COMPUTATIONAL APPROACHES TO EXPLORING PERSIAN-ACCENTED ENGLISH

COREY MILLER
University of Maryland
cmiller6@umd.edu

Abstract
Methods involving phonetic speech recognition are discussed for detecting Persian-accented English. These methods offer promise for both the identification and mitigation of L2 pronunciation errors. Pronunciation errors, both segmental and suprasegmental, particular to Persian speakers of English are discussed.

1. Introduction

Automatic pronunciation error detection (Strik, Truong, de Wet, & Cucchiarini, 2009) is a technique using automatic speech recognition (ASR) technology that can form part of a computer-aided pronunciation training (CAPT) regimen. We outline an approach to automatic detection of pronunciation errors using an off-the-shelf speech recognizer based on the principles of Miller, Strong, Jones, and Vinson (2014). We then describe typical English pronunciation errors by speakers of Persian and show how they can be detected using our approach. We conclude with a discussion of how the technology can be improved and its area of application widened.

2. Automatic pronunciation error detection

Eskenazi (2009) distinguishes between two types of approaches employing speech technology that differentiate native from nonnative speech: pronunciation assessment and individual error detection. Pronunciation assessment attempts to provide a global characterization of fluency over some stretch of speech. In contrast, individual error detection focuses on the pronunciation of particular speech sounds. This paper will focus on a method for individual error detection. An automatic pronunciation error detector can make two kinds of errors: false positives and false negatives (Eskenazi, 2009). A false positive is a case where an error was identified but the pronunciation was correct, and a false negative is a case where no error was identified despite an incorrect pronunciation.

As Krasnova and Bulgakova (2014) note, "usual ASR systems cannot distinguish sounds of a native and a foreign language that are similar, for example, between English aspirated alveolar [t] and Russian dental dorsal [t]". Figure 1 illustrates a "usual" ASR

system, known as speech-to-text (STT). Such a system generates an output text transcript (in a given language) from an input speech file. The reason that STT is language-specific is that it requires a language-specific acoustic and language model. The acoustic model contains representations of the phonemes of the language and the language model is trained on a text corpus of the language. The lexicon mediates between the two by mapping each word to the phonemes of which it is composed. Since STT generally works on a single language at a time, it is not designed to distinguish between native and nonnative renditions of particular targets.

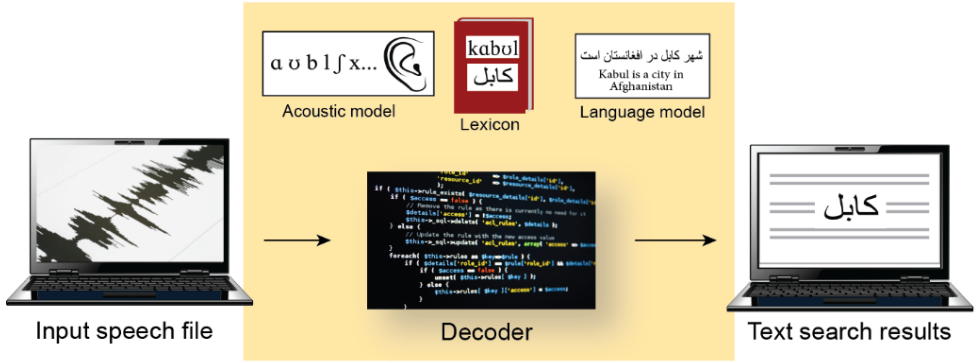

Figure 1. STT in Persian.

Given the inability of STT to distinguish between native and nonnative renditions, alternative approaches must be considered for the purpose of automatic pronunciation error detection. Strik et al. (2009) describe the use of various classifiers trained on particular pronunciation errors. While such an approach shows promise, it suffers from the need to develop a potentially infinite number of classifiers depending on the particular L1/L2 pronunciation mismatch being explored.

Miller et al. (2014) employed the Nexidia phonetic speech recognizer (PSR) (Gavalda & Schlueter, 2010) to perform language and dialect specific audio search of Afghan toponyms in Dari (the Afghan variety of Persian) and Pashto. Figure 2 shows how PSR works. An acoustic model of a particular language in combination with an indexing engine operating on an input speech file creates a phonetic lattice indicating the relative probabilities of the different acoustic targets for each timepoint. During a searching phase, the user can specify a phonetic string and the system will return sequences of timepoints by which the phonetic lattice can be traversed through the sounds of that string above a user-specified level of confidence.

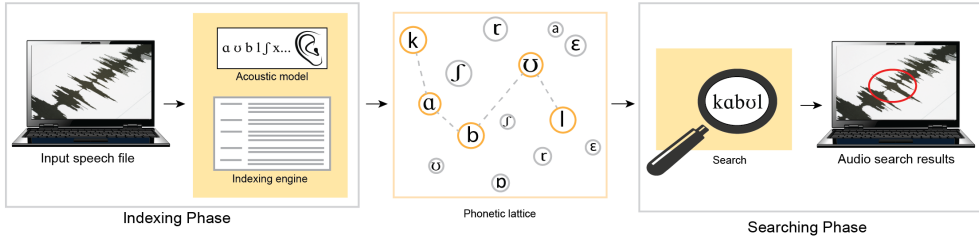

Figure 2. PSR.

Miller et al. (2014) used such a system to search for the Dari and Pashto pronunciations of particular toponyms using both Dari and Pashto acoustic models. Such searches depended on crucial differences in pronunciation of particular speech sounds between Dari and Pashto and their dialects. As expected, Dari pronunciations were matched at higher confidence when using the Dari models and Pashto pronunciations were matched at higher confidence when using the Pashto models.

In the experiments described below on Persian-accented English, we employ both PSR's English models in addition to its multilingual models, which have been trained on a variety of languages (excluding English), and allow for search using a broad set of IPA symbols.

3. Persian-accented English

We consider two major national varieties of Persian, Farsi (spoken in Iran) and Dari (spoken in Afghanistan). While the majority of studies of the Persian accent in English have focused on segmental issues (Aghai, 2011; Hall, 2007; Mirhassani, 2003; Yavaş 2006; Keshavarz & Ingram, 2002; L. Wilson & M. Wilson, 2001), suprasegmental phenomena such as stress (Hayati, 1997), intonation (Soltani, 2007) and syllable structure (Boudaoud & Cardoso, 2009) have also received attention. The related field of native language identification (NLID) should also be noted here – Perkins (2012) employed this for Persian, focusing on text and the clues offered by word choice and syntactic structure. In this paper, we concentrate on a subset of segmental issues and a subset of suprasegmental issues involving syllable structure.

Table 1 summarizes the consonantal differences between Persian and English. The Farsi (Windfuhr, 1997) and Dari (Farhadi, 1955) phoneme inventories contain neither /θ/ nor /ð/, and these are often replaced by /t/ and /d/ in the L2 English of L1 Farsi and Dari speakers (Aghai, 2011). The Persian letter و is pronounced /v/ or /ʊ/ in Farsi and /w/ in Dari. Since there is no /w/ in Farsi, English /w/ often surfaces as /v/ or /ʊ/ among L2 English/L1 Farsi speakers (Aghai, 2011). In contrast, Dari speakers have /w/ in their inventory and would not be expected to have problems with English /w/. Rafat (2010) describes the complex allophony and sociophonetics of /r/ in Persian which may surface as [r] or [ɾ]; neither of which approximate English [ɹ], accounting for yet another possible L1 Farsi/Dari characteristic in English.

L1 English	L2 English/L1 Farsi	L2 English/L1 Dari
θ	t	
ð	d	
w	ʊ, v	w
ɹ	r, ɾ	

Table 1. L2 English consonantal substitutions by L1 Farsi and Dari speakers.

Table 2 summarizes the vowel differences between Persian and English. As can be seen, the Dari vowel inventory is richer than that of Farsi, for example by employing both /i/ and /ɪ/ and /u/ and /ʊ/ in contrast to Farsi's /i/ and /u/. This theoretically would make it easier for Dari speakers to approximate English short vowels.

L1 English	L2 English/L1 Farsi	L2 English/L1 Dari
i	i	i
ɪ		ɪ, ɛ
e	e(j)	e
ɛ	e, æ	ɛ, a
æ	æ	a
ɑ	ɒ	ɒ
ɔ	ɒ, o	ɒ, o
o	o(w)	o
u	u	u
ʊ		ʊ
aw	ɒw	aw
aj	ɒj	aj
ə	ɒ	a

Table 2. L2 English vowel substitutions by L1 Farsi and Dari speakers.

The suprasegmental phenomenon that we examine here involves a phonotactic constraint on syllable patterns whereby Persian prohibits /s/ + consonant clusters (sC), e.g. */sn,sk,sl,sm,sp/. As a consequence, L1 Persian speakers may use e-epenthesis when speaking English, e.g. pronouncing 'snow' [esno], 'school' [eskul] etc.

4. Experiments

In order to explore the traits of Persian-accented English computationally, we used data from the Speech Accent Archive ("SAA", Weinberger, 2015): 8 Dari speakers, 18 Farsi speakers and 2 native American English speakers reading the "Stella" paragraph in English along with IPA transcriptions. Along the lines of Franco et al. (2010), we divide our experiments between using native and nonnative baselines. Franco et al. (2010) used native and nonnative acoustic models; this aspect of their work is reflected in our use of both English models and nonnative multilingual models (trained with a number of languages besides English). When using native models, we search only for native

English phonetic strings. When using nonnative multilingual models, we search for both English and Persian-accented phonetic strings, as shown in Table 3.

Acoustic models	Phonetic strings
Native English	English
Nonnative Multilingual	English
	Persian-accented

Table 3. Combination of acoustic models and phonetic strings.

When searching for English phonetic strings, we expect correctly pronounced English speech to have higher confidence and Persian-accented speech to have lower confidence. When searching for Persian-accented phonetic strings, we expect Persian-accented speech to have higher confidence and correctly pronounced English speech to have lower confidence. At this stage in our research, we have not yet established what the optimal length of the phonetic string to be searched is; PSR's manual mentions that "up to a point" the longer the search string, the better the results (Nexidia, 2013). Another factor currently under experimentation is the effect of grouping various nonnative traits together in a single string. The first set of experiments we report on involve pronunciations of the English interdental fricatives /θ/ and /ð/. Evanini and Huang (2012) report on a set of experiments exploring the pronunciation of English /θ/ by native Mandarin speakers. In contrast to Mandarin speakers who typically substitute /s/ for /θ/, Persian speakers typically substitute /t/ for /θ/ and /d/ for /ð/. Table 4 shows the results of searching for the native English phonetic string /ðiz θɪŋz/ ('these things') on the combined Farsi and Dari corpus (including 2 English controls) from SAA using multilingual models. All utterances returned over confidence level 60 are shown. As can be seen, the phonetic transcriptions provided by SAA do not always match the search string; however, if this approach is to show promise, there will be a good correlation between what is searched and what is retrieved at high confidence. 75% of the utterances returned over confidence 60 contain at least 1 interdental fricative, and 50% have both. Note that neither of the native English speakers' utterances were retrieved with high confidence. We presume this is due to the fact that the multilingual models used in this experiment were not created using native English data, thus reducing potential matches.

Speaker	Transcription	Confidence
farsi9	ðis θiŋz	71
farsi7	diz θɪŋz	67
farsi7	diəs tɪŋs	63
farsi8	ðis θəɪŋs	61

Table 4. Results of searching for native /ðiz θɪŋz/ using multilingual models.

The next experiment can be seen as the inverse: searching for the Persian-accented version of 'these things': /dis tinks/. Note that this phonetic string contains both the substitutions for the interdentals as well as the substitution of /i/ for /ɪ/. All utterances returned over confidence 80 are shown, and all of them have at least one nonnative substitution for an interdental; 60% have both substitutions for the interdentals.

Speaker	Transcription	Confidence
farsi1	dis θiŋks	94
dari3	diz tɪŋz	92
dari2	diz θɪŋs	91
farsi1	dɪs tɪŋks	88
farsi11	dɪz tɪŋz	88

Table 5. Results of searching for nonnative /dis tinks/ using multilingual models.

As is well known, achieving native-like renditions of individual speech sounds is only part of the picture when aiming for native-like pronunciation. Manipulation of connected speech processes (CSPs) is an important element of native production and perception (Alameen, 2014). PSR can be employed to explore the application of CSPs by nonnative speakers. In this experiment we employ the English acoustic models, since it was felt that would enable the highest matching native-like utterances to emerge. The CSP examined here is the nasal assimilation of /ð/ to [n] following [ŋ], as reflected in the phonetic search string [bɹɪŋ niz θɪŋz] 'bring these things'. As seen in the results of this experiment in Table 6, only the native English speakers employed this CSP, and indeed their utterances were matched with the highest confidence.

Speaker	Transcription	Confidence
english160	bɹɪŋ niz θɪŋz	97
english147	bɹɪŋ niz θɪŋs	97
farsi7	bɹɪŋ diəs tɪŋs	97
farsi1	bɹɪŋ dɪs tɪŋks	96
farsi4	bɹiŋ diz θiŋz	94
farsi6	bɹɪŋ ðiz θɪŋz	94
dari3	bɹɪŋ diz tɪŋz	92

Table 6. Results of searching for CSP-influenced [bɹɪŋ niz θɪŋz] using English models.

In the next set of experiments, we explore Dari and Farsi differences in English pronunciation. As mentioned above, Dari has a /w/ in its inventory, but Farsi does not. Using the multilingual models, we first attempt a search for the native English phonetic string /wi wɪl go/ 'we will go'. As seen in Table 7, all of the utterances returned over confidence 60 have at least 1 /w/, and 80% have both /w/'s.

Speaker	Transcription	Confidence
english147	wi wɪl go	86
farsi1	wɪ wil go	66
farsi4	vi wɪl go	65
farsi7	wi wɪl goʊ	65
dari2	wi wɪl go	63

Table 7. Results of searching for /wi wɪl go/ using multilingual models.

We next search for the nonnative phonetic string /vi vɪl go/ ‘we will go’. As can be seen in Table 8, all of the utterances over confidence 70 contain /v/ substituting for native /w/ in both words. In contrast to the search with /w/ above, no Dari utterances were retrieved above confidence 70.

Speaker	Transcription	Confidence
farsi4	vi vɪl go	96
farsi10	vi vɪl go	94
farsi2	vi vɪl gou	78
farsi3	vi vɪl go	75

Table 8. Results of searching for /vi vɪl go/ using multilingual models.

The next experiment explores the syllable structure constraint whereby Persian prohibits /s/ + consonant clusters (sC), e.g. */sn,sk,sl,sm,sp/. Boudaoud and Cardoso (2009) found that English /st/ and /sn/ environments were the most likely to show epenthesis among nonnative Persian speakers, especially following pause or consonants. We thus chose to use the environment ‘call Stella’, in which an /st/ cluster follows the consonant /l/, in order to explore differences in native and nonnative phonological behavior. We first searched for the native-like phonetic string /kɔl stɛlə/ ‘call Stella’ using multilingual models. As can be seen in Table 9, 83% of the utterances retrieved with confidence over 60 followed native norms and did not exhibit epenthesis.

Speaker	Transcription	Confidence
farsi5	kɔl əstɛlə	75
farsi1	kɔl stɛlə	74
farsi9	kol stɛlə	62
dari2	kɔl stɛlə	61
farsi6	kɑl stɛlʌ	61
farsi8	kɑl stɛlʌ	60

Table 9. Results of searching for /kɔl stɛlə/ using multilingual models.

We next searched for the nonnative phonetic string /kɔl ɛstɛlə/ ‘call Stella’, exhibiting epenthesis. As can be seen in Table 10, half of the utterances retrieved over confidence 60 exhibit epenthesis.

Speaker	Transcription	Confidence
farsi5	kɔl əstɛlə	72
farsi2	kɑl istɛla	66
farsi1	kɔl stɛla	63
farsi4	kal əstela	61
farsi9	kol stɛlɑ	60
farsi8	kɑl stɛlʌ	60

Table 10. Results of searching for /kɔl ɛstɛlə/ using multilingual models.

5. Conclusion

We have shown some promising results for detecting nonnative English segmental and suprasegmental characteristics using an off-the-shelf phonetic speech recognition system. While the experiments described here were limited to the English spoken by native Persian speakers, in principle any combination of native and nonnative languages would be possible to investigate using the same process, since the multilingual models allow for arbitrary search using IPA symbols. While the results are not without errors, both false positives and false negatives, we have shown that the results trend in the expected direction; that is, with native phonetic string searches identifying native-like pronunciations with higher confidence and nonnative phonetic string searches identifying nonnative pronunciations with higher confidence.

In order to further refine this process so that it can be useful for teachers and students in identifying pronunciation errors, whether in Persian-accented English or ultimately in any nonnative variety of any language, it will be important both to understand the consequences of employing different search strings better and to refine the scoring methodology.

Regarding the choice of search strings, it will be beneficial to establish the optimal length of a search string, as well as how to deal with both native and nonnative variation which may be possible in search strings of any length. For example, if we are focusing on the /w/~/v/ distinction, it cannot be searched in a vacuum, and as we have seen above, we must make choices about which vowels to include in our search, whether or not we are concerned with their pronunciation at a given stage. Part of this effort will involve more comprehensive scoring, accounting for the full set of false positives and negatives across a wider set of data. Concurrent with improvements in these areas, we hope to explore a wider set of pronunciation phenomena, especially in the area of CSPs, in order to establish the potential of this method.

References

Aghai, L. (2011). The Acquisition of the English Phonological System by Persian Speakers. M.A. Thesis, University of Texas at San Antonio.

Alameen, G. (2014). The effectiveness of linking instruction on NNS speech perception and production. Ph.D. Thesis, Iowa State University.

Boudaoud, M., & Cardoso, W. (2009). The variable acquisition of /s/ + consonant onset clusters in Farsi-English interlanguage. In M. Bowles et al. (Eds.), *Proceedings of the 10th Generative Approaches to Second Language Acquisition Conference (GASLA 2009)* (pp. 86-104). Somerville, MA: Cascadilla Proceedings Project.

Eskenazi, M. (2009). An overview of spoken language technology for education. *Speech Communication 51*, 832-844.

Evanini, K., & Huang, B. (2012). Automatic detection of [θ] pronunciation errors for Chinese learners of English. *Proceedings of the International Symposium on Automatic Detection of Errors in Pronunciation Training*, 71-74.

Farhadi, R. (1955). *Le persan parlé en Afghanistan*. Paris: Klincksieck.

Franco, H., Bratt, H., Rossier, R., Gadde, V. R., Shriberg, E., Abrash, V., & Precoda, K. (2010). EduSpeak®: A speech recognition and pronunciation scoring toolkit for computer-aided language learning applications. *Language Testing 27(3)*, 401-418.

Gavalda, M., & Schlueter, J. (2010). "The truth is out there": Using advanced speech analytics to learn why customers call help-line desks and how effectively they are being served by the call center agent. In A. Neustein (Ed.), *Advances in Speech Recognition: Mobile Environments, Call Centers and Clinics* (pp 221-243). New York: Springer.

Hall, M. (2007). Phonological Characteristics of Farsi Speakers of English and L1 Australian English Speakers' Perceptions of Proficiency. M.A. Thesis, Curtin University.

Hayati, J. (1997). A constrastive analysis of English and Persian stress. *Papers and Studies in Contrastive Linguistics 32*, 51-56.

Keshavarz, M. H., & Ingram, D. (2002). The early phonological development of a Farsi-English bilingual child. *International Journal of Bilingualism 6(3)*, 255-269.

Krasnova, E., & Bulgakova, E. (2014). The use of speech technology in computer assisted language learning systems. In A. Ronzhin et al. (Eds.), *Speech and Computer: 16th International Conference, SPECOM 2014*. Springer.

Miller, C., Strong, R., Jones, E., & Vinson, M. (2014). Employing phonetic speech recognition for language and dialect specific search. In *Proceedings of VarDial Workshop @ COLING 2014*, Dublin.

Mirhassani, A. (2003). *A Constrative Analysis of Persian and English Parts of Speech*. Tehran: Zabankadeh.

Nexidia, T. (2013). North American English Guide. Product documentation.

Perkins, R. C. (2012). Linguistic identifiers of L1 Persian speakers writing in English: NLID for authorship analysis. Ph.D. Dissertation, Aston University.

Rafat, Y. (2010). A socio-phonetic investigation of rhotics in Persian. *Iranian Studies 43(5)*, 667-682.

Soltani, M. A. (2007). *Contrastive analysis of English-Persian intonation*. Azad University-Tehran School of Medicine.

Strik, H., Truong, K., de Wet, F., & Cucchiarini, C. (2009). Comparing different approaches for automatic pronunciation error detection. *Speech Communication 51*, 845-852.

Weinberger, S. (2015). Speech Accent Archive. George Mason University. Retrieved from http://accent.gmu.edu.

Wilson, L., & Wilson, M. (2001). Farsi speakers. In M. Swan & B. Smith (Eds.), Learner English: *A teacher's guide to interference and other problems* (pp. 179-194). Cambridge.

Windfuhr, G. (1997). Persian phonology. In A. S. Kaye (Ed.), *Phonologies of Asia and Africa.Winona Lake* (pp. 663-677). Indiana: Eisenbrauns.

Yavaş, M. (2006). *Applied English Phonology*. Oxford: Blackwell.