A Posterior-Based Multi-Stream Formulation for G2P Conversion

Marzieh Razavi, *Student member, IEEE,* and Mathew Magimai.-Doss, *Member, IEEE*

Abstract

In the literature, a number of approaches have been proposed for learning grapheme-to-phoneme (G2P) relationship and inferring pronunciations. In this paper, we present a novel multi-stream framework for G2P conversion where various machine learning techniques providing different estimates of probability of phonemes given graphemes can be effectively combined during pronunciation inference. More precisely, analogous to multi-stream automatic speech recognition, 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 probabilities estimated by the state-of-the-art conditional random field-based G2P conversion approach and acoustic data-driven G2P conversion approach in the Kullback-Leibler divergence based hidden Markov model framework on the PhoneBook 600 words task.

Index Terms

grapheme-to-phoneme conversion, automatic speech recognition, multi-stream combination, Kullback-Leibler divergence based HMM, conditional random fields

I. Introduction

Lexicon development is one of the key steps in development of human language technologies such as automatic speech recognition (ASR) systems and text-to-speech synthesis (TTS) systems. This is

M. Razavi (email: marzieh.razavi@idiap.ch) and M. Magimai.-Doss (email: mathew@idiap.ch) are both affiliated to Idiap Research Institute, CH-1920 Martigny, Switzerland. This work was supported by Hasler foundation through the grant AddG2SU. Thanks to the reviewers for their valuable comments.

Manuscript received ????.

typically achieved in a semi-automatic manner by development of a seed lexicon followed by application of grapheme-to-phoneme (G2P) conversion techniques [1–6]. Another approach to infer a pronunciation model for a word would be to employ a phoneme recognition technique to get a phonetic transcription given its acoustic realization(s) [7–9]. In that respect, phoneme recognition can be regarded as acoustic-to-phoneme (A2P) conversion.

The paper builds on the parallels between G2P conversion and A2P conversion to propose a novel multi-stream formulation for G2P conversion, in a more general sense an approach that unifies G2P conversion and A2P conversion for pronunciation inference. More precisely,

1) In both G2P conversion and A2P conversion tasks, the goal is to predict or infer a phoneme sequence given an input observation sequence, i.e., sequence of graphemes in the case of G2P conversion task and sequence of acoustic features in the case of A2P conversion. In other words, both tasks need sequence modeling techniques.
2) In both tasks, the relationship between the observations (graphemes or acoustic features) and the phonemes is not deterministic. Thus, there is a need for statistical techniques to learn the relationship between the observations and phonemes. Towards that, different approaches have been proposed in the literature. In the case of G2P conversion, the G2P relationship can be captured through (a) counting methods [4]; (b) local classification techniques (e.g., artificial neural networks (ANN) [2], decision trees [1]); or (c) global classification techniques (e.g., conditional random fields (CRFs) [5]). Similarly, in the case of A2P conversion, the A2P relationship can be captured via ANNs [10] or Gaussian mixture models [11] to name a prominent few.

In the literature, one of the best methods for A2P conversion is based on hybrid hidden Markov model/ANN (HMM/ANN) approach [10, 12]. In this method, phoneme class conditional probabilities estimated using an ANN are decoded by a fully connected HMM to infer the phoneme sequence. A distinctive advantage of posterior probabilities is that they can be enhanced or refined by combination of multiple complementary estimates [13, 14]. In ASR community, the approach of combining multiple probability estimates, also known as multi-stream combination, has been found to be beneficial [15–19].

Given the parallels between the G2P conversion and A2P conversion, an interesting question arising is that whether the multi-stream combination method can be exploited to improve G2P conversion. Towards that, we first present a posterior based G2P conversion formalism analogous to hybrid HMM/ANN ASR approach, originally proposed in [20] (*Section II*). We then show how multiple estimates of $P(f|g)$, probability of phoneme f given grapheme g, can be estimated through different techniques and combined in a multi-stream fashion, exactly as done in ASR, for G2P conversion. Specifically, in this paper we study estimation and combination of $P(f|g)$ using CRF-based G2P conversion approach and acoustic data-driven G2P conversion approach using Kullback-Leibler divergence HMM (KL-HMM) (*Section III*). We evaluate the multi-stream formulation on speaker-independent task-independent setup of PhoneBook corpus (*Section IV*). Our experimental studies show that despite inferior performance at pronunciation level, the proposed formulation leads to significant improvements at ASR level (*Section V*).

II. Posterior-based G2P conversion formalism

Given a grapheme sequence $G = (g_1, \ldots, g_n, \ldots, g_N)$, G2P conversion 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 $s_n \in \mathcal{F} = \{f_1, \ldots, f_k, \ldots, f_K\}$, K being the number of phoneme units. For convenience, hereafter we drop Θ from the equations. By applying *i.i.d.* and first order Markov assumptions, Eqn. (1) can be simplified as:

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

Applying the Bayes' rule and ruling out the parameters not affecting the maximization lead to,

$$S^* = \arg\max_{S \in \mathcal{S}} \prod_{n=1}^{N} \frac{\overbrace{P(s_n = f_k|g_n)}^{\text{Posterior probability}}}{\underbrace{P(s_n = f_k)}_{\text{Prior probability}}} \cdot \underbrace{P(s_n = f_k|s_{n-1} = f_{k'})}_{\text{Transition probability}}. \tag{3}$$

As in the case of A2P conversion, the scaled-likelihoods are decoded by an ergodic HMM to infer a phoneme sequence.

III. Multi-stream combination of G2P relationship learning techniques

The posterior probability $P(s_n = f_k|g_n)$ in Eqn. (3) can be estimated by combining streams of phoneme posterior probabilities obtained from different G2P conversion techniques. In this paper, we validate such a multi-stream approach by combining estimates from the CRF-based approach, which learns the G2P relationship using only seed lexicon, with acoustic data-driven G2P conversion approach, which learns the G2P relationship using both seed lexicon and acoustics.

A. CRF-Based G2P conversion approach estimate

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 [21]. 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 [21]. In other words, each time instance n will yield a probability vector $[P_{crf}(s_n = f_1|G) \cdots P_{crf}(s_n = f_K|G)]^{\mathrm{T}}$.

B. Acoustic data-driven G2P conversion approach estimate

The acoustic data-driven G2P conversion approach is a particular case of the posterior-based G2P conversion formalism presented in Section II, 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 [22]. Briefly, this involves first training of an ANN to classify phonemes. This is then followed by training of a HMM using the phoneme posterior probabilities estimated by the ANN as feature observations, with an objective function based on KL-divergence [23]. Each KL-HMM state represents a context-dependent (CD) 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 obtain a 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 is decoded according to Eqn. (3). For more details the readers are referred to [6, 20].

C. Multi-stream combination

Given estimates from the two techniques for each time instance n in the input, the posterior probability in Eqn. (3) can be estimated by applying probability combination rules [13, 14], namely product rule (*Comb-prod*) and sum rule (*Comb-sum*) with weights assigned to each stream as shown in Eqn. (4) and

Eqn. (5) respectively:

$$P_{prod}(s_n = f_k|g_n) = \frac{1}{Z_p(n)} \cdot [P_{crf}(s_n = f_k|G)^{w_{crf}} \quad \cdot$$
$$P_{ag2p}(s_n = f_k|g_n)^{w_{ag2p}}] \tag{4}$$
$$P_{sum}(s_n = f_k|g_n) = \frac{1}{Z_s(n)} \cdot [w_{crf} \cdot P_{crf}(s_n = f_k|G) \quad +$$
$$w_{ag2p} \cdot P_{ag2p}(s_n = f_k|g_n)], \tag{5}$$

where $Z_p(n)$ and $Z_s(n)$ are normalization factors at time instance n, 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 $w_{crf} + w_{ag2p} = 1$. The weights w_{crf} and w_{ag2p} can be statically or dynamically estimated.

Similar combinations based on estimates of $P(f|g)$ through other G2P relationship learning techniques, such as ANNs [2] or decision trees (DTs) can be as well realized. In case of DTs, the estimates are Kronecker delta distributions [20], as DTs map a central grapheme with contextual information deterministically onto a phoneme.

IV. EXPERIMENTAL SETUP

We evaluate the proposed method on the English PhoneBook corpus [24]. The G2P conversion task on PhoneBook is difficult as 1) in English the G2P relationship is highly irregular; 2) the corpus contains uncommon English words and proper names (e.g., Witherington, Gargantuan, etc); 3) the number of words in the seed lexicon is relatively small, thus emulates a resource-constrained scenario which makes reliable estimation of $P_{crf}(s_n = f_k|G)$ and $P_{ag2p}(s_n = f_k|g_n)$ really challenging; and 4) the words in the test set are unseen. Furthermore, the reader is pointed to an existing literature [25] that also shows the difficulty of G2P conversion on PhoneBook.

We use the medium size vocabulary task with 602 unique words setup defined for speaker-independent task-independent isolated word recognition in [26]. Table I gives an overview of the dataset. All the words and speakers across train, development and test set are entirely different. The pronunciation lexicon is transcribed using 42 phonemes (including silence).

A. Lexicon generation

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 [27] to determine the G2P alignment. To train and decode the CRF, we used the publicly available CRF++ software [28]. We used bigram features and set the grapheme context to 9, i.e., four preceding and following graphemes as done in [29].

TABLE I: Overview of the PhoneBook corpus.

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

2) Acoustic data-driven G2P conversion approach: To learn the probabilistic G2P relationship, we first trained a 5-layer multilayer perceptron (MLP) using the Quicknet software [30]. The input to the MLP was 39-dimensional PLP cepstral features with four preceding and four following frame context. The MLP output units were 313 clustered CD phonemes derived by clustering CD phonemes in HMM/Gaussian mixture model (GMM) 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 [31]. 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.

3) Multi-stream combination and inference: The weights w_{crf} and w_{ag2p} were estimated by running the multi-stream combination based pronunciation inference on the training data and selecting the one yielding the highest percentage of correct phonemes. In our studies, for the product rule (*Comb-prod*) $w_{crf} = 0.8$ and for the sum rule (*Comb-sum*) $w_{crf} = 0.9$.

For the pronunciation inference, estimation of the prior probability $P(s_n = f_k)$ and the transition probability $P(s_n = f_k | s_{n-1} = f_{k'})$ from the seed lexicon may not be robust, since in the PhoneBook corpus the train and test lexicons are very different and contain uncommon words, and the seed lexicon is relatively small. Therefore, rather than estimating the prior and transition probabilities, we consider the probability distributions to be uniform. With these assumptions, Eqn. (3) can be rewritten as:[1]

$$S^* = \arg\max_{S \in \mathcal{S}} \prod_{n=1}^{N} \overbrace{P(s_n = f_k | g_n)}^{\text{Posterior probability}}. \tag{6}$$

[1] We have indeed ascertained the benefit of a flat prior model over a phone transition model estimated from the seed lexicon through experiments.

B. ASR systems

To evaluate the proposed approach at the application level, in our case ASR, we built CD phoneme-based HMM/GMM system and hybrid HMM/ANN system. The acoustic feature was 39 dimensional PLP cepstral features ($c_0 - c_{12} + \Delta + \Delta\Delta$) extracted using HTK [32]. Following the observations made in [20], we used G2P generated lexicons to train the ASR system, as it yields better systems than the case when trained with manual lexicon and tested with G2P lexicon. The number of tied states were between 2174 and 2270. Each tied state in the HMM/GMM system was modeled by 8 Gaussians. In the case of hybrid HMM/ANN, we trained a five layer multilayer perceptron to classify the tied states using Quicknet [30].

V. Results and analysis

In this section we first present pronunciation level evaluation followed by ASR level evaluations and analysis.

A. Pronunciation level evaluation

Table II provides the pronunciation level evaluation results in terms of number of deletions, substitutions, insertions and phoneme recognition rate (PRR), i.e., 1-`phoneme error rate`. It can be observed that the proposed multi-stream combination method leads to significant improvements at the pronunciation level compared to the acoustic G2P conversion approach. However, it performs worse than the CRF-based approach, mainly due to insertions.

TABLE II: Pronunciation level results in terms of number of deletions (D), substitutions (S), insertions (I) and PRR.

Approach	D	S	I	PRR
CRF	78	364	56	88.5
Acoustic G2P	111	644	245	76.9
Comb-sum	49	379	201	85.5
Comb-prod	52	377	127	87.1

B. ASR level evaluation

Table III presents the ASR level evaluation results in terms of word accuracy (WA), i.e., 1 - `word error rate`. It can be observed that, irrespective of the ASR framework used, the lexicon based on the proposed multistream combination approach leads to the best system. † denotes that the performance gain is statistically significant [33] with 95% confidence interval against the best performing individual G2P conversion approach. The difference between systems using lexicons based on *Comb-sum* and *Comb-prod* rules is not statistically significant. Interestingly, despite performing poor at pronunciation level, acoustic G2P approach when compared to CRF-based approach yields better system in the framework of hybrid HMM/ANN and inferior system in the framework of HMM/GMM. In both cases though the performance is statistically comparable. This trend is more attributed to the fact that acoustic G2P conversion approach typically leads to acoustically confusable substitutions [20], which a discriminative acoustic model (ANN) seems to handle better than a generative acoustic model (GMM). Finally, the best performance of 93.1% is considerably lower than manual dictionary based best system performance of 98.9%. This indicates the difficulty of G2P conversion task.

TABLE III: ASR level evaluations in terms of WA.

	Manual	Acoustic G2P	CRF G2P	*Comb-sum*	*Comb-prod*
HMM/GMM	98.2	88.5	89.2	90.4†	89.9
Hybrid HMM/ANN	98.9	92.7	92.1	93.1	93.1

C. Comparison to combination of lexicons

An alternative approach for exploiting different G2P conversion approaches would be to obtain pronunciation lexicons by combining lexicons generated by the individual G2P conversion approaches. Table IV presents the 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 with two-best pronunciations. It can be seen that ASR systems using the multi-stream combination lexicon perform better than the systems using merged lexicon. Specifically, the differences between the systems using *Comb-sum* and *Acoustic G2P + CRF* lexicons are statistically significant.

D. Analysis

In order to understand if the multi-stream approach is indeed effective, we computed the confusion matrix for the generated pronunciations through each of the approaches. Figure 1 presents the percentage correctly labeled for a few example phonemes. It can be seen that, in most cases, the CRF-based G2P conversion approach is the best individual model. However, there are cases where the acoustic G2P

TABLE IV: Lexical level combination versus multi-stream combination. ‡ denotes that the performance gain is statistically significant with 95% confidence interval.

	Acoustic G2P +CRF	*Comb-sum*	*Comb-prod*
HMM/GMM	91.7	93.0‡	92.4
Hybrid HMM/ANN	94.2	94.9‡	94.4

conversion approach performs better, despite its overall relatively poor PRR. Nevertheless the proposed multi-stream approach is able to perform better than or equal to the best individual models.

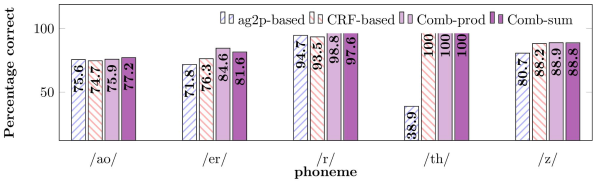

Fig. 1: Percentage correct for few selected phonemes according to the confusion matrix.

Table V presents a few example pronunciations inferred by the different G2P conversion techniques investigated. It can be observed that the multi-stream combination is able to leverage from both the G2P relationship learning techniques.

TABLE V: Pronunciations generated by different G2P conversion approaches along with the manual pronunciations.

Pronunciation	attribution	orion	exorbitant
CRF-based	ae t r aa b uw sh aa n	ao r aa n	aa k s ao r b aa t aa n t
Acoustic G2P	ae t r ay b ah sh aa n	ao r iy aa n	aa g z ao r b aa t ae n t
Combination	ae t r aa b y uw sh aa n	ao r ay aa n	aa g z ao r b aa t aa n t
Manual	ae t r aa b y uw sh aa n	ao r ay aa n	aa g z ao r b aa t aa n t

These analyses show that indeed the multi-stream combination is exploiting the complementarities of the two G2P relationship learning techniques. However, it does not explain the difference in the trend observed at PRR level and ASR level, i.e., at pronunciation level the CRF-based lexicon yields a better PRR than the multi-stream combination based lexicons, but at ASR level it yields inferior performance. One plausible reason could be that PRR is measured with a single manual pronunciation as a reference, while uncommon English words and proper names can exhibit more pronunciation variability. Another reason could also be that the multi-stream G2P conversion is making systematic errors which the ASR system is able to compensate. To further understand that aspect, we examined the pronunciation level errors closely. It can be observed in Table II that low PRR for multi-stream combination is mainly due to insertions. So, we examined the generated pronunciations to investigate the type of insertions. We found that several of the insertions were due to systematic insertion of acoustically close phonemes, such as /axr/ $\rightarrow$ /axr/ /r/ or /ey/ $\rightarrow$ /ey/ /iy/. We speculate that the ASR level trend is a combination of these two factors: pronunciation variation and the ability of ASR system development to handle systematic errors. We aim to investigate it further in our future work.

VI. Conclusion and future directions

Grapheme-to-phoneme conversion can be achieved using different techniques. These techniques primarily differ in the manner the G2P relationship is learned and in the sequential modeling approach employed. The central premise of the present paper is that we can exploit various G2P relationship modeling techniques in order to estimate complementary multiple streams of $P(f|g)$. These streams can then be combined, in a manner analogous to multi-stream speech recognition approach, to improve G2P conversion. We validated the proposed approach by investigating combination of $P(f|g)$ estimates obtained from the CRF-based approach and acoustic data-driven approach. Our studies showed that the lexicons based on the proposed multi-stream approach consistently lead to better ASR systems across different frameworks.

In our future work, in addition to investigating the proposed approach in conjunction with other G2P relationship modeling methods to estimate $P(f|g)$, we intend to focus on unification of acoustic based and G2P conversion based pronunciation model inferences. More precisely, as noted in Section I, in abstract terms A2P conversion and G2P conversion differ mainly in terms of the input. The two techniques can be combined in the same multi-stream formulation where, (a) an acoustic model such as an ANN yields phoneme class conditional probabilities, and (b) the issue related to unequal sequence lengths is handled through dynamic programming.

References

[1] 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.

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

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

[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] 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.

[6] R. Rasipuram and M. Magimai.-Doss, “Acoustic Data-driven Grapheme-to-Phoneme Conversion using KL-HMM,” in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2012.

[7] T. Sloboda, “Dictionary Learning: Performance Through Consistency,” in *Proceedings of ICASSP*, 1995.

[8] M. Ravishankar and M. Eskenazi, “Automatic Generation of Context-Dependent Pronunciations,” in *Proceedings of Eurospeech*, 1997.

[9] 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.

[10] N. Morgan and H. Bourlard, “Continuous Speech Recognition: An Introduction to the Hybrid HMM/Connectionist Approach,” *IEEE Signal Processing Magazine*, pp. 25–42, 1995.

[11] L.R. Rabiner, “A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition,” *Proceedings of IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[12] A. Mohamed, G.E. Dahl, and G. Hinton, “Acoustic Modeling Using Deep Belief Networks,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 14 –22, jan. 2012.

[13] 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.

[14] 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.

[15] A. Janin, D. Ellis, and N. Morgan, “Multi-Stream Speech Recognition: Ready for Prime Time?,” in *Proceedings of Eurospeech*. 1999, ISCA.

[16] H. Misra, H. Bourlard, and V. Tyagi, "New Entropy Based Combination Rules in HMM/ANN Multi-stream ASR," in *Proceedings of ICASSP*, 2003.

[17] F. Valente, "Multi-Stream Speech Recognition Based on Dempster-Shafer Combination Rule," *Speech Communication*, vol. 52, no. 3, pp. 213–222, 2010.

[18] Y. Sun et al., "Combination of Sparse Classification and Multilayer Perceptron for Noise Robust ASR," in *Proceedings of Interspeech*, 2012.

[19] E. Variani, F. Li, and H. Hermansky, "Multi-Stream Recognition of Noisy Speech with Performance Monitoring," in *Proceedings of Interspeech*, 2013.

[20] M. Razavi, R. Rasipuram, and M. Magimai.-Doss, "Acoustic Data-Driven Grapheme-to-Phoneme Conversion in the Probabilistic Lexical Modeling Framework," *Speech Communication*, vol. 80, 2016.

[21] 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*, 2001, pp. 282–289.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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*, 2007, pp. 372–379.

[28] "CRF++: Yet Another CRF toolkit," https://taku910.github.io/crfpp/, Accessed: 2016-02-21.

[29] 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.

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

[31] 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.

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

[33] M. Bisani and H. Ney, "Bootstrap Estimates for Confidence Intervals in ASR Performance Evaluation," in *Proceedings of ICASSP*, May 2004, vol. 1, pp. 409–412.