INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Unsupervised Joint Estimation of Grapheme-to-phoneme Conversion Systems and Acoustic Model Adaptation for Non-native Speech Recognition

Satoshi Tsujioka, Sakriani Sakti, Koichiro Yoshino, Graham Neubig, Satoshi Nakamura

Graduate School of Information Science, Nara Institute of Science and Technology (NAIST), Japan
`{tsujioka.satoshi.tl4, ssakti}@is.naist.jp`

Abstract

Non-native speech differs significantly from native speech, often resulting in a degradation of the performance of automatic speech recognition (ASR). Hand-crafted pronunciation lexicons used in standard ASR systems generally fail to cover non-native pronunciations, and design of new ones by linguistic experts is time consuming and costly. In this work, we propose acoustic data-driven iterative pronunciation learning for non-native speech recognition, which automatically learns non-native pronunciations directly from speech using an iterative estimation procedure. Grapheme-to-Phoneme (G2P) conversion is used to predict multiple candidate pronunciations for each word, occurrence frequency of pronunciation variations is estimated from the acoustic data of non-native speakers, and these automatically estimated pronunciation variations are used to perform acoustic model adaptation. We investigate various cases such as learning (1) without knowledge of non-native pronunciation, and (2) when we adapt to the speaker's proficiency level. In experiments on speech from non-native speakers of various levels, the proposed method was able to achieve an 8.9% average improvement in accuracy.

Index Terms: Non-native Speech Recognition, Iterative Pronunciation Learning, Lexical Modeling, Probabilistic Pronunciation Modeling

1. Introduction

Due to globalization, it is more and more common for speakers to communicate in languages that are not their mother tongue. This need for speakers to obtain non-native language skills has led to development of CALL (Computer Assisted Language Learning) systems that evaluate learners' pronunciation and grammar by using ASR techniques [1]. In these systems, it is necessary to recognize non-native speech accurately.

However, non-native ASR is difficult because the features of non-native speech differ from those of native speech in several ways. Perhaps the most prominent is pronunciation, which is affected by the accent of the speaker's mother tongue or dialect. In addition, non-native speech corpora that can be used for supervised training are limited because it is very expensive and time consuming to collect a large amount non-native speech with the corresponding transcriptions [2–4].

Given this background, there are a number of works on methods to adapt ASR models to non-native speech, which generally focus on adaptation of the acoustic model or the pronunciation dictionary. Methods for acoustic modeling range from standard methods for speaker adaptive training (SAT) such as maximum a posteriori (MAP) [5] and maximum likelihood linear regression (MLLR) [6] specifically tailored for the variance found in non-native speech [7, 8]. In models of the pronunciation dictionary, which we focus on here, early work proposed knowledge-based approaches, which try to find the best pronunciation transformation rules using phonological and linguistic knowledge [9–16]. Data-driven approaches to non-native speech recognition further build upon rule-based approaches by using rule-based methods to generate multiple pronunciation candidates, then using databases of real acoustic evidence to verify which of the pronunciation variations are most appropriate [17–25]. This allows for more accurate estimation of the probability of each pronunciation variation, allowing for more effective use of the pronunciation dictionary at recognition time.

In this work, we focus on two weaknesses of the current approaches to data-driven non-native pronunciation modeling. First, previous work has mainly focused on using non-native data as a way to adjust probabilities of candidates built using either rule-based methods, or estimated from hand-crafted native or non-native pronunciation dictionaries. Rule-based systems and non-native dictionaries require significant effort to create, and data from native speakers is not guaranteed to cover pronunciation phenomena witnessed in non-native speakers. Second, previous work on non-native pronunciation estimation has not covered iterative adaptation of both the dictionary and the acoustic model.

In this paper, we propose a method for unsupervised data-driven iterative pronunciation learning that has the potential to resolve both of these problems. Specifically, we adapt a method [26] that uses a trainable G2P converter [27] to generate multiple latent pronunciation candidates and estimate their occurrence frequency from acoustic data to update the pronunciation lexicon. In order to solve the problem of requiring resources specific to the non-native accent in question, our model creates a seed lexicon directly from acoustic evidence by performing phoneme recognition on non-native speech, and taking phoneme-to-word alignment using Levenshtein distance. In addition, our method is able to perform iterative training to update both the pronunciation dictionary and we adopt SAT included in the iterative training for the acoustic model training, leading to further improvements. The results reveal that the proposed method is able to achieve about 8.9% improvement on average in accuracy for various non-native speakers.

2. Probabilistic Pronunciation Model [26]

2.1. Formulation

A conventional ASR system obtains the optimal word sequence $\hat{\mathbf{W}}$ given the acoustic observations $\mathbf{X}$.

$$\hat{\mathbf{W}} = \underset{\mathbf{W}}{\operatorname{argmax}}\, P(\mathbf{X}|\mathbf{W})P(\mathbf{W}). \tag{1}$$

http://dx.doi.org/10.21437/Interspeech.2016-919

In the probabilistic pronunciation model, Equation (1) can be converted to:

$$\hat{\mathbf{W}} = \operatorname*{argmax}_{\mathbf{W}} P(\mathbf{W}) \sum_{\mathbf{B}\in\Psi_{\mathbf{W}}} P(\mathbf{X}|\mathbf{B})P(\mathbf{B}|\mathbf{W}), \tag{2}$$

where $\mathbf{B} = \{b_1, ..., b_n\}$ denotes a valid pronunciation sequence for word sequence $\mathbf{W} = \{w_1, ..., w_n\}$, $P(\mathbf{B}|\mathbf{W})$ denotes its probability. b_i is the pronunciation of word w_i. $\Psi_{\mathbf{W}}$ denotes the set of all the possible pronunciation sequences of word sequence $\mathbf{W}$. This is decomposed as the product of pronunciation probabilities of each word:

$$P(\mathbf{B}|\mathbf{W}) = P(b_1|w_1)\cdots P(b_n|w_n). \tag{3}$$

When each word has multiple latent pronunciation candidates, they are given weights:

$$P(b_i = \mathbf{p}_j|w_i) = \theta_{ij}, \quad j = 1, \ldots, J_i \tag{4}$$

$$\sum_{j=1}^{J_i} \theta_{ij} = 1, \tag{5}$$

where J_i is the number of alternate pronunciations of w_i, and $\mathbf{p}_j$ denotes one of those pronunciations with a weight θ_{ij}.

2.2. Updating Pronunciation Probabilities

Generally, a G2P converter is used to estimate pronunciation candidates for unseen words. However, G2P converters are not perfect, sometimes predicting incorrect candidates, which can degrade recognition accuracy. In order to solve this problem, Lu et al. proposed a method to obtain correct pronunciation candidates using real acoustic evidence by acoustic data-driven iterative pronunciation learning [26]. This method is called "Supervised G2P" in this paper. The overview of this iterative pronunciation learning method is as follows (Fig. 1):

1. Train the G2P converter on pronunciation candidates in an expert seed lexicon.
2. Use the G2P converter generates multiple latent pronunciation candidates (5 pronunciations for each word) and give equal pronunciation weight to each candidate. These pronunciation candidates are defined as *Initial*, an example of which is shown in Table 1.
3. Train the acoustic model using the *Initial* lexicon.
4. Recognize training speech and obtain the pronunciation used for each word.
5. Update the pronunciation weights by calculating the number of appearances of each pronunciation the recognized word divided by each appearance of the recognized word. When updating the pronunciation weight, prune those pronunciations whose weight is below a threshold. These pronunciation candidates are defined as *Updated*, also shown in Table 1.
6. We also examine the following optional steps:
   - Go to step 3 to re-train the acoustic model with the updated lexicon and re-recognize.
   - Go to step 7 to re-train the G2P converter using the updated lexicon.

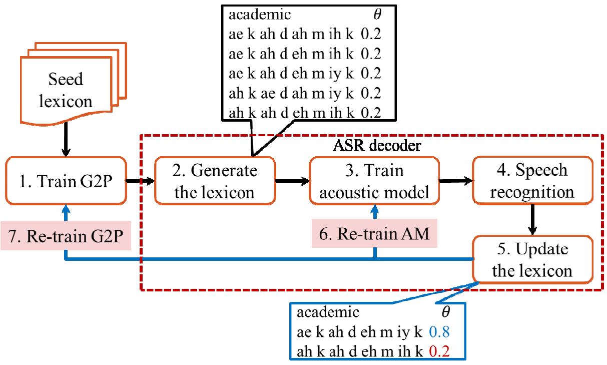

Figure 1: Overview of iterative pronunciation learning [26].

Table 1: Example of pronunciation learning using the probabilistic pronunciation model (threshold: 0.2).

Word	*Initial*	θ	*Updated*	θ
bathroom	b aa th r uw m	0.2	b ae th r uw m	1.0
	b ae th r uw m	0.2		
	b et dh r uh m	0.2		
	b ey dh r uw m	0.2		
	b ey th r uw m	0.2		
academic	ae k ah d ah m ih k	0.2	ae k ah d eh m iy k	0.58
	ae k ah d eh m ih k	0.2	ah k ah d eh m ih k	0.42
	ae k ah d eh m iy k	0.2		
	ah k ae d ah m iy k	0.2		
	ah k ah d eh m ih k	0.2		
trouble	t r ah b ah l	0.2	t r ah b ah l	0.63
	t r ah b ah l iy	0.2	t r aw b ah l	0.37
	t r ah b ah l n	0.2		
	t r aw b ah l	0.2		
	t r aw b ah l n	0.2		

3. Iterative Pronunciation Learning for Non-Native Speakers

3.1. Unsupervised Data-driven G2P

As mentioned in the previous section, the supervised data-driven method requires a seed lexicon. This lexicon will either be a lexicon containing native pronunciations, which may not have the appropriate coverage for non-native speakers, or a rule-based or manually created non-native lexicon that needs to be hand-crafted using the knowledge of linguistic experts.

In this work, we propose a method that is both effective and flexible: using phoneme recognition results of non-native speech to train the G2P converter (Shown in Fig. 2) and use of SAT of acoustic models in the iterative training process. This method has the benefit that it can generate multiple latent pronunciation candidates for the G2P converter that are not constrained by a native seed lexicon, and without explicit knowledge of non-native pronunciation used in rule-based methods. Phoneme recognition can be performed with any acoustic model, but once results are created it is necessary to align the recognized phonemes to words in the transcript. This is done by selecting an alignment with the smallest Levenshtein

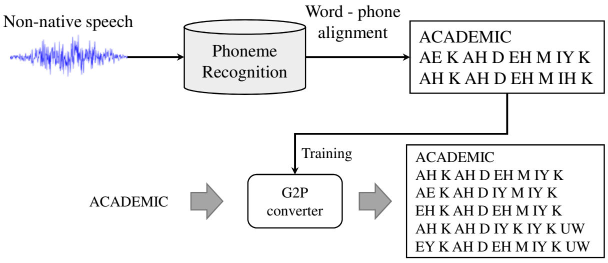

Figure 2: Overview of unsupervised G2P.

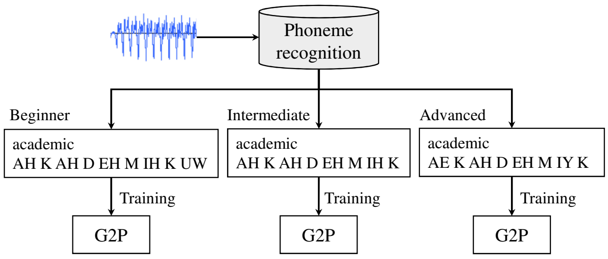

Figure 3: Overview of unsupervised G2P on matched proficiency level.

distance between the recognized phonemes and the estimated phonemes for each word in the transcript, then assigning the recognized phonemes to the corresponding word of the transcript phoneme to which they are aligned. This method is called "Unsupervised G2P" in this paper.

3.2. Unsupervised Data-driven G2P on Matched Proficiency Level/Language

"Unsupervised G2P" used phoneme recognition results of non-native speech to train the G2P converter. However, it has been noted that the pronunciation characteristics of non-native speakers vary depending on their proficiency level or native language [28]. "Unsupervised G2P" does not consider English proficiency or native language, so it cannot learn individual models for speakers with these varying traits.

To solve this problem, we examine a method "Unsupervised G2P on Matched Proficiency Level" that uses phoneme recognition results of non-native speech for speakers of each English proficiency to train the G2P converter. This method also can generate pronunciations that depend on the speaker's proficiency level without explicit knowledge or rules tailored to the specific level. Specifically, we test on low, middle, and high-proficiency speakers, and each method is called "LOW Un-sup G2P," "MID Un-sup G2P," and "HIGH Un-sup G2P." An example of this setup is shown in Fig. 3.

Moreover, we also apply a similar approach to non-native English ASR for various languages. We test on English spoken by native speakers of French, Italian, Greek, and Spanish, which are respectively denoted as "FR Un-sup G2P," "IT Un-sup G2P," "GR Un-sup G2P," and "SP Un-sup G2P."

4. Experimental Evaluation

4.1. Experimental Conditions

Table 2 shows the details of training and test data. In this work, we use a part of the English Read by Japanese (ERJ) corpus [29] for training and evaluation. This corpus contains read English speech data of Japanese college students (190 speakers), with 15,275 utterances for training and 7,952 utterances for testing. Moreover, this speech data is scored for fluency from 1.0 to 5.0 by five native English speakers. The criteria for score evaluation are (1) phoneme generation, (2) rhythm generation, and (3) intonation generation. We take the average of three criteria of score evaluation for each speaker and we divided three English proficiency levels as follows: LOW (scored from 1.0 to 2.5), MID (scored from 2.5 to 3.5) and HIGH (scored from 3.5 to 5.0). For our multi-lingual database, we use the HIWIRE (Human Input that Works In Real Environments) database. This is a database designed to be used as a tool for development and test of speech processing and recognition techniques dealing with robust non-native speech recognition, and contains data from the aforementioned languages. Specifically, it contains 900-3000 utterances from French, Greek, Italian, and Spanish.

Table 2: Experimental data.

Training data		# Speakers	# Utterances	# Hour
WSJ		282	37318	82.9
ERJ	LOW	6	736	1.0
	MID	93	11364	14.3
	HIGH	26	3175	4.2
HIWIRE		57	4560	6.0
Test data		# Speakers	# Utterances	# Hour
ERJ	LOW	5	610	0.8
	MID	40	4889	6.6
	HIGH	20	2453	3.3
HIWIRE		24	480	0.6

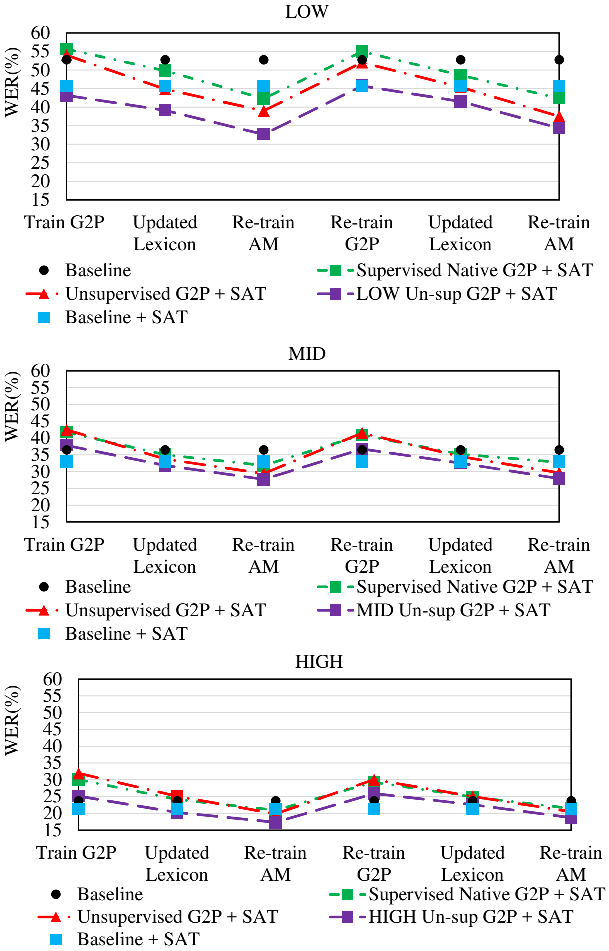

Figure 4: Result of unsupervised G2P on matched proficiency level + SAT.

We use the Kaldi speech recognition toolkit [30] for training and testing of the ASR system. For the acoustic model training data, we use WSJ and part of ERJ corpus (not including the test data). We use a context-dependent GMM-HMM acoustic model with 3-states per phoneme. The acoustic features used 39 dimensional MFCC$+\Delta + \Delta\Delta$, and we also perform feature transformation using LDA (Linear Discriminative Analysis) and MLLT (Maximum Linear Likelihood Transformation) to reduce feature dimensionally. In addition, we evaluate systems that perform SAT using fMLLR (feature-space Maximum Linear Likelihood Regression) [31] to adapt the acoustic model

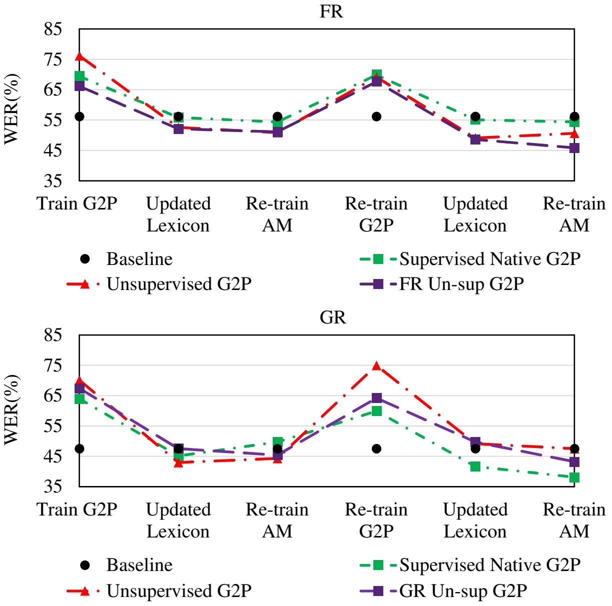

Figure 5: Result of French and Greek.

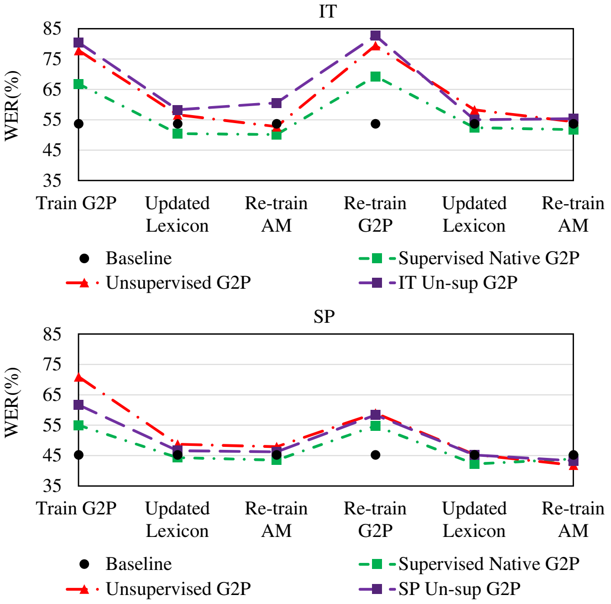

Figure 6: Result of Italian and Spanish.

to non-native speech. For the language model training data, we also use WSJ and part of the ERJ corpus (not including the test data). These language models: 3-gram with Kneser-Ney and Witten-Bell discounting, are generated by the SRILM toolkit [32]. As a native expert lexicon, we use the CMU lexicon, which has about 130K words and a 39 phoneme set. For G2P, we use Sequitur [27], a trainable data-driven G2P tool based on joint n-gram models. The baseline uses a G2P trained on the CMU lexicon for conventional ASR without iterative pronunciation learning. In addition, we compare the results when using our unsupervised seed lexicon with the results of iterative learning using the supervised lexicon as a seed.

4.2. Experimental Results

4.2.1. Results for Speakers of Different Proficiencies

First, we show results for Japanese speakers of differing proficiency levels in Fig. 4. The WERs are 52.8% in LOW, 36.5% in MID, and 23.8% in HIGH for the baseline model (Baseline). In addition, when we apply SAT to the baseline, we achieve WERs of 45.6% in LOW, 33.0% in MID, and 21.3% in HIGH (Baseline + SAT). We perform the evaluation and analysis of each iterative pronunciation learning method using these results as a reference.

First, comparing the results of “Supervised Native G2P+SAT” with the baselines, we can confirm previous work in finding iterative pronunciation learning useful. Next, comparing these results with “Unsupervised G2P+SAT,” we can further see that the proposed method outperforms the supervised native lexicon, particularly for speakers of level LOW. Finally, all of the systems trained on speakers of similar proficiency levels improved 6.1% in LOW, 1.7% in MID, and 2.7% in the HIGH level compared to “Unsupervised G2P.” This indicates the usefulness of training different pronunciation lexicons for speakers of different levels of proficiency.

4.2.2. Results for Speakers of Various Languages

Next, we show results for speakers of various European languages. In this case, the baseline WER of various native languages is 56.1% for French, 47.6% for Greek, 53.8% for Italian, and 45.2% for Spanish. Moreover, we attempted to apply SAT, and the WERs degraded to 66.4%, 52.4%, 62.5%, and 50.0% respectively, likely because the multi-lingual non-native speech data is too small to properly adapt the acoustic model, as the HIWIRE corpus is much smaller than the ERJ corpus. Accordingly, we perform the evaluation and analysis of each method using the Baseline method as a reference.

The results of each language are shown in Fig. 5 and Fig. 6. For the native French Speakers, “FR Un-sup G2P” is the most effective method, likely because the amount of training data is larger (3100 utterances) than that of the corpora in the other languages (a maximum of 2000). However, results in other languages are less conclusive, suggesting that this method will depend on the amount of training data.

5. Conclusions

In this work, we proposed a method to solve the problems caused in non-native speech recognition due to mismatch between native and non-native pronunciation. Specifically, we proposed an acoustic data-driven iterative pronunciation learning method that can be learned directly from phoneme transcripts of non-native speech. This unsupervised method proved beneficial, improving accuracy over a baseline system using supervised G2P, and in several cases improving over an iterative pronunciation estimation method using the supervised dictionary as a seed lexicon. Finally, we found that matching the English proficiency level can help to improve non-native speech recognition accuracy significantly, and that the amount of improvement may be dependent on the data size used in training.

One interesting feature of the proposed method is that it is more effective for non-native speakers of lower proficiency levels. It is conceivable that by combining supervised and unsupervised seed lexicons, we could create a method that works more effectively for all proficiency levels. This is one interesting avenue for future work.

6. Acknowledgements

Part of this work was supported by JSPS KAKENHI Grant Number 24240032 and 26870371.

7. References

[1] C.Chapelle, and J.Jamieson. "Computer Assisted Language Learning as a Predictor of Success in Acquiring English as a Second Language." TESOL Quarterly, vol.20, No.1, pp.27–46, 1986.

[2] L.Besacier, et al. "Automatic speech recognition for under-resourced languages: A survey," Speech Communication vol. 56 pp. 85–100, 2014.

[3] Tanja Schultz and Tim Schlippe, " GlobalPhone: Pronunciation Dictionaries in 20 Languages, " in The 9th edition of the Language Resources and Evaluation Conference (LREC 2014), Reykjavik,Iceland, pp. 26–31, May. 2014.

[4] Tim Schlippe, Sebastian Ochs, and Tanja Schultz, "Wiktionary as a Source for Automatic Pronunciation Extraction, " in The 11th Annual Conference of the International Speech Communication Association (Interspeech), Makuhari, Japan, pp. 26–30, September. 2010.

[5] JL.Gauvain and CH.Lee. "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," Speech and audio processing, IEEE transactions, vol.2, No.2 pp.291-298, 1994.

[6] C.J.Leggetter and P.C.Woodland. "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," Computer Speech and Language, vol.9, No.2, pp. 171–185, 1995.

[7] Z.Wang, T.Schultz, and A.Waibel. "Comparison of acoustic model adaptation techniques on non-native speech," in Proc. ICASSP 2003, vol.1, pp.540–543, 2003.

[8] D.Imseng, R.Rasipuram, M.Magimai.-Doss. "Fast and Flexible Kullback-Leibler Divergence Based Acoustic Modeling for Non-native Speech Recognition,", in Proc. ASRU, pp. 348–353, Dec. 2011.

[9] M.Wester. "Pronunciation modeling for ASR-knowledge-based and data-derived methods," in Proc. Computer Speech and Language vol.17, pp.69–85, 2003.

[10] M.Lehr, K.Gorman, I.Shafran. "Discriminative Pronunciation Modeling for Dialectal Speech Recognition,", in Proc. INTERSPEECH 2014, pp. 1458–1562, Sept. 2014.

[11] M.Finke and A.Waibel. "Speaking mode dependent pronunciation modelling in large vocabulary conventional speech recognition," in Proc. EUROSPEECH 1997, pp. 2379–2382, 1997.

[12] W.Byrne, et al. "Pronunciation modelling using a hand-labelled corpus for conversational speech recognition," in Proc. ICASSP 1998, vol.1, pp.313–316, 1998.

[13] M.Riley, W.Byrne, M.Finke, S.Khudanpur, A.Ljolje, J.McDonough, and G.Zavaliagkos. "Stochastic pronunciation modelling from hand-labelled phonetic corpora,". Speech Communication, vol.29, pp.209–224, 1999.

[14] J.M.Kessens, M.Wester, and H.Strik. "Improving the performance of a Dutch CSR by modeling within-word and cross-word pronunciation variation," Speech Communication, vol.29, pp.193–207, 1999.

[15] M.Wester, J.M.Kessens, and H.Strik. "Pronunciation variation in ASR: Which variation to model?," in Proc. ICSLP 2000, pp. 488–491, 2000.

[16] S.Schaden. "Generating non-Native pronunciation lexicons by phonological rule," in Proc. ICSLP. 2004.

[17] T.Holter and T.Svendsen. "Incorporating linguistic knowledge and automatic baseform generation in acoustic subword unit based speech recognition," in Proc. EUROSPEECH 1997, pp.1159–1162, 1997.

[18] T.Holter and T.Svendsen. "Maximum likelihood modelling of pronunciation variation," Speech Communication, vol.29, pp.177–191, 1999.

[19] I.Amdal, F.Korkmazskiy, and A.C.Surendran. "Data-driven pronunciation modelling for non-native speakers using association strength between phones." ASR2000-Automatic Speech Recognition: Challenges for the new Millenium ISCA Tutorial and Research Workshop (ITRW). 2000.

[20] I.Amdal, F.Korkmazskiy, and A.C.Surendran. "Joint pronunciation modelling of non-native speakers using data-driven methods," In Proc. INTERSPEECH 2000, pp.622–625, Oct. 2000.

[21] J.M.Kessens, C.Cucchiarini, and H.Strik. "A data-driven method for modeling pronunciation variation," Speech Communication, vol.40, pp.517–534, 2003.

[22] S.Matsunaga, A.Ogawa, Y.Yamaguchi and A.Imamura. "Non-native English speech recognition using bilingual English lexicon and acoustic models." in Proc. ICASSP 2003 IEEE International Conference vol.1, pp.625–628, 2003.

[23] M.Kim, YR.Oh, and HK.Kim. "Non-native pronunciation variation modelling using an indirect data driven method," in Proc. ASRU 2007, pp.231–236, 2007.

[24] R.Rasipuram, M.Razavi, M.Magimai-Doss, "Integrated Pronunciation Learning for Automatic Speech Recognition Using Probabilistic Lexical Modeling,", in Proc. ICASSP 2015, 2015.

[25] Tim Schlippe, Wolf Quaschningk, and Tanja Schultz, " Combining Grapheme-to-Phoneme Converter Outputs for Enhanced Pronunciation Generation in Low-Resource Scenarios, " in The 4th Workshop on Spoken Language Technologies for Under-resourced Languages (SLTU 2014), St. Petersburg, Russia, pp. 14–16, May. 2014.

[26] L.Lu，A.Ghosal, S.Renals, "Acoustic Data-driven Pronunciation Lexicon for Large Vocabulary Speech Recognition,", in Proc. ASRU 2013, pp.374–379, 2013.

[27] M. Bisani，H. Ney, "Joint-sequence Models for Grapheme-to-phoneme Conversion," Speech Communication, vol. 50, no.5, pp. 434–451, 2008.

[28] X.Wang, and S.Yamamoto. "Second Language Speech Recognition Using Multiple-Pass Decoding with Lexicon Represented by Multiple Reduced Phoneme Sets." in Proc. INTERSPEECH 2015, pp.1265–1269, 2015.

[29] N.Minematsu, et, al., "English Speech Database Read by Japanese Learners for CALL System Development,", in Proc. LREC 2002, pp.896–903, 2002

[30] D.Povey, A.Ghoshal, G.Boulianne, L.Burget, O.Glembek, N.Goel, M.Hannemann, P.Motlicek, Y.Qian, P.Schwarz, J.Silovsk´ y, G.Semmer, K.Vesel´ y, "The Kaldi Speech Recognition Tool kit," in Proc. ASRU 2011, 2011.

[31] D.Povey, and S.George. "Feature and model space speaker adaptation with full covariance Gaussians." in Proc. INTERSPEECH 2006, 2006.

[32] A.Stolcke. "SRILM-an extensible language modeling toolkit," in Proc. INTERSPEECH 2002, 2002.