ASR results on various QUAERO English development and test corpora. The first line represents the baseline system, where the Sequitur G2P model was only used iff the respective word was not in the Beep lexicon. This "hierarchical" lookup is denoted by "→". The "∪"-symbol denotes a merge of the models' hypotheses.

system number	pronunciation lookup	WER[%] dev10	eval10	eval10	eval11	eval11	eval12	# pronunciation variants
1	Beep → Seq	16.46	16.41	16.40	21.78	21.72	18.66	181.604
2	Seq	16.57	16.38	16.33	21.74	21.61	18.44	184.313
3	Seq ∪ Beep	16.44	16.31	16.19	21.54	21.26	18.28	189.303
4	HCRF	16.39	16.39	16.22	21.34	21.20	18.28	182.644
5	HCRF ∪ Beep	16.42	16.17	16.10	**21.20**	21.07	18.36	183.882
6	HCRF ∪ Seq ∪ Beep	16.32	16.23	16.16	21.18	21.04	18.08	239.150
7	Beep → HCRF	16.31	**16.28**	16.14	21.28	21.02	**18.04**	178.588

Table 6: Results for the ASR systems based on Seq and HCRF G2P modeling only. The (fixed) number of pronunciations per word has been optimized on the dev sets and varies with the type of pronunciation score and the G2P method.

	pron scores	# prons/ word	WER[%] dev10	eval10	eval10	eval11
HCRF	none	2	17.57	17.43	17.35	22.70
	G2P	4	16.70	16.60	16.53	21.70
	train	5	16.32	**16.39**	16.32	**21.41**
Seq	none	2	17.99	17.95	17.95	23.10
	G2P	3	16.70	16.99	16.92	22.27
	train	3	16.68	**16.69**	16.69	**21.80**

also the use of G2P confidence scores as pronunciation scores. Therefor, we have chosen the two ASR systems which rely only on a G2P model for pronunciation modeling (cf. systems 2 and 4 in Tab. 5). We optimized the number of (fixed) pronunciations per word on the dev10 set without using the confidences for thresholding. The results are shown in Tab. 6. Again, the HCRF model outperforms the Sequitur model, independent of the type of pronunciation score. Within each method, the pronunciation scores calculated on the alignment of the training data outperform the system's confidence scores. Additionally, if no pronunciation scores are used, more than 2 pronunciations per word lead to worse systems whereas the optimal number of pronunciations per word is higher if pronunciation scores are used. Using pronunciation scores leads to a gain in all cases. The boldface numbers refer to the best setup per G2P strategy.

4.4. LVCSR - Varying Pronunciation Scores

With a last set of experiments, we wanted to overcome one drawback when using the training alignment as the only source for pronunciation scores: this is only possible for words (more precisely: pronunciations) which occur in the acoustic training data. We wanted to assign pronunciation scores to all words in the recognition lexicon. Additionally, we wanted to verify the gain by using pronunciation scores without varying the number of pronunciations.

As baseline systems, we again use the systems based on Sequitur and HCRFs only, i.e. without Beep lookup. The results are presented in Tab. 7. Whereas lines 2 and 6 ("train align") show the baseline results which are taken from Tab. 5, line 1 and 5 ("none") show results without using pronunciation scores at all, which means that all variants are weighted equally. To include pronunciation scores does apparently help. But when

Table 7: Results for the ASR systems based on Seq and HCRF G2P modeling only. Here, only the pronunciation scores have been varied. For the "mix" lines, the G2P system's confidence score has been used as pronunciation score iff the pronunciation did not occur in the training alignment.

setup	pron scores	WER[%] dev10	eval10	eval10	eval11
HCRF	none	16.58	16.40	16.35	21.65
	train align	16.39	16.39	16.22	**21.34**
	G2P scores	16.61	16.48	16.31	21.63
	mix	16.38	**16.39**	16.24	21.33
Seq	none	16.66	16.50	16.50	21.94
	train align	16.57	16.38	16.33	21.74
	G2P scores	16.87	16.82	16.58	22.07
	mix	16.57	16.38	16.30	21.80

raw G2P posterior scores are used as pronunciation scores (lines 3 and 7, "G2P scores"), the quality of the ASR system drops. Even a combination of the G2P system with the scores from the training alignment ("mix") does not help. It should be noted that the posterior scores are always normalized per word, i.e. across all pronunciation variants per word. Boldface numbers represent the best performing system per data set, again after tuning the systems on the respective dev set.

5. Conclusion

In this paper, we have shown that G2P modeling using HCRFs can outperform a generative Sequitur G2P model within LVCSR experiments, even is the PER of the HCRF model on text data is worse than the Sequitur approach. Improvements of 1-3% could be achieved across a number of test sets from the English QUAERO tasks. To include pronunciation variants is also helpful, especially if pronunciation scores are used. We could also verify that pronunciation weights calculated on the training alignment improve performance. To include the posterior scores from G2P systems as pronunciation weights does not seem to help.

6. Acknowledgements

This work was partly realized as part of the Quaero Programme, funded by OSEO, French State agency for innovation. H. Ney was partially supported by a senior chair award from DIGITEO, a French research cluster in Ile-de-France.

7. References

[1] S. Deligne, F. Yvon, and F. Bimbot, "Variable-Length Sequence Matching for Phonetic Transcription Using Joint Multigrams," in *Proceeding of the Fourth European Conference on Speech Communication and Technology (EUROSPEECH)*, Madrid, Spain, Sep. 1995, pp. 2243–2246.

[2] S. F. Chen, "Conditional and Joint Models for Grapheme-to-Phoneme Conversion," in *Proc. European Conf. on Speech Communication and Technology*, Geneva, Switzerland, Sep. 2003, pp. 2033 – 2036.

[3] P. Vozila, J. Adams, Y. Lobacheva, and R. Thomas, "Grapheme to Phoneme Conversion and Dictionary Verification Using Graphonemes," in *European Conf. on Speech Communication and Technology*, Geneva, Switzerland, Sep. 2003, pp. 2469 – 2472.

[4] M. Bisani and H. Ney, "Joint-Sequence Models for Grapheme-to-Phoneme Conversion," *Speech Communication*, vol. 50, no. 5, pp. 434–451, May 2008.

[5] J. R. Novak, P. R. Dixon, N. Minematsu, K. Hirose, C. Hori, and H. Kashioka, "Improving WFST-based G2P Conversion with Alignment Constraints and RNNLM N-best Rescoring," in *Interspeech*, Portland, OR, USA, Sep. 2012.

[6] M. Bisani, "Sequitur G2P," 2008, http://www-i6.informatik.rwth-aachen.de/web/Software/g2p.html.

[7] J. Novak, "Phonetisaurus: A WFST-driven Phoneticizer," 2011, http://code.google.com/p/phonetisaurus/.

[8] S. Hahn, P. Vozila, and M. Bisani, "Comparison of Grapheme-to-Phoneme Methods on Large Pronunciation Dictionaries and LVCSR Tasks," in *Interspeech*, Portland, OR, USA, Sep. 2012.

[9] S. Jiampojamarn, C. Cherry, and G. Kondrak, "Integrating Joint n-Gram Features into a Discriminative Training Framework," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, ser. HLT '10. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010, pp. 697–700. [Online]. Available: http://dl.acm.org/citation.cfm?id=1857999.1858102

[10] S. Hahn, M. Dinarelli, C. Raymond, F. Lefevre, P. Lehnen, R. De Mori, A. Moschitti, H. Ney, and G. Riccardi, "Comparing Stochastic Approaches to Spoken Language Understanding in Multiple Languages," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 6, pp. 1569–1583, Aug. 2011.

[11] D. Wang and S. King, "Letter-to-Sound Pronunciation Prediction Using Conditional Random Fields," *IEEE Signal Processing Letters*, vol. 18, no. 2, pp. 122–125, 2011.

[12] P. Lehnen, S. Hahn, V.-A. Guta, and H. Ney, "Hidden Conditional Random Fields with M-to-N Alignments for Grapheme-to-Phoneme Conversion," in *Interspeech*, Portland, OR, USA, Sep. 2012.

[13] J. Lafferty, A. McCallum, and F. Pereira, "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data," in *Proceedings of the Eighteenth International Conference on Machine Learning (ICML)*, Williamstown, MA, USA, Jun. 2001, pp. 282–289.

[14] A. Quattoni, S. Wang, L.-P. Morency, M. Collins, and T. Darrell, "Hidden Conditional Random Fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, pp. 1848–1852, 2007.

[15] T. Koo and M. Collins, "Hidden-Variable Models for Discriminative Reranking," in *HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*. Morristown, NJ, USA: Association for Computational Linguistics, 2005, pp. 507–514.

[16] X. Yu and W. Lam, "Hidden Dynamic Probabilistic Models for Labeling Sequence Data," in *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence*, Chicago, IL, USA, Jul. 2008, pp. 739–745.

[17] T. Lavergne, O. Cappé, and F. Yvon, "Practical Very Large Scale CRFs," in *Proceedings the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*. Association for Computational Linguistics, July 2010, pp. 504–513. [Online]. Available: http://www.aclweb.org/anthology/P10-1052

[18] M. Sundermeyer, M. Nußbaum-Thom, S. Wiesler, C. Plahl, A. El-Desoky Mousa, S. Hahn, D. Nolden, R. Schlüter, and H. Ney, "The RWTH 2010 Quaero ASR Evaluation System for English, French, and German," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Prague, Czech Republic, May 2011, pp. 2212–2215.

[19] T. Robinson, "BEEP - The British English Example Pronunciation Dictionary," 1995, ftp://svr-ftp.eng.cam.ac.uk/comp.speech/dictionaries/.

[20] C. Gollan and H. Ney, "Towards Automatic Learning in LVCSR: Rapid Development of a Persian Broadcast Transcription System," in *Interspeech*, Brisbane, Australia, Sep. 2008, pp. 1441–1444.