POSTERIOR-BASED MULTI-STREAM FORMULATION TO COMBINE MULTIPLE GRAPHEME-TO-PHONEME CONVERSION TECHNIQUES

Marzieh Razavi Mathew Magimai.-Doss

Idiap-RR-33-2015

OCTOBER 2015

Centre du Parc, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

POSTERIOR-BASED MULTI-STREAM FORMULATION TO COMBINE MULTIPLE GRAPHEME-TO-PHONEME CONVERSION TECHNIQUES

Marzieh Razavi[1,2] *and Mathew Magimai.-Doss*[1]

[1] Idiap Research Institute, CH-1920 Martigny, Switzerland
[2] Ecole Polytechnique Fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland
{marzieh.razavi, mathew}@idiap.ch

ABSTRACT

In the literature, a number of approaches have been proposed for learning grapheme-to-phoneme (G2P) relationship and inferring pronunciations. The paper presents a multi-stream framework where different G2P relationship learning techniques can be effectively combined during pronunciation inference. Specifically, analogous to multi-stream automatic speech recognition in the literature, the framework involves (a) obtaining different streams of estimates of probability of phonemes given graphemes; (b) combining them based on probability combination rules; and (c) inferring pronunciations by decoding the probabilities resulting after combination. We demonstrate the potential of the proposed approach by combining state-of-the-art CRF-based G2P conversion approach and acoustic data-driven G2P conversion approach in the Kullback-Leibler divergence based HMM framework on the PhoneBook 600 words task.

***Index Terms*—** grapheme-to-phoneme conversion, automatic speech recognition, Kullback-Leibler divergence based HMM, conditional random fields, multi-stream framework

1. INTRODUCTION

State-of-the-art automatic speech recognition (ASR) and text-to-speech synthesis (TTS) systems are based on phonemes/phones. This necessitates having a well developed phonetic lexicon which transcribes each word as a sequence of phonemes. Development of a phonetic lexicon is a semi-automatic process. More precisely, given an initial hand crafted seed lexicon based on linguistic expertise in the target language, grapheme-to-phoneme (G2P) conversion techniques are used to generate pronunciations for new words.

In sequence processing terms, the goal of G2P conversion is to predict a sequence of phonemes given a sequence of graphemes (obtained from orthography of the word). In the literature, this problem has been approached in knowledge-driven manner [1, 2] and in data-driven manner through application of different statistical pattern recognition methods, namely, decision trees [3, 4, 5], artificial neural networks (ANNs) [6], hidden Markov models (HMMs) [7], joint multigram modeling [8], conditional random fields (CRFs) [9], hidden CRFs [10] and bidirectional long short-term (BLSTM) neural networks [11]. These approaches tend to achieve G2P conversion solely based on the seed lexicon. Recently, a G2P conversion approach in the framework of Kullback-Leibler divergence based HMM (KL-HMM) has been proposed, which uses both acoustic data and seed lexicon for G2P conversion [12].

More recently, it was elucidated that G2P conversion can be formulated in more abstract terms as estimation of sequence of probability of phonemes given grapheme input and decoding the phoneme posterior probabilities through an ergodic HMM to infer a phoneme sequence [13] (*Section 2*). It was shown that the decision tree based approach, ANN-based approach and acoustic G2P conversion approach are particular cases of such an abstract formulation. The present paper builds on that abstract formulation to show that the formulation can be effectively exploited to combine different G2P conversion approaches. More precisely, the sequence of phoneme posterior probabilities estimated from different G2P conversion techniques are treated as multiple streams, which are combined and a phoneme sequence is then inferred. This is analogous to multi-stream ASR framework where phoneme posterior probabilities estimated by classifiers with different feature inputs are combined and then used for speech recognition [14, 15, 16] (*Section 3*)

Our motivation to combine multiple G2P conversion techniques stems from the following reasons. *Firstly*, learning the relationship between graphemes and phonemes lies at the core of any G2P conversion technique. Such a learning, in statistical

This work was supported by Hasler foundation through the grant Flexible acoustic data driven grapheme to acoustic unit conversion.

terms, can be seen as training of probability of phoneme f_k given grapheme input g_n $P(f_k|g_n)$ estimator, where $k \in \{1, \cdots K\}$ and K is the number of phonemes. Given the alignment between the grapheme sequence and phoneme sequence, there are many methods to learn the probabilistic relationship $P(f_k|g_n)$, such as by counting, by training a locally discriminative classifier [6] or by training a globally discriminative classifier [9]. *Secondly*, as pointed out earlier, there are approaches such as acoustic data-driven G2P conversion approach that, unlike conventional G2P conversion techniques, employ acoustic information in addition to the seed lexicon to learn the G2P relationship. *Finally*, none of the G2P conversion techniques could be outrightly seen as the best method. This comes from the observation that differences in the pronunciation level performance between G2P conversion techniques may not necessarily translate as end use case (e.g., ASR) performance differences [10, 13]. Therefore, there can be benefits in inferring pronunciations by combining multiple estimates of $P(f_k|g_n)$. We demonstrate that through an investigation on combination of CRF-based G2P conversion technique and acoustic G2P conversion technique (*Section 4* and *Section 5*).

2. POSTERIOR-BASED G2P CONVERSION FORMALISM

Given a sequence of graphemes $G = (g_1, \ldots, g_n, \ldots, g_N)$, the G2P conversion problem in an HMM-based framework can be expressed as finding the most probable phoneme sequence F^* that can be achieved by finding the most likely state sequence S^*:

$$S^* = \arg\max_{S \in \mathcal{S}} P(G, S|\Theta) = \arg\max_{S \in \mathcal{S}} P(G|S, \Theta)P(S|\Theta) \tag{1}$$

where Θ denotes the parameters of the system, $\mathcal{S}$ denotes the set of possible HMM state sequences, and $S = (s_1, \cdots, s_n, \cdots, s_N)$ denotes a sequence of HMM states which corresponds to a phoneme sequence hypothesis with $s_n \in \mathcal{F} = \{f_1, \ldots, f_k, \ldots, f_K\}$ where K is the number of phoneme units. By applying *i.i.d.* and first order Markov assumption, Equation 1 can be simplified as:

$$S^* = \arg\max_{S \in \mathcal{S}} \prod_{n=1}^{N} P(g_n|s_n = f_k, \Theta)P(s_n = f_k|s_{n-1} = f_{k'}, \Theta) \tag{2}$$

Then through applying Bayes rule and ruling out the parameters that do not affect the maximization, i.e., $P(g_n|\Theta)$, Equation 2 can be written as:

$$S^* = \arg\max_{S \in \mathcal{S}} \prod_{n=1}^{N} \underbrace{\frac{P(s_n = f_k|g_n, \Theta)}{P(s_n = f_k|\Theta)}}_{\text{local emission score}} \underbrace{P(s_n = f_k|s_{n-1} = f_{k'}, \Theta)}_{\text{transition probability}}. \tag{3}$$

Estimation of the prior probability $P(s_n = f_k|\Theta)$ is a challenging problem as we have access to only a few words (not all the words in the language) in the seed lexicon. Therefore, rather than estimating $P(s_n = f_k|\Theta)$ we assume equal phoneme prior probabilities:

$$S^* = \arg\max_{S \in \mathcal{S}} \prod_{n=1}^{N} \underbrace{P(s_n = f_k|g_n, \Theta)}_{\text{local emission score}} \underbrace{P(s_n = f_k|s_{n-1} = f_{k'}, \Theta)}_{\text{transition probability}}. \tag{4}$$

Finally, if the transition probabilities are assumed to be uniform, i.e., ergodic HMM, then,

$$S^* = \arg\max_{S \in \mathcal{S}} \prod_{n=1}^{N} \underbrace{P(s_n = f_k|g_n, \Theta)}_{\text{local score}}. \tag{5}$$

Such an assumption is reasonable as robust estimation of transition probabilities from the few pronunciations present in the seed lexicon is not trivial.

In this paper, we will see that $P(s_n = f_k|g_n, \Theta)$ can be estimated as combination of estimates obtained from different G2P conversion techniques, which eventually yields a pronunciation lexicon that helps in building better ASR systems.

3. COMBINATION OF G2P RELATIONSHIP LEARNING TECHNIQUES AND PRONUNCIATION INFERENCE

In this paper, we demonstrate the potential of the multi-stream formulation through an investigation on combining CRF-based approach and acoustic data-driven G2P conversion approach. This section gives a brief overview about the G2P conversion approaches investigated and the multi-stream combination mechanism for pronunciation inference. In addition to that, we provide a theoretical insight into the investigated combination through a link to the ASR literature.

3.1. CRF-Based G2P Conversion Approach

The CRF-based G2P conversion approach is a probabilistic sequence modeling-based approach which enables global inference, discriminative training and relaxing the independence assumption existing in HMMs [17]. In the case of G2P conversion, the input to the CRF is the grapheme sequence obtained from the orthography of the word, and the CRF output is the predicted phoneme sequence. In this approach, the posterior probability for each phoneme f_k given the entire grapheme sequence G denoted as $P_{crf}(s_n = f_k|G)$ can be efficiently estimated using the well-known forward-backward algorithm [17]. In other words, each time instance n will yield a probability vector $[P_{crf}(s_n = f_1|G) \cdots P_{crf}(s_n = f_K|G)]^{\mathrm{T}}$.

3.2. Acoustic Data-Driven G2P Conversion Approach

The acoustic data-driven G2P conversion approach is a particular case of the posterior-based G2P conversion formalism presented in Section 2, in which estimation of probability of each phoneme f_k given a local grapheme context g_n, denoted as $P_{ag2p}(s_n = f_k|g_n)$, at each time instance n is done in two stages. In the first stage, a probabilistic grapheme-to-phoneme relationship is learned through acoustic data using KL-HMM [18, 19]. Briefly, this involves first training of an ANN to classify phonemes. This is then followed by training of KL-HMM, in which phoneme posterior probabilities estimated by ANN are used as feature observations. Each KL-HMM state represents a context-dependent grapheme and is parameterized by a categorical distribution of phonemes. The KL-HMM parameters are estimated using Viterbi Expectation-Maximization algorithm with a cost function based on KL-divergence. In the second stage, given a word, the KL-HMM is used to generate sequence of probability vectors $[P_{ag2p}(s_n = f_1|g_n) \cdots P_{ag2p}(s_n = f_K|g_n)]^{\mathrm{T}}, \forall n$ based on the sequence of graphemes in the orthography of the word. In order to infer the pronunciation of the word, the sequence of probability vectors are decoded according to Equation (5). For more details the readers are referred to [12, 13].

3.3. Multi-Stream Combination

Figure 1 depicts a schematic view of the multi-stream combination. Briefly, given the two sequences of phoneme posterior probabilities estimated by the two approaches, at each time instance n the phoneme probability estimates $[P_{crf}(s_n = f_1|G) \cdots P_{crf}(s_n = f_K|G)]^{\mathrm{T}}$ and $[P_{ag2p}(s_n = f_1|g_n) \cdots P_{ag2p}(s_n = f_K|g_n)]^{\mathrm{T}}$ are combined using probability combination rules [20]. The resulting sequence of phoneme probabilities is then decoded according to Equation (5).

In addition to the fact that the CRF-based approach and the acoustic data-driven approach use different statistical models and information to learn the G2P relationship, theoretically, the combination presented here is synonymous to an approach studied in the literature to combine ”global/hierarchical” phoneme posterior probability estimates with local phoneme posterior probability estimates [21] to improve performance of the ASR system. Specifically, in comparison to that approach, $[P_{crf}(s_n = f_1|G) \cdots P_{crf}(s_n = f_K|G)]^{\mathrm{T}}$ is synonymous to global phoneme posterior probability, i.e., phoneme posterior probabilities estimate given the whole acoustic feature sequence using forward-backward algorithm [22]. While $[P_{ag2p}(s_n = f_1|g_n) \cdots P_{ag2p}(s_n = f_K|g_n)]^{\mathrm{T}}$ is synonymous to phoneme posterior probabilities given a local acoustic feature input (in simple terms, the output of ANN given a local acoustic feature input), as the KL-HMM states only model a local grapheme context. Thus, we hypothesize that the combination investigated in this paper should be beneficial.

4. EXPERIMENTAL SETUP

We evaluated the G2P conversion task on the PhoneBook corpus [23]. It is a challenging task for several reasons: 1) in English the G2P relationship is highly irregular; 2) the training and test vocabulary sets are entirely different; 3) the corpus contains uncommon English words and proper names (e.g. Witherington, Gargantuan, etc); and 4) the number of words in the seed lexicon is relatively small which makes reliable estimation of $P_{crf}(s_n = f_k|G)$ and $P_{ag2p}(s_n = f_k|g_n)$ really challenging.

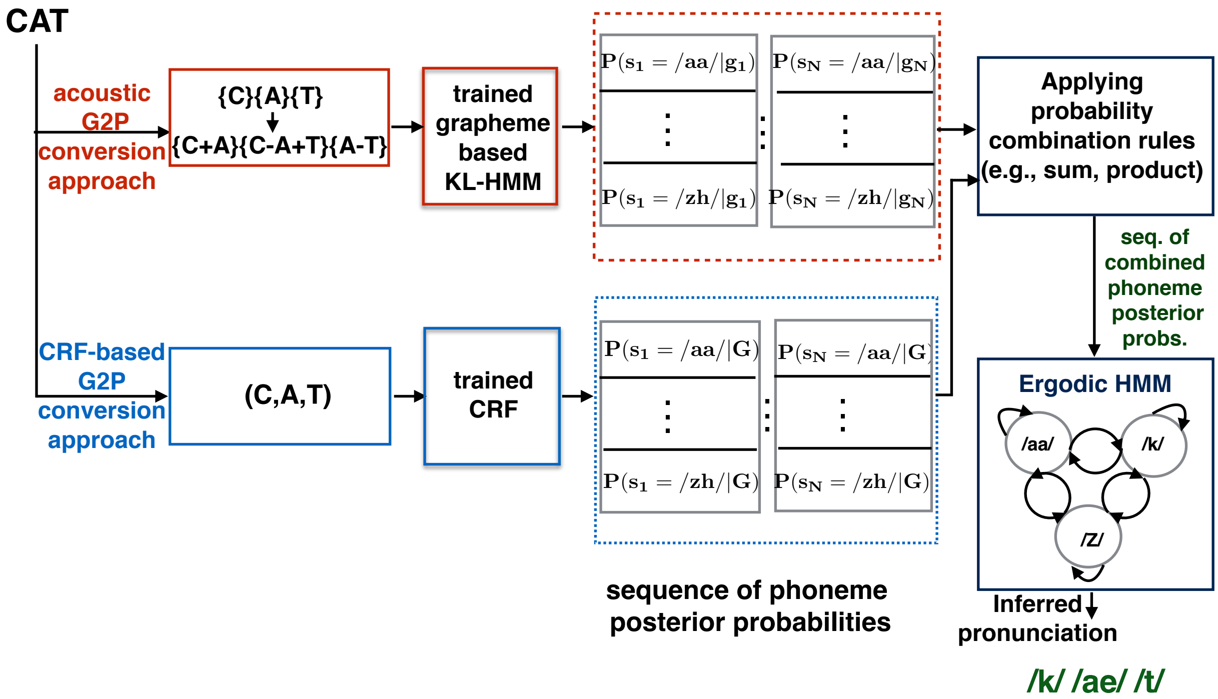
CAT
acoustic G2P conversion approach
{C}{A}{T}
{C+A}{C-A+T}{A-T}
trained grapheme based KL-HMM
P(s1 = /aa/|g1)
P(s1 = /zh/|g1)
P(sN = /aa/|gN)
P(sN = /zh/|gN)
CRF-based G2P conversion approach
(C,A,T)
trained CRF
P(s1 = /aa/|G)
P(s1 = /zh/|G)
P(sN = /aa/|G)
P(sN = /zh/|G)
sequence of phoneme posterior probabilities
Applying probability combination rules (e.g., sum, product)
seq. of combined phoneme posterior probs.
Ergodic HMM
/aa/
/k/
/z/
Inferred pronunciation
/k/ /ae/ /t/

Fig. 1: Illustration of pronunciation inference using multi-stream combination of CRF-based phoneme posterior probabilities sequence and acoustic data-driven G2P-based phoneme posterior probabilities sequence.

4.1. Dataset

We use the medium size vocabulary task with 602 unique words setup defined for speaker-independent task-independent isolated word recognition in [24]. Table 1 gives an overview of the dataset. All the words and speakers across train, cross-validation and test set are entirely different. The PhoneBook pronunciation lexicon is transcribed using 42 phonemes (including silence).

Number of	Train	Cross-validation	Test
Utterances	19421	7290	6598
Hours	7.7	2.9	2.6
Speakers	243	106	96
Words	1580	603	602

Table 1: Overview of the PhoneBook corpus.

4.2. Lexicon Generation

This section explains the different lexicon generation setups studied.

4.2.1. CRF-based G2P conversion approach

In order to train the CRFs, a preliminary alignment between the graphemes and phonemes in the training lexicon is required. In this paper, we use the m2m-aligner [25] to determine the G2P alignment. To train and decode the CRF, we used the publicly available CRF++ software[1]. We used bigram features and set the grapheme context to 9, i.e., four preceding and following graphemes as done in [26].

4.2.2. Acoustic data-driven G2P conversion approach

To learn the probabilistic grapheme-to-phoneme relationship, we first trained a 5-layer multilayer perceptron (MLP) using the Quicknet software [27]. The input to the MLP was 39-dimensional PLP cepstral features with four preceding and four

[1] https://taku910.github.io/crfpp/

following frame context. The MLP output units were 313 clustered context-dependent (CD) phonemes derived by clustering CD phonemes in HMM/Gaussian mixture model framework. We then trained a single preceding and following CD grapheme-based KL-HMM system. In the cost function based on the KL-divergence, the output of MLP was used as the reference distribution. To handle unseen contexts, we used the KL-divergence based decision tree state tying method proposed in [28]. After the KL-HMM training, as we are interested in inferring context-independent phoneme sequence, the clustered CD phoneme categorical distribution estimated for each state was marginalized based on the central phoneme information.

4.2.3. Multi-stream combination and pronunciation inference

We investigated two probability combination rules, namely product rule and sum rule [20, 29], with static weighting. More precisely,

$$\text{Comb-prod} = \frac{1}{Z} \cdot \prod_{k=1}^{K} P_{crf}(s_n = f_k|G)^{w_{crf}} \quad \cdot P_{ag2p}(s_n = f_k|g_n)^{w_{ag2p}} \tag{6}$$

$$\text{Comb-sum} = \frac{1}{Z} \cdot \sum_{k=1}^{K} w_{crf} \cdot P_{crf}(s_n = f_k|G) \quad + w_{ag2p} \cdot P_{ag2p}(s_n = f_k|g_n) \tag{7}$$

where Z is a normalization factor, w_{crf} is the weight given to CRF G2P relationship stream and w_{ag2p} is the weight given to acoustic data driven G2P relationship stream, $0 \leq w_{crf}, w_{ag2p}, \leq 1$ and $w_{crf} + w_{ag2p} = 1$. w_{crf} and w_{ag2p} were estimated by running the multi-stream combination-based pronunciation inference on the training data and selected the one yielded the lowest phone error rate.

4.3. Evaluation

We evaluated the performance of the generated lexicons at two different levels, namely, at pronunciation level as conventionally done in G2P conversion literature and at ASR level. One of the easiest approaches to combine CRF-based G2P conversion approach and acoustic data-driven G2P conversion approach is to combine the lexicons generated by the two methods. So, at ASR level we compared the multi-stream approach against lexicon combination approach by generating 2-best pronunciations.

For the ASR study, we trained standard cross-word context-dependent phoneme-based HMM/Gaussian mixture model (GMM) systems for each of the phonetic lexicons generated through the G2P conversion approaches using HTK [30]. We used 39 dimensional PLP cepstral features (static+dynamic features). Each subword unit was modeled with three HMM states. Each HMM state was modeled by a mixture of 8 Gaussians. The HMM states were tied using singleton question set.

5. ANALYSIS AND RESULTS

This section first provides a brief analysis about the possible benefits of combining G2P conversion techniques. Then it presents the evaluation results at both pronunciation and ASR levels.

5.1. Analysis

In the proposed approach, the first question that arises is: how different are $[P_{crf}(s_n = f_1|G) \cdots P_{crf}(s_n = f_K|G)]^{\mathrm{T}}$ and $[P_{ag2p}(s_n = f_1|g_n) \cdots P_{ag2p}(s_n = f_K|g_n)]^{\mathrm{T}}$? To understand that we estimated the entropy of these distributions on the training set. Figure 2 plots a histogram of it. As it can be seen, the probabilities estimated by CRF have low entropy compared to acoustic G2P conversion approach. There is very little overlap between the distributions. In other words, CRF output has high confidence. The static weight capture this difference. For the product rule (Comb-prod) $w_{crf} = 0.8$ and for the sum rule (Comb-sum) $w_{crf} = 0.7$.

Given that the entropies are so different, a question that arises is: would we get any different pronunciation than the one estimated by CRF by combination? Table 2 presents a few of the generated pronunciations through each approach together with the manual pronunciation. It can be observed that indeed the multi-stream combination paradigm is able to exploit the merits of both approaches.

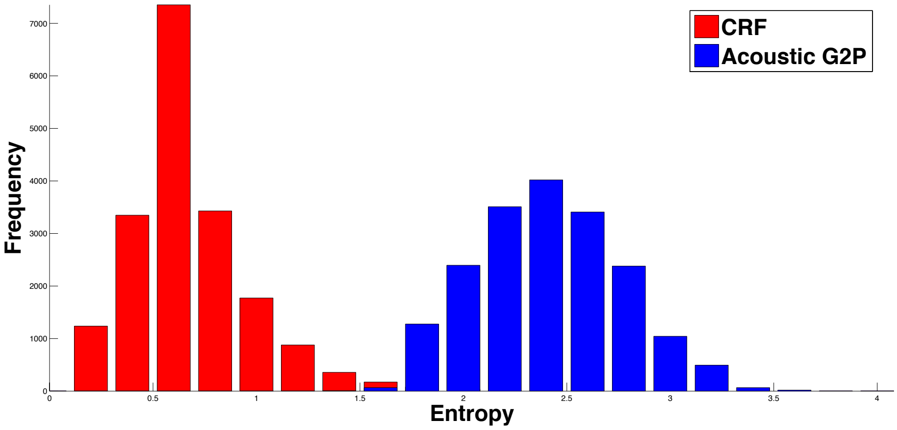

Fig. 2: Histogram of entropy of phoneme posterior probability distributions for the acoustic G2P conversion and the CRF-based G2P conversion approaches.

Word	CRF-based pronunciation	ag2p-based pronunciation	Comb-prod pronunciation	Manual pronunciation
attribution	@ t r x b u S x n	@ t r Y b ˆS x n	@ t r x b y u S x n	@ t r x b y u S x n
beirut	b i r ˆt	b Y r u t	b i r u t	b e r u t
exorbitant	x k s c r b x t x n t	x g z c r b x t @ n t	x g z c r b x t x n t	x g z c r b x t x n t

Table 2: Pronunciations generated by different G2P approaches along with the manual pronunciations.

5.2. Pronunciation Level Results

Table 3 provides pronunciation level evaluation results in terms of phone error rate (PER) and word error rate (WER). It can be observed that the proposed method leads to significant improvements at the pronunciation level compared to the acoustic G2P conversion approach. However, there is no gain at the pronunciation level over the CRF-based G2P conversion approach.

	Acoustic G2P	CRF G2P	Comb-sum	Comb-prod
PER	23.1	11.5	12.4	12.9
WER	82.4	49.2	53.8	55.7

Table 3: Pronunciation level evaluations in terms of PER and WER

5.3. Comparison Across G2P Conversion Approaches

Table 4 presents the ASR evaluation results in terms of word error rate (WER). Comparison across individual G2P conversion approaches shows that the lexicon based on CRF approach yields the best system. This performance is similar to that of joint multigram approach (WER of 10.6%[13]), which is another the state-of-the-art G2P conversion approach. We also ran an experiment where the probabilities obtained from the CRF are decoded according to Equation (5), which assumes uniform transition probability. Interestingly, we obtained a performance of 10.9% WER, thus suggesting that the posterior-based formulation is quite generic. It can be observed that despite wide difference in PER and WER at pronunciation level we see that the lexicon from the acoustic G2P conversion approach yields a system that is not too far from the CRF-based lexicon. Such a trend has been observed before in comparison to other G2P conversion approaches [13].

Though the multi-stream approaches perform poor at the pronunciation inference level when compared to CRF-based approach, at ASR level we see improvements. The improvements obtained with product rule (Comb-prod) are statistically significant, while with sum rule (Comb-sum) the improvements are marginal. The trend could be related to the fact that the entropies of the probability distributions estimated by the two approaches are very different. We speculate that the product rule is giving more importance to the estimates of CRF. Finally, these results also show that the pronunciation level performance is not necessarily indicative of the performance at the recognition level. Such a trend has been observed in the G2P conversion literature [26, 10, 13].

	Acoustic G2P	CRF G2P	Comb-sum	Comb-prod	Manual lexicon [13]
WER	11.5	10.8	10.6	**10.1**	1.8

Table 4: ASR level evaluations in terms of WER.

5.4. Comparison to Combination at Lexicon Level

Table 5 presents results of the ASR study comparing lexical level combination of CRF-based approach and acoustic G2P conversion approach, i.e. simply merging the lexicons, (Acoustic G2P+CRF) against the multi-stream approach (Comb-prod and Comb-sum) with 2-best pronunciations. We observe that the multi-stream approach yields better systems. Again the lexicon based on product rule yields a significantly better system.

	Acoustic G2P +CRF	Comb-sum	Comb-prod
WER	8.3	8.2	**7.6**

Table 5: Lexical level combination versus 2-best.

6. CONCLUSION AND FUTURE DIRECTIONS

The paper presented a posterior-based formulation to combine multiple estimates of phoneme probabilities conditioned on graphemes obtained by applying different G2P relationship learning mechanisms proposed in the literature for pronunciation lexicon development. Our study on combining posterior probability estimates obtained through CRF-based approach and acoustic G2P conversion approach showed that combining multiple estimates can yield pronunciation lexicons, which despite being relatively poor at pronunciation level, can help in building better ASR systems. As an extension to the present work, we aim to investigate: (a) combination with other G2P conversion approaches; (b) applying dynamic weighting techniques for combining probability distributions [15, 16]; and (c) evaluation on large vocabulary tasks and other languages.

In the literature, one technique to improve a pronunciation lexicon obtained with G2P conversion is to use acoustic realization of words either to select from pronunciation variants inferred by G2P conversion [31, 32] or to adapt graphoneme model parameters [33]. In Equation (5), it can be observed that if the input g_n is replaced by an acoustic feature input from a speech signal (in other words, if the input grapheme sequence is replaced by an acoustic speech signal), then the formulation reduces to acoustic data-driven pronunciation variant extraction framework [34]. Alternately, the orthographic information (i.e., sequence of probability of phonemes given graphemes) and the information from the acoustic realizations of the words (i.e., sequence of probability of phonemes given acoustic features) could be trivially combined in the proposed multi-stream formulation. Such a method could have potential implications towards development of lexicon for names, children speech and accented speech. Our future work will also focus along this direction together with the extensions pointed out earlier.

7. REFERENCES

[1] R.M. Kaplan and M. Kay, "Regular Models of Phonological Rule Systems," *Computational Linguistics*, vol. 20, pp. 331–378, 1994.

[2] M. Davel and E. Barnard, "Pronunciation prediction with Default&Refine," *Computer Speech and Language*, vol. 22, pp. 374–393, 2008.

[3] A. W. Black, K. Lenzo, and V. Pagel, "Issues in Building General Letter to Sound Rules," *ESCA Workshop on Speech Synthesis*, pp. 77–80, 1998.

[4] W. Daelemans and A. Van Den Bosch, "Language Independent Data-Oriented Grapheme-to-Phoneme Conversion," *Progress in speech synthesis*, pp. 77–89, 1997.

[5] V. Pagel, K. Lenzo, and A. W. Black, "Letter to sound rules for accented lexicon compression," in *Proceedings of ICSLP*, 1998, vol. 5, pp. 2015–2020.

[6] T. J. Sejnowski and C. R. Rosenberg, "Parallel Networks that Learn to Pronounce English Text," *Complex Systems*, vol. 1, pp. 145–168, 1987.

[7] P. Taylor, "Hidden Markov Models for Grapheme to Phoneme Conversion.," in *Proceedings of Interspeech*, 2005, pp. 1973–1976.

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

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

[10] S. Hahn, P. Lehnen, S. Wiesler, R. Schlter, and H. Ney, "Improving LVCSR with Hidden Conditional Random Fields for Grapheme-to-Phoneme Conversion.," in *Proceedings of Interspeech*, 2013, pp. 495–499.

[11] K. Yao and G. Zweig, "Sequence-to-Sequence Neural Net Models for Grapheme-to-Phoneme Conversion," in *Proceedings of Interspeech*, May 2015.

[12] R. Rasipuram and M. Magimai-Doss, "Acoustic Data-driven Grapheme-to-Phoneme Conversion using KL-HMM," in *Proceedings of ICASSP*, Mar. 2012.

[13] M. Razavi, R. Rasipuram, and M. Magimai.-Doss, "Acoustic Data-Driven Grapheme-to-Phoneme Conversion in the Probabilistic Lexical Modeling Framework," Idiap-RR Idiap-RR-10-2015, Idiap, 5 2015.

[14] A. Janin, D. Ellis, and N. Morgan, "Multi-stream speech recognition: ready for prime time?," in *EUROSPEECH*. 1999, ISCA.

[15] H. Misra, H. Bourlard, and V. Tyagi, "New entropy based combination rules in HMM/ANN multi-stream ASR," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2003.

[16] F. Valente, "Multi-stream speech recognition based on Dempster-Shafer combination rule," *Speech Communication*, vol. 52, no. 3, pp. 213–222, 2010.

[17] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data," in *Proceedings of ICML*, San Francisco, CA, USA, 2001, ICML '01, pp. 282–289, Morgan Kaufmann Publishers Inc.

[18] M. Magimai.-Doss, R. Rasipuram, G. Aradilla, and H. Bourlard, "Grapheme-based Automatic Speech Recognition using KL-HMM," in *Proceedings of Interspeech*, 2011, pp. 445–448.

[19] G. Aradilla, H. Bourlard, and M. Magimai Doss, "Using KL-Based Acoustic Models in a Large Vocabulary Recognition Task ," in *Proceedings of Interspeech*, 2008, pp. 928–931.

[20] C. Genest and J. V. Zidek, “Combining Probability Distributions: A Critique and an Annotated Bibliography,” *Statist. Sci.*, vol. 1, no. 1, pp. 114–135, 02 1986.

[21] H. Ketabdar and H. Bourlard, “Enhanced Phone Posteriors for Improving Speech Recognition Systems.,” *IEEE Transactions on Audio, Speech & Language Processing*, vol. 18, no. 6, pp. 1094–1106, 2010.

[22] H. Bourlard, S. Bengio, M. Magimai Doss, Q. Zhu, Mesot B., and N. Morgan, “Towards Using Hierarchical Posteriors for Flexible Automatic Speech Recognition Systems,” in *Proceedings of RT04*, 2004.

[23] J. Pitrelli, C. Fong, S.H. Wong, J.R. Spitz, and H.C. Leung, “PhoneBook: a Phonetically-Rich Isolated-Word Telephone-Speech Database,” in *Proceedings of ICASSP*, 1995, vol. 1, pp. 101–104.

[24] S. Dupont, H. Bourlard, O. Deroo, V. Fontaine, and J. M. Boite, “Hybrid HMM/ANN Systems for Training Independent Tasks: Experiments on ’Phonebook’ and Related Improvements,” in *Proceedings of ICASSP*, 1997.

[25] S. Jiampojamarn, G. Kondrak, and T. Sherif, “Applying Many-to-Many Alignments and Hidden Markov Models to Letter-to-Phoneme Conversion,” in *Proceedings of NAACL*, Rochester, New York, April 2007, pp. 372–379, Association for Computational Linguistics.

[26] D. Jouvet, D. Fohr, and I. Illina, “Evaluating Grapheme-to-Phoneme Converters in Automatic Speech Recognition Context,” in *Proceedings of ICASSP*, 2012, pp. 4821–4824.

[27] D. Johnson et al., “ICSI Quicknet Software Package,” http://www.icsi.berkeley.edu/Speech/qn.html, 2004.

[28] D. Imseng, J. Dines, P. Motlicek, P. N. Garner, and H. Bourlard, “Comparing Different Acoustic Modeling Techniques for Multilingual Boosting,” in *Proceedings of Interspeech*, Sept. 2012.

[29] D. M.J. Tax, M. van Breukelen, R. P.W. Duin, and J. Kittler, “Combining multiple classifiers by averaging or by multiplying?,” *Pattern Recognition*, vol. 33, no. 9, pp. 1475 – 1485, 2000.

[30] S.J. Young et al., *The HTK Book (for HTK Version 3.4)*, Cambridge University Engineering Department, UK, 2006.

[31] I. McGraw, I. Badr, and J.R. Glass, “Learning Lexicons From Speech Using a Pronunciation Mixture Model,” *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 21, no. 2, pp. 357–366, 2013.

[32] L. Lu, A. Ghoshal, and S. Renals, “Acoustic Data-Driven Pronunciation Lexicon For Large Vocabulary Speech Recognition,” in *Proceedings of ASRU*, 2013, pp. 374–379.

[33] L. Xiao, A. Gunawardana, and A. Acero, “Adapting Grapheme-to-Phoneme Conversion for Name Recognition,” in *Proceedings of ASRU*, 2007, pp. 130–135.

[34] H. Mokbel and D. Jouvet, “Derivation of the Optimal Set of Phonetic Transcriptions for a Word from its Acoustic Realizations ,” *Speech Communication*, vol. 29, no. 1, pp. 49 – 64, 1999.