Automatic Detection of [θ] Pronunciation Errors for Chinese Learners of English

Keelan Evanini and Becky Huang
Educational Testing Service
Princeton, NJ 08541

***Abstract*—This study examines the types of errors produced by Chinese learners of English when attempting to pronounce [θ] in reading passages and presents a system for automatically detecting these pronunciation errors. The system achieves an accuracy of 79.8%, compared to the inter-annotator exact agreement rate of 83.1%. In addition, speaker-level scores based on the total number of correct productions of [θ] made by each speaker are generated from both the human and machine error annotations, and these are shown to have a strong correlation with each other (0.797).**

I. Introduction

The English voiceless interdental fricative, [θ], is a difficult sound for many non-native speakers to master, and is quite rare cross-linguistically. The difficulty of acquiring a native-like pronunciation of [θ] can be shown by the fact that many long-term learners of English residing in English-speaking countries continue to make this error. In the case of fossilized errors like this, explicit instruction and intense individual practice with the target phone is required to help the language learner achieve a native-like pronunciation. This is thus an ideal application for an automated pronunciation error detection system, since the level of individual attention required to change a learner's behavior would be more than is possible in a typical instructional environment.

In this study, we focus on the specific L1 background of Mandarin Chinese. Studies have shown that Mandarin learners of English have difficulty acquiring [θ] in English and typically use the phone [s] as a substitution [1], [2]. This study focuses on a set of adult speakers of Mandarin Chinese who have been residing in the United States for extended periods, and examines the performance of an automated system for detecting [θ] pronunciation errors on this group of speakers.

There have been many prior approaches to automated pronunciation error detection. The most widespread method uses confidence scores obtained from the ASR system, as in [3] and [4]. Other, more recent approaches, have investigated the use of classifiers based on spectral characteristics, as in [5], or a combination of both approaches, as in [6]. In this study, we adopt a simple approach based on modifying the pronunciation dictionary to contain pronunciations with errors and using forced alignment to select the variant with the highest acoustic score, as in [7]. This approach was used since it can be done relatively easily using open-source capabilities; thus, it has the potential to be used in a wide variety of Computer-Assisted Language Learning applications.

This paper is organized as follows: first, Section II presents the materials that were used in collecting the data for this study and the characteristics of the speakers; Section III describes the annotation procedure that was followed to produce phonetic transcriptions for the learners' tokens of [θ]; Section IV describes the methodology that was used to automatically detect [θ] pronunciation errors; Section V presents analyses of the error detection results; finally, Section VI summarizes the study and describes future related work.

II. Data Collection

This study used three isolated sentences and a paragraph as stimuli. The three sentences were designed by [8] for a foreign accent rating experiment, and each contains one word with the target phone, [θ]. These three sentences are listed below, with the word containing the target phone in bold:

1) *Ron set a **thick** rug in the sun.*
2) *You should **thank** Sam for the food.*
3) *It is fun to play chess **with** a rook.*

In addition, the study also used the *Stella* paragraph [9], a reading passage that is commonly used in accent research. The reading passage contained five instances of the target phone:

> *Please call Stella. Ask her to bring these **things with** her from the store: six spoons of fresh snow peas, five **thick** slabs of blue cheese, and maybe a snack for her brother Bob. We also need a small plastic snake and a big toy frog for the kids. She can scoop these **things** into **three** red bags, and we will go meet her Wednesday at the train station.*

Thus, there were a total of 8 tokens containing the target phone [θ] in this study, and 5 lexical types (*thick*, *with*, and *things* each appeared twice).

36 native speakers of Mandarin (13 male, 23 female) who are long-term residents in the USA participated in the study. All participants arrived in the USA after the age of 18, and all have lived in the USA for a minimum of 7 years (range = 7 - 26; mean = 10, st. dev. = 4). The participants read each of the three sentences and the paragraph out loud twice. Their speech was recorded using a headset microphone (Shure SM 10A) and Audacity (v. 1.2.5) in a quiet location; the audio files were sampled at 16 kHz and saved as uncompressed WAV files. The total number of non-native productions of [θ] investigated in this study is thus 576 (36 participants * 8 tokens * 2 repetitions).

In addition, 22 native speakers of American English (10 male, 12 female) were included in the study as a control group. They read the same materials as the non-native speakers and were recorded under the same conditions. The total number of native productions of [θ] in this study is 352 (22 * 8 * 2).

III. Annotation

Both of the authors of this paper[1] independently listened to the recordings of the sentences and paragraphs produced by the non-native speakers and provided phonetic transcriptions for each of the 576 instances of the target phone. In addition to using perceptual cues to produce the transcriptions, the annotators also incorporated information from the waveform and the spectrogram when the perceptual cues were ambiguous.

The phonetic transcription process revealed that the participants in this study produced a wide range of substitutions to replace the target phone [θ]. In addition to the expected variant [s], the following English phones were also used occasionally as substitutes: [d], [ð], [f], [t], [tʃ], and [z]. Finally, two further sounds which are somewhat harder to characterize were occasionally used as substitutes for [θ]. First, some speakers produced a sound which clearly started out as [s], but then ended with an interdental release (either a stop or a fricative). In these cases, it appeared that the speaker first substituted [s] for [θ], but then became conscious of this mispronunciation and attempted a strategy for correcting it. These tokens are labeled as [sθ], and they only occurred word-initially (i.e. not in the word *with* in this data set). It is likely that their frequency would be much lower in unmonitored spontaneous speech. The other variant that is problematic to describe sounds like an interdental stop. In these cases, there is no sustained portion of aperiodic noise that would be characteristic of a fricative, but the place of articulation sounds quite different from a canonical alveolar stop, [t]. We labeled these interdental stops as [t̪].[2]

TABLE I
Confusion Matrix for Human Annotations

		Annotator BH										
		[d]	[ð]	[f]	[s]	[sθ]	[ʃ]	[t]	[t̪]	[tʃ]	[θ]	[z]
Annotator KE	[d]											
	[ð]		1								3	2
	[f]										1	
	[s]				126	7	2			1	30	
	[sθ]				3	10			1		6	
	[ʃ]											
	[t]		1					3				
	[t̪]	1				1		2	18		6	
	[tʃ]									1		
	[θ]	1	2		28	16		4	35		250	
	[z]				1						1	10

Table I presents the confusion matrix for the two annotators. The inter-annotator exact agreement rate was 72.7% with $\kappa = 0.55$ (the total number of phonetic symbols used in the annotation task was 11). After the two annotators completed annotating the tokens independently, all cases of disagreement were adjudicated by the two annotators together. For the adjudication round, the annotators did not have access to their original annotations, but listened to each audio sample and examined the spectrogram together to come to an agreement. Table II presents the distribution of the annotations on the 576 tokens after adjudication.

TABLE II
Distribution of Annotations after Adjudication

Annotation	Frequency	Annotation	Frequency
[θ]	289	[ð]	4
[s]	166	[tʃ]	2
[t̪]	69	[ʃ]	1
[sθ]	24	[f]	1
[z]	12	[d]	1
[t]	7		

IV. Methodology

To detect pronunciation errors in the non-native productions of [θ], we used the Penn Phonetics Lab Forced Aligner [10]. This open-source forced alignment toolkit is based on HTK [11] and contains monophone acoustic models trained on 25.5 hours of native speech. To model the most frequent type of [θ] error produced by the non-native speakers in this data set, we modified the pronunciation dictionary to include additional pronunciations of the target words containing the phone S instead of TH.[3] For example, the modified dictionary contained the following two entries for the word *thick*:

THICK TH IH1 K
THICK S IH1 K

Then, the recorded utterances were subjected to forced alignment with the stimulus texts using this modified pronunciation dictionary; no modifications were made to the transcriptions to account for disfluencies or reading errors. During the process of forced alignment, the system selects the pronunciation from the dictionary containing either TH or S based on which phone's model is the closest match to the acoustic features. When the forced aligner outputs TH for one of the tokens, this is categorized as a correct pronunciation of the target phone [θ]; alternatively, an output of S for a given token is categorized as a pronunciation error. In the following section, we will compare these machine classifications with the gold standard labels provided by the human annotators.

V. Results

This section presents the results of [θ] pronunciation error detection. However, evaluating the results is not straightforward, due to the fact that the non-native speakers produced a large number of different pronunciation variants for [θ], but

[1]The first author is a native speaker of English with no knowledge of Mandarin and the second author is a native speaker of Mandarin.

[2]This variant is also relatively common in speech produced by native speakers–it occurred several times in a random sample of the responses from the native speaker control group.

[3]Additional experiments were conducted with multiple pronunciation errors in addition to [s] included in the dictionary; however, this approach decreased the performance of the system.

the system only predicts two phones ([θ] and [s]). In Section V-A we first present the results on two different subsets that contain only the two most frequent variants: [θ] and [s]. Then, in Section V-B we present the results for all tokens in order to estimate the performance in an actual application where a decision must be made about every token.

A. Tokens Annotated as [s] and [θ]

As described in Section III, the speakers in this study substituted a wide range of pronunciation variants for the target [θ]. Since the pronunciation error detection system is only designed to classify pronunciations as either the target [θ] or the variant [s], and since the human annotations included several phones in addition to these two, the evaluation of the system's performance is not a straightforward task. Therefore, we first evaluate its performance on the following two subsets of the data containing only adjudicated annotations of [θ] or [s] for which the evaluation task is more straightforward:

- ADJ_TH_S: This subset contains the 455 tokens that received an adjudicated annotation of either [θ] or [s].
- ANN_TH_S: This subset contains the 429 tokens that received an annotation of [θ] or [s] from both annotators during the round of independent annotation (this set is a subset of ADJ_TH_S). This subset was thus intended to only include the tokens which were unambiguous instances of voiceless fricatives so that the system's performance could be examined on the most prototypical cases.

For these two subsets, a direct comparison between the adjudicated annotation and the phone output by the forced aligner is thus possible. Table III presents the inter-annotator agreement and the automatic error detection results for these two subsets of tokens. The inter-annotator agreement statistics were computed by comparing the two sets of independent annotations (before adjudication). The machine detection accuracy results were calculated by comparing the output of the forced aligner with the gold standard adjudicated annotations. The precision and recall values show how well the machine system detected errors; that is, these values were computed with respect to the [s] category.

TABLE III
[θ] ERROR DETECTION RESULTS ON TWO SUBSETS

Experiment	N	Task	% Agree	κ	Prec.	Rec.
ADJ_TH_S	455	human	0.826	0.65	–	–
		machine	0.813	0.60	0.748	0.734
ANN_TH_S	429	human	0.876	0.73	–	–
		machine	0.811	0.59	0.740	0.735

As Table III shows, the performance of the error detection system was similar to the human-human agreement for the ADJ_TH_S subset, but the human-human agreement was higher on the ANN_TH_S subset. The higher human agreement on the ANN_TH_S subset can be attributed to the fact that the tokens included in it were likely more distinct perceptually, since they all received an initial annotation of either [θ] or [s].

B. All Tokens

In this section, we present the error detection results on all of the tokens in the data set. As discussed above, the evaluation of the system's performance on this task is less straightforward, since the set of adjudicated annotations contains more phones than were used by the error detection system. So, it is first necessary to merge the adjudicated annotations into two categories that correspond to the two phones produced by the system. Therefore, the human annotations were divided into a group consisting of *correct* tokens and *errors*. The *correct* category included the target phone [θ] along with the two pronunciation variants that native speakers may also produce: [t̪] and [ð].[4] The *error* category included all other variants produced by the participants, none of which would be expected from a native speaker: [s], [sθ], [z], [tʃ], [ʃ], [f], [d]. The two phones output by the forced aligner corresponded in a straightforward manner to these two categories: TH corresponded to the *correct* category and S corresponded to the *error* category. This experiment in which the token labels were merged to create a binary distinction will be referred to as ALL_BINARY below.

Table IV first presents the human-human agreement results for all of the tokens before they were merged (ALL). Then, the results are presented using the annotations that were merged into the *correct* and *error* categories (ALL_BINARY).

TABLE IV
[θ] ERROR DETECTION RESULTS ON ALL TOKENS

Experiment	N	Task	% Agree	κ	Prec.	Rec.
ALL	576	human	0.727	0.55	–	–
ALL_BINARY	576	human	0.831	0.64	–	–
		machine	0.798	0.56	0.766	0.658

As Table IV shows, the accuracy rate achieved by the error detection system on all of the tokens (0.798) is only 3.3% lower than the exact agreement rate achieved by the two human annotators (0.831). However, the κ value is 8% lower; furthermore, the recall of the error detection system declines substantially when all tokens are included. This indicates that the performance of the error detection system suffered on the categories that were labeled as *error* in the ALL_BINARY set but were excluded from the ANN_TH_S and ADJ_TH_S sets. This higher incidence of false negatives can be shown in the three highest frequency annotations in the *error* category after [s]: the error detection system only predicted 50% of the [sθ] variants as errors (12 / 24), 50% of the [z] variants (6 / 12), and 0% of the [t] variants (0 / 7).

C. Native Speaker Results

As an additional test of the validity of the automated error detection system, it was also applied to the set of native speaker responses. No annotation was conducted for this experiment, since it was assumed that all native speaker tokens

[4]The 4 tokens with the [ð] variant produced by the non-native participants occured word-finally before a vowel in the function word *with*, an environment in which native speakers may also produce the voiced [ð].

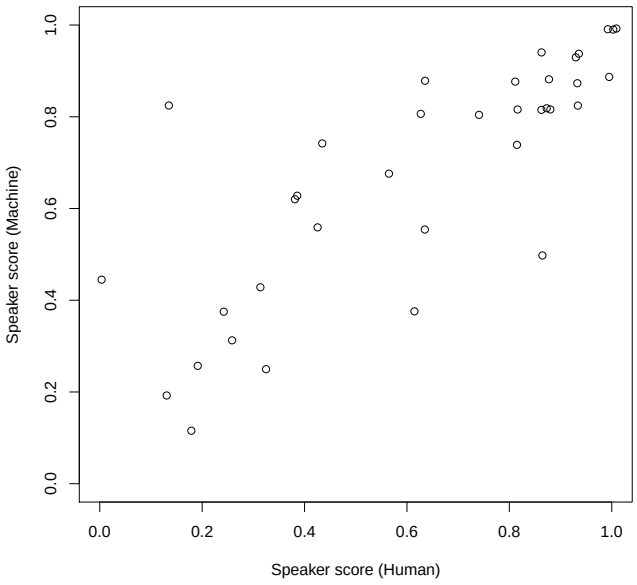
would fall into the *correct* category (i.e., would be one of the following three variants: [θ], [t̪], or [ð]). Out of the 352 tokens in this data set, 337 were classified by the system as TH and 15 were classified as S. Asssuming that none of the tokens should have been classified as an *error*, this amounts to a 4.3% false positive rate for this data set.

D. Speaker-level Results

In order to evaluate the usefulness of the automated error detection system for diagnostic or placement purposes, speaker-level [θ] scores were calculated for each non-native participant in the study based on the percent of *correct* tokens they produced. Scores were produced based on both the human annotations and the machine error detection results by dividing the total number of tokens labeled as *correct* in the ALL_BINARY condition by the total number of tokens produced by the speaker (16). This value thus provides a holistic speaker-level score for each participant's proficiency in producing the phone [θ].

Figure 1 shows that the speaker-level [θ] scores produced by the automated system correspond well with the scores from the human annotations.[5] The Pearson correlation between the speaker-level [θ] scores based on human annotations and those based on machine predictions was 0.797 ($p < 0.001$).

Fig. 1. Speaker-level scores for % of tokens produced correctly



As Figure 1 shows, the speaker with the largest divergence between the human and machine [θ] scores had a machine score of 0.813 (13 / 16 *correct*) and a human score of 0.125 (2 / 16 *correct*). One possible explanation for the poor performance of the error detection system on this speaker is the fact that the audio quality of all of the responses for this speaker was severely degraded by the presence of a constant source of static in the signal. This additional noise in the signal may have caused the spectral characteristics of the speaker's productions of the variant [s] to be more similar to the forced aligner's models for [θ], thus causing a large number of false negatives for this speaker.

[5]The values for points in the figure that are represented by multiple speaker-level scores, such as (1.0, 1.0), were slightly perturbed so that all points could be seen.

VI. CONCLUSION

In this study we have demonstrated that a simple [θ] pronunciation error detection system based on forced alignment with a modified pronunciation dictionary and open-source native-speaker acoustic models achieves a level of performance that is close to the inter-annotator agreement rate for this data set. In addition, we showed that a speaker-level [θ] production accuracy scores based on the automated error detections has a strong correlation with the scores based on human annotations. These results indicate that the error detection system can provide valid feedback to Chinese learners of English in terms of their production accuracy.

The process of pronunciation error annotation and adjudication used in this study provides a rich foundation of knowledge on which analyses of the performance of an error detection system can be based. The fact that several pronunciation variants occurred that were not expected based on the literature suggests that researchers should always use real learner corpora (not artificial errors) and provide detailed transcriptions of their data so they can fully evaluate the performance of their systems. Future research will incorporate more state-of-the-art error detection techniques and apply them to detecting [θ] errors from speakers with a variety of first languages.

REFERENCES

[1] D. V. Rau, H.-H. A. Chang, and E. E. Tarone, "Think or sink: Chinese learners' acquisition of the English voiceless interdental fricative," *Language Learning*, pp. 581–621, 2009.

[2] J. Xiao and Y. Zhang, "A study of Chinese EFL learners' acquisition of English fricatives," in *Proceedings of the 16th Conference of the Pan-Pacific Association of Applied Linguistics*, 2011.

[3] S. Witt, "Use of the speech recognition in computer-assisted language learning," Ph.D. dissertation, Cambridge University, 1999.

[4] H. Franco, L. Neumeyer, M. Ramos, and H. Bratt, "Automatic detection of phone-level mispronunciation for language learning," in *Proceedings of Eurospeech*, 1999.

[5] H. Strik, K. Truong, F. de Wet, and C. Cucchiarini, "Comparing classifiers for pronunciation error detection," in *Proceedings of Interspeech*, 2007.

[6] S.-Y. Yoon, M. Hasegawa-Johnson, and R. Sproat, "Automated pronunciation scoring using confidence scoring and landmark-based svm," in *Proceedings of Interspeech*, 2009.

[7] D. Herron, W. Menzel, E. Atwell, R. Bisiani, F. Daneluzzi, R. Morton, and J. A. Schmidt, "Automatic localization and diagnosis of pronunciation errors for second-language learners of English," in *Proceedings of Eurospeech*, 1999.

[8] J. E. Flege, G. H. Yeni-Komshian, and S. Liu, "Age constraints on second-language acquisition," *Journal of Memory and Language*, vol. 41, pp. 78–104, 1999.

[9] S. Weinberg, "Speech Accent Archive," http://accent.gmu.edu, 2011.

[10] J. Yuan and M. Liberman, "Speaker identification on the SCOTUS corpus," in *Proceedings of Acoustics '08*, 2008.

[11] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book*, 2006. [Online]. Available: http://htk.eng.cam.ac.uk/docs/docs.shtml