G2P variant prediction techniques for ASR and STD

Marelie H. Davel, Charl van Heerden and Etienne Barnard

Multilingual Speech Technologies, North-West University, Vanderbijlpark, South Africa
{marelie.davel, cvheerden, etienne.barnard}@gmail.com

Abstract

Introducing pronunciation variants into a lexicon is a balancing act: incorporating necessary variants can improve automatic speech recognition (ASR) and spoken term detection (STD) performance by capturing some of the variability that occurs naturally; introducing superfluous variants can lead to increased confusability and a decrease in performance. We experiment with two very different grapheme-to-phoneme *variant* prediction techniques and analyze the variants generated, as well as their effect when used within fairly standard ASR and STD systems with unweighted lexicons. Specifically, we compare the variants generated by joint sequence models, which use probabilistic information to generate as many or as few variants as required, with a more discrete approach: the use of pseudo-phonemes within the default-and-refine algorithm. We evaluate results using three of the 2013 Babel evaluation languages with quite different variant characteristics – Tagalog, Pashto and Turkish – and find that there are clear trends in how the number and type of variants influence performance, and that the implications for lexicon creation for ASR and STD are different.
Index Terms: pronunciation variants, speech recognition, spoken term detection, grapheme-to-phoneme

1. Introduction

The same word may be pronounced in different ways for various reasons: differences in semantics, accent, dialect and speaking style all influence the pronunciation of words. These variants can occur in a continuum ranging from generally accepted alternate pronunciations to barely perceptible phonetic variations. While incorporating necessary variants can improve both automatic speech recognition (ASR) and spoken term detection (STD) performance [1, 2], introducing superfluous variants can lead to increased confusability and a decrease in performance [3]. The use of pronunciation variants in state-of-the-art speech-recognition systems has received significant attention in recent years, with a number of studies investigating the effect of variant prediction on either ASR or STD accuracy [4, 5, 1, 3, 6]. We review the most relevant of these below (in Section 2).

In this paper, we are specifically interested in ASR and STD systems that function with unweighted lexicons: that is, where the pronunciation variants themselves are all considered equally likely. This approach is not as well-studied as systems with weighted pronunciation lexicons, even though it is often used in practice, especially where computational efficiency is a concern or pronunciation probabilities are difficult to estimate. We experiment with two grapheme-to-phoneme (G2P) variant prediction techniques that use quite different methods to handle variants: joint sequence models (JSMs) [7] use probabilistic information to generate as many or as few variants as required, while default-and-refine (D&R) [8] first encodes variants in a discrete manner before generalizing from the encoded lexicon in a single, fixed way. We experiment with data from three languages: Turkish, Pashto and Tagalog. These languages have different variant characteristics, as described in section 3.2. We analyze the variants generated, as well as their effect when used within fairly standard ASR and STD systems with unweighted lexicons. Note that we are particularly interested in the modelling of pronunciation *variants* rather than the predictive ability of G2P algorithms in general.

2. Background

In this section we review some of the most relevant studies related to the analysis of pronunciation variants in ASR and STD, and discuss how pronunciation variants are handled by two different G2P prediction algorithms: JSMs [7] and D&R [8].

2.1. Pronunciation variant analysis

Hahn, Vozila and Bisani [4] compared various joint n-gram models on large vocabulary speech recognition tasks in five European languages, in which large pronunciation lexicons (20k to 126k words) were available. As part of their comparison, they investigated whether the addition of pronunciation variants could improve recognition accuracy. Up to three variants could be generated, based on the posterior probability mass of the selected variants. When substantial numbers of words had to be predicted by the G2P system, small gains were achieved in all languages, but in a more natural condition (where out-of-vocabulary words were treated realistically), the variants were slightly detrimental for some languages (and slightly helpful for others). Other approaches to variant generation were not considered in [4].

Jouvet, Fohr and Illina [5] performed a systematic evaluation of ASR accuracy as a function of the degree of pronunciation variability allowed, using both JSMs and Conditional Random Field (CRF) models. The evaluation was performed on French broadcast news data from the ESTER2 corpus [9]. A threshold on the posterior probabilities was used to select the variants retained; in the range of thresholds considered, this procedure produced an average of 1.07 to 1.66 variants per word. The best results were generally achieved when the average number of variants per word is quite low (around 1.1 or less).

For STD, Wang, King and Frankel [1] show that significant gains in performance are achieved when a large number of JSM-generated pronunciations, weighted by their 'corrected' posterior probabilities, are included in a phonetic search lattice. (This correction factor is a linear rescaling of the computed posteriors, with the rescaling parameters computed to optimize the Actual Term-Weighted Value (ATWV) on a validation set.) Since the correction process rescales the pronunciation probabilities, up to 50 pronunciations per word can be allowed; when smaller numbers of variants are allowed, the performance benefits are proportionally smaller.

Whereas the correction factor for pronunciations probabilities in [1] are based on the optimization of an STD measure, McGraw, Badr and Glass [2] employ a data-driven approach to learn these probabilities directly from alignments during ASR. These more accurate probability estimates also allow them to retain a large number of pronunciations per word. On a corpus of English weather requests, [1] reports a significant gain in ASR accuracy over a hand-crafted lexicon when as many as an average of three pronunciations per word are retained. Taken together, these results suggest that both ASR and STD can benefit from the use of pronunciation variants, if a large number of *properly weighted* variants are permitted. However, many systems in practice do not use weighted pronunciation variants; also, considerations of efficiency may conflict with the use of copious pronunciation variants. In the domain of unweighted variants, with a relatively small number of variants per word, it is not clear what the best approach to variant selection is – and also whether such an approach can produce significant performance gains for ASR or STD.

2.2. Variant prediction in JSM

JSMs [7] naturally imply a probabilistic approach to variant prediction, since the JSM training process assigns a likelihood to each of the pronunciations that it hypothesizes. The most likely pronunciation is selected if only one pronunciation is allowed per word, and for multiple variants the likelihoods provide a rank ordering of possible pronunciations. This ordered list is typically truncated when the total probability mass of the selected variants exceeds a predetermined threshold [7].

2.3. Creating pseudo-phonemes

Encoding variants as pseudo-phonemes is a general technique that can be used to incorporate variants in G2P approaches that typically do not allow variants [10]. In particular, a pseudo-phoneme is used to model a phoneme that is consistently realized as two or more variants. These variants are discovered by extracting all the words giving rise to pronunciation variants in an aligned lexicon, and then analyzing these words one grapheme at a time. For each word, any grapheme that can be realized as two or more phonemes (in a set of variant pronunciations of the same word) is considered, and the associated set of phonemes mapped to a new single pseudo-phoneme. If a set of phonemes has been seen before, the existing pseudo-phoneme – already associated with this set – is used. These pseudo-phonemes are then utilized during the extraction of pronunciation rules with no changes required to the underlying G2P algorithm; when predictions are made for lexicon entries, each pseudo-phoneme is expanded to create two or more pronunciations with regular phonemes.

3. Experimental design

3.1. Overview

For this set of experiments, we develop a single ASR and STD system per language using a manually created pronunciation lexicon. We then create different versions of an evaluation lexicon: each version contains exactly the same words, but with pronunciations predicted in different ways. We also produce an 'oracle' lexicon: an evaluation lexicon that consists fully of manually created/verified pronunciations. We compare and analyze the variants obtained using different pronunciation prediction approaches and then report on ASR and STD accuracy.

3.2. Data

The IARPA 2013 Babel corpus is used for all experiments. This corpus consists of both scripted (elicited / read) and conversational speech in several languages; in this paper we focus on Pashto (pus), Tagalog (tgl) and Turkish (tur), and specifically use the babel104b-v0.4aY, babel105b-v0.4 and babel106-v0.2g data releases. For each language, a limited language pack (about 10 hours of transcribed training data per language) and a full language pack (approximately 120 hours of transcribed training data per language) are released. Each training set contains a manually created lexicon covering most of the words in the training corpora. In order to focus on the impact of pronunciation prediction specifically, we create an artificial set-up: we use the limited pack audio data and lexicons during system development, and use the full pack lexicon as our evaluation lexicon. This implies that we only allow words occurring in the full pack lexicon in either our language model or keyword list. As test audio we use the Babel development set and select a subset of the Babel development keyword lists for STD (only those keywords with known pronunciations).

The three languages studied have somewhat different variant characteristics: the Pashto data reflects four different dialects which creates additional variants (a pronunciation per dialect for some words) while Tagalog and Turkish both include a large percentage of words of foreign origin. The Tagalog lexicon includes many words of English and Spanish origin; the Turkish data set words of Arabic and English origin. In Table 1 we analyze the variants occurring within the studied lexicons and show how the average number of pronunciations per word as well as maximum number of variants observed for a single word differ across the languages. The variant percentage provides an indication of the number of pronunciations that are variants per lexicon (if each word has 2 pronunciations, this measure would be 50%).

Table 1: *Number of unique words and pronunciations in the training and evaluation lexicons used, as well as the average number of pronunciations per word (p/w) per lexicon, percentage of lexicon that consists of variants (var %) and maximum number of variants observed for a single word.*

Language	#words	#prons	p/w	var%	max
pus-trn	6 998	9 103	1.30	23.12	9
pus-eval	22 019	28 652	1.30	23.15	12
tgl-trn	6 291	9 789	1.56	35.73	6
tgl-eval	24 650	37 796	1.53	34.78	6
tur-trn	11 992	13 658	1.14	12.2	4
tur-eval	47 327	52 949	1.12	10.62	4

3.3. ASR/STD system

Our baseline ASR/STD systems are fairly standard Kaldi-based [11] systems: triphone models are represented as 3-state left-to-right hidden Markov models, and are tied at the state level using decision tree clustering. Speaker adaptive training (constrained maximum likelihood linear regression) and semi-tied covariance transforms are employed. Standard Mel frequency cepstral coefficients (13 static, 13 delta and 13 delta-delta) with cepstral mean normalization are used as features.

The STD system is lattice based with word posteriors used for confidence scoring. A trigram language model with modified Kneser-Ney discounting [12] was built using SRILM [13]. This language model is used to generate lattices during decoding, which are subsequently used to search for spoken terms.

The language model weights for ASR and STD are fixed for each language and chosen to optimize the baseline systems using the manually developed pronunciation lexicons.

3.4. Performance measures

Standard measures are employed to quantify ASR and STD performance. Word error rate (WER) is used to report ASR performance, whereas the actual and maximum term-weighted values (ATWV, MTWV) as defined in [14] are used to measure STD performance. We calculate WER in the standard way, by dividing the sum of all substitutions, insertions and deletions by the number of reference words in the evaluation set.

Calculating G2P accuracy is not straightforward when variants are involved, and there seems to be no single clear standard for scoring systems where the number of reference pronunciations and hypothesized pronunciations differ. In [4], an hypothesis is considered correct if it equals any one of the given reference variants. Since such a measure does not penalize overgeneration of variants when measuring G2P accuracy, we find it useful to, in addition, report on the percentage of variants generated. Specifically, we calculate the expected percentage of variants on the training lexicon, and then measure the matching variant percentage (M-VAR): the ratio between the obtained and required variants, expressed as a percentage.

Per single hypothesis-reference pair, our G2P phone-based error rate mirrors the ASR word error rate (total number of errors at the phone level, divided by the number of phones in the reference). Per word, we evaluate each reference pronunciation against the best-matching hypothesized pronunciation, and then average over all the reference pronunciations. This results in a variant-based phone error rate per word which can only be 0 if all reference pronunciations are present in the lexicon being evaluated. Phone error rates reported on in this paper make use of this variant-based phone error rate (V-PER). Overall results are obtained by averaging over all word scores, resulting in a score that is not dominated by words with multiple variants, but weighs the contributions of all words equally. While more recently proposed measures weigh pronunciation errors based on the probability of confusion of two phonemes [15], we did not experiment with such weighted measures here.

4. Results

4.1. Non-standard words

The Babel lexicons include a number of 'spelled words' that are clearly marked as such (for example 'B_B_C'); these have pronunciations consisting of spelled out letters. Before performing G2P prediction, all spelled words and unknown one-character words are removed from the set of words to be predicted. (One-character words that are not spelled words typically occur frequently in a language and are therefore included in the training lexicon.) The pronunciations for spelled words are predicted using symbol matching: all spelled word examples are extracted from the training lexicon and character pronunciations obtained through automated matching of the characters with the phone strings using dynamic programming. For all required characters, pronunciations could be obtained this way. The number of variants differ between what can be obtained from the training and evaluation lexicons, namely 22 vs 46 for Pashto, 26 vs 34 for Tagalog and 20 vs 31 for Turkish. These variants are easily handled and excluded from the remainder of the analysis.

Table 2: *Comparison of variant-based G2P phone error rate (V-PER), variant-based G2P word error rate (V-WER), matching variant percentage (MVP), ATWV and MTVW for different variant generation techniques.*

	V-PER	V-WER	MVP	WER	ATWV	MTWV
Pashto						
oracle	0.00	0.00	100.00	69.2	0.1841	0.1963
no-var	8.92	37.52	76.68	72.2	0.1334	0.1584
JSM-var4	3.10	15.57	306.71	72.1	0.1402	0.1582
JSM-mass-100	8.27	36.20	100.12	71.6	0.1453	0.1676
JSM-thr-100	7.07	29.40	99.71	70.4	0.1635	0.1824
JSM-thr-200	3.23	16.01	207.77	70.2	0.1744	0.1892
D&R-pp1	8.15	34.72	106.80	69.7	0.1761	0.1894
Tagalog						
oracle	0.00	0.00	100.00	68.4	0.2602	0.2664
no-var	10.76	43.71	64.94	72.7	0.1899	0.2039
JSM-var4	3.26	15.15	259.74	70.9	0.2247	0.2353
JSM-mass-100	9.20	40.63	100.03	72.7	0.1944	0.2045
JSM-thr-100	6.87	28.06	99.93	69.7	0.2373	0.2400
JSM-thr-200	3.15	15.10	205.33	69.8	0.2446	0.2499
D&R-pp1	7.17	30.76	128.95	68.6	0.2480	0.2533
Turkish						
oracle	0.00	0.00	100.00	68.5	0.2242	0.2284
no-var	2.74	14.50	89.45	70.2	0.2067	0.2140
JSM-var4	0.40	2.28	357.78	71.9	0.2019	0.2056
JSM-mass-100	2.06	11.55	100.45	69.9	0.2125	0.2184
JSM-thr-100	1.76	9.60	101.45	69.2	0.2183	0.2271
JSM-thr-200	0.39	2.21	201.47	69.6	0.2216	0.2284
D&R-pp1	1.55	9.38	101.99	69.0	0.2186	0.2239

4.2. Variant analysis: JSM

We first evaluate the variants predicted using JSMs. We optimise for JSM model order (M=7 for all three languages) and require no restrictions with regard to context (L). We report on both V-PER and M-VAR (as introduced in section 3.4) when:

- no variants are generated (JSM-no-var),
- when the number of variants generated are fixed at various numbers (JSM-vnum-2 to JSM-vnum-12),
- when variants are selected based on the allowed probability mass (JSM-mass-x), and
- when a threshold mechanism is employed: when variants are first over-generated according to a fixed number, and only those variants with a probability higher than a threshold value included (JSM-thr-x).

Since JSM fails to output a transcription if an unknown character is encountered, unseen characters are added manually prior to model training. (This is only required for two characters in Turkish, but results in a fairer comparison.) In Fig. 3 we plot phone error rate against matching variant percentage (M-VAR) for a subset of the techniques evaluated on Pashto (broadly similar trends are observed for Tagalog and Turkish). In order to compare error rates directly, we specifically select experiments where the M-VAR is 100% (indicated by JSM-mass-100% and JSM-thr-100%). Results for selected systems are also included in the first three columns of Table 2. Results for the three languages studied are quite similar: it is clear that the threshold mechanism provides the best V-PER for given M-VAR, and the probability mass method, the worst (of the JSM techniques).

4.3. Variant analysis: D&R

A similar analysis is performed for D&R. All possible pseudo-phones are generated, and subsets created for pseudo-phones that occur n times or more in the training corpus, with n ranging from 2 to 4 (DR-pp-2 to DR-pp-4). The standard D&R algorithm is used, without allowing group formation. In addition,

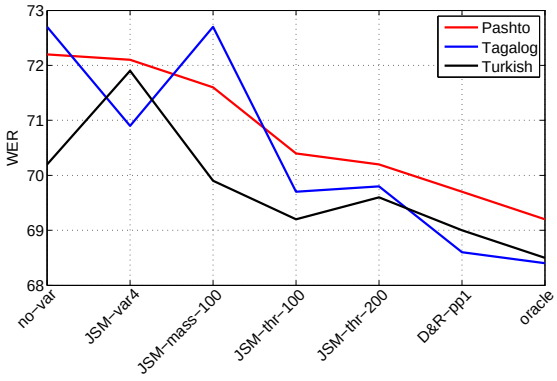

Figure 1: WER for the different strategies.

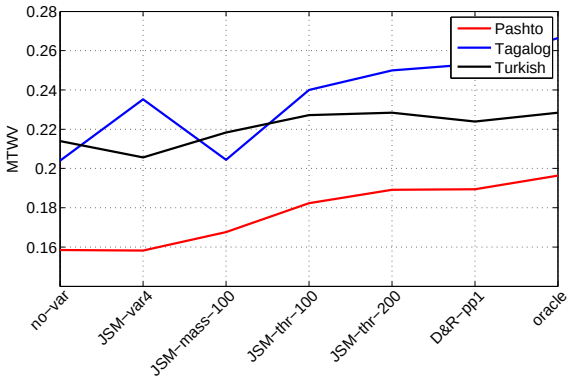

Figure 2: MTWV for the different strategies.

a set of rules are created for a version of the training lexicon where only the first variant (as selected by the JSM technique) is retained. These results are also included in Fig. 4 and Table 2. Compared to the JSM results, three interesting observations are apparent: (1) the D&R techniques perform worse when considering V-PER than JSM, (2) results for DR-pp-2 to DR-pp-4 are very similar, with limited change in either the number or nature of variants generated and (3) the results obtained with JSM-no-var and DR-no-var are startlingly similar. When analyzing the latter observation, it is clear that both algorithms generate almost identical lexicons. This is only the case if JSM is used to select the most probable variant - if random pronunciations are used, this symmetry is not retained.

4.4. Effect on ASR and STD performance

Finally, we consider the effect of the different variant prediction strategies on ASR and STD system performance. Results for selected systems are listed in Table 2. In Fig. 1 and 2 we plot MTWV and WER for the different strategies. An initial ASR and STD system were developed using the oracle lexicon (obtained from the full pack lexicon as described in section 3.2). This system ('oracle') produces the best results across all languages. From Fig. 1 and 2 it can be seen that the trends observed in V-PER for JSM-based techniques match the results obtained: threshold-based selection performs best, resulting in the lowest WER and highest MTWV. Surprisingly, the best results overall (apart from the oracle lexicon) are obtained using the pseudo-phoneme technique. (Note that 'no-var' summarise both 'JSM-no-var' and 'DR-no-var' results, as these were identical at the reported level of precision.)

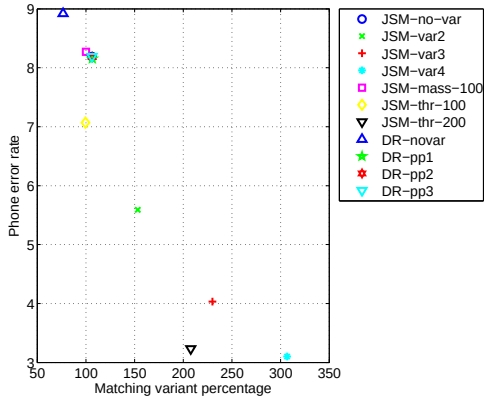

Figure 3: Phone error rate vs matching variant percentage for the different prediction strategies for Pashto.

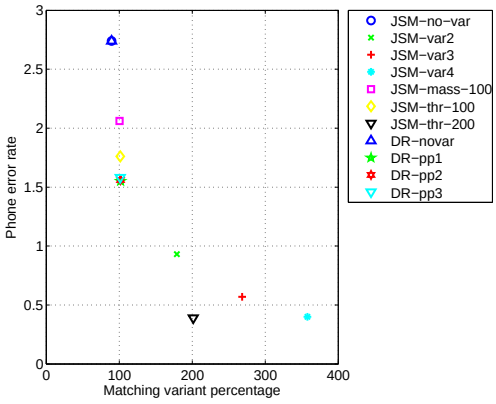

Figure 4: Phone error rate vs matching variant percentage for the different prediction strategies for Turkish.

5. Conclusion

In this paper we review two approaches to variant prediction with G2P algorithms. We show that different types of variants are generated, and that, while both variant behavior and system accuracies are language-specific, some trends emerge. Results for the three languages studied were quite similar: for JSM-based variant prediction, a threshold mechanism provided the best variant-based phone error rate (V-PER) for given matching variant percentage (M-VAR), and the probability mass method, the worst. None of the D&R techniques achieved similarly high V-PER as JSMs, but results were surprisingly similar when comparing JSMs and D&R in the 'no variants' scenario, if the most likely variants were selected by JSM. In future work we would like to analyse this further, and also experiment with JSMs developed for lexicons encoded using the pseudo-phoneme technique, which can also be used with JSMs.

With regard to both ASR and STD with unweighted lexicons, the oracle system (developed using a manually created pronunciation lexicon) produced the best results across all languages. From Fig. 1 and 2 it can be seen that the trends observed in V-PER for JSM-based techniques match the ASR and STD results obtained: threshold-based selection performs best, resulting in the lowest WER and highest MTWV. Although the ASR and STD trends are broadly similar, they are not identical, with ASR systems, for example, performing in a more unpre-

dictable manner when a fixed number of variants are used. Surprisingly, the best overall results (apart from the oracle lexicon) are obtained using the pseudo-phoneme technique: a simple, discrete approach to incorporating variant pronunciations into a general G2P algorithm.

6. Acknowledgement

This work was supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Defense U.S. Army Research Laboratory contract number W911NF-12-C-0013. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either express or implied, of IARPA, DoD/ARL, or the U.S. Government.

7. References

[1] D. Wang, S. King, and J. Frankel, "Stochastic pronunciation modeling for out-of-vocabulary spoken term detection," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 688–698, 2011.

[2] I. McGraw, I. Badr, and J. Glass, "Learning lexicons from speech using a pronunciation mixture model," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 21, no. 2, pp. 357–366, 2013.

[3] M. Adda-Decker and L. Lamel, "Pronunciation variants across system configuration, language and speaking style," *Speech Communication*, vol. 29, no. 2, pp. 83–98, 1999.

[4] S. Hahn, P. Vozila, and M. Bisani, "Comparison of grapheme-to-phoneme methods on large pronunciation dictionaries and LVCSR tasks," in *Proc. Interspeech*, Portland, Oregon, USA, September 2012, pp. 2538–2541.

[5] D. Jouvet, D. Fohr, and I. Illina, "Evaluating grapheme-to-phoneme converters in automatic speech recognition context," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Kyoto, Japan, March 2012, pp. 4821–4824.

[6] R. G. Brunet and H. A. Murthy, "Impact of pronunciation variation in speech recognition," in *IEEE Int. Conf. on Signal Processing and Communications (SPCOM)*, Bangalore, India, July 2012, pp. 1–5.

[7] M. Bisani and H. Ney, "Joint-sequence models for grapheme-to-phoneme conversion," *Speech Communication*, vol. 50, no. 5, pp. 434–451, 2008.

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

[9] S. Galliano, G. Gravier, and L. Chaubard, "The ESTER 2 evaluation campaign for the rich transcription of French radio broadcasts," in *Proc. Interspeech*, Brighton, UK, September 2009, pp. 2583–2586.

[10] M. Davel and E. Barnard, "Developing consistent pronunciation models for phonemic variants," in *Proc. Interspeech*, Pittsburgh, PA, USA, September 2006, pp. 1760–1764.

[11] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Hawaii, USA, December 2011.

[12] S. F. Chen and J. Goodman, "An empirical study of smoothing techniques for language modeling," in *Proceedings of the 34th annual meeting on Association for Computational Linguistics*. Santa Cruz, California, USA: Association for Computational Linguistics, June 1996, pp. 310–318.

[13] A. Stolcke *et al.*, "SRILM-an extensible language modeling toolkit," in *Proceedings of the international conference on spoken language processing*, vol. 2, Denver, Colorado, USA, September 2002, pp. 901–904.

[14] J. G. Fiscus, J. Ajot, J. S. Garofolo, and G. Doddingtion, "Results of the 2006 spoken term detection evaluation," in *Proceedings of ACM SIGIR Workshop on Searching Spontaneous Conversational*, Amsterdam, The Netherlands, July 2007, pp. 51–55.

[15] B. Hixon, E. Schneider, and S. L. Epstein, "Phonemic similarity metrics to compare pronunciation methods," in *Proc. Interspeech*, Florence, Italy, August 2011, pp. 825–828.