Adapting the Acoustic Model of a Speech Recognizer for Varied Proficiency Non-Native Spontaneous Speech Using Read Speech with Language-Specific Pronunciation Difficulty

Klaus Zechner, Derrick Higgins, René Lawless, Yoko Futagi, Sarah Ohls and George Ivanov

Educational Testing Service
Princeton, NJ, USA
{kzechner,dhiggins,rlawless,yfutagi,sohls,givanov}@ets.org

Abstract

This paper presents a novel approach to acoustic model adaptation of a recognizer for non-native spontaneous speech in the context of recognizing candidates' responses in a test of spoken English. Instead of collecting and then transcribing spontaneous speech data, a read speech corpus is created where non-native speakers of English read English sentences of different degrees of pronunciation difficulty with respect to their native language. The motivation for this approach is (1) to save time and cost associated with transcribing spontaneous speech, and (2) to allow for a targeted training of the recognizer, focusing particularly on those phoneme environments which are difficult to pronounce correctly by non-native speakers and hence have a higher likelihood of being misrecognized. As a criterion for selecting the sentences to be read, we develop a novel score, the "phonetic challenge score", consisting of a measure for native language-specific difficulties described in the second-language acquisition literature and also of a statistical measure based on the cross-entropy between phoneme sequences of the native language and English.

We collected about 23,000 read sentences from 200 speakers in four language groups: Chinese, Japanese, Korean, and Spanish. We used this data for acoustic model adaptation of a spontaneous speech recognizer and compared recognition performance between the unadapted baseline and the system after adaptation on a held-out set from the English test responses data set.

The results show that using this targeted read speech material for acoustic model adaptation does reduce the word error rate significantly for two of four language groups of the spontaneous speech test set, while changes of the two other language groups are not significant.

Insdex Terms: acoustic model adaptation, non-native spontaneous speech, cross-lingual phonetic difficulty

1. Introduction

Recognizing non-native speech has proven itself to be significantly harder than native speech, mostly due to a wider variation in many elements of natural language, such as fluency (e.g., speaking rate, disfluencies), pronunciation, pauses, stress and intonation, vocabulary choice and grammatical structure.

The goal of this research is to accurately recognize spontaneous responses from candidates of a test of English, the Test of English as a Foreign Language (TOEFL®) Academic Speaking Test (TAST), a pre-cursor to the Speaking section of the TOEFL internet-based test ((TOEFL iBT).

In this scenario, the challenge of non-native ASR is greatly magnified by at least three exacerbating factors:

1. The speakers' proficiencies vary widely between almost native-like and completely incomprehensible.
2. The speakers come from many different native languages with very uneven frequency distributions.
3. The recordings were made over the phone, i.e. we have to use a narrow-band 8 kHz recognition system.

All these factors contribute to the rather high word error rate (WER) of 65.9% for the baseline system, measured on an independent test set. (We observed WERs of similar magnitude in other English test corpora, as well [1]). In this context, we are looking at ways to reduce the WER by using acoustic model (AM) adaptation.

For acoustic model adaptation, a variety of methods have been used in previous research, such as Maximum A Posteriori (MAP) adaptation, Maximum Likelihood Linear Regression (MLLR), Singular Value Decomposition (SVD), and phone confusion matrices, among others (e.g., [2], [3], [4]).

While normally one would use spontaneous speech for AM adaptation in our situation, the purpose of this paper is to investigate to what extent a carefully selected sample of read speech (in otherwise equivalent acoustic conditions) might be a possible choice for AM adaptation as well.

The two main arguments for using read as opposed to spontaneous speech are (1) the substantial savings in human transcription time and labor, and (2) the possibility of targeting specific L1-dependent phonetic contexts that typically pose challenges for the learner of English and hence may often be misrecognized by the ASR system.

Against this background, [5] presented a study that indicates that phonetically-rich read speech can indeed be successfully used for acoustic model adaptation for a spontaneous speech corpus. However, their corpus of mostly very short utterances in a dialog system is quite different from our spontaneous speech corpus where test candidates talk freely for up to a minute on a given subject. This study should demonstrate whether effects similar to those reported by [5] can be observed for spontaneous speech corpora with much longer utterances and much more diverse vocabulary. While [5] used phonetic strings that are difficult in general, we selected sentences on the basis of native language-specific phonetic difficulty.

In this paper, we chose to take a new approach which targets phonetic environments that pose certain levels of difficulty for non-native speakers.

In a first step, we develop a language-specific "phonetic challenge score" which we validated by human rankings of sentences with varying pronunciation difficulty. This allows a native language-specific and targeted adaptation of the recognizer where the focus is on phonetic environments which are difficult to pronounce for a non-native speaker of a particular native language and which therefore are also more likely to be misrecognized in the absence of this adaptation (Section 2)

The second step is to select sentences according to this criterion from material designed for classroom reading.

Thirdly, we recruit students in four groups with the following native languages: Chinese[1], Japanese, Korean and Spanish. All participants had to read 120 sentences each to a phone-based automated interactive voice response (IVR) recording system (Section 3).

Finally the speech samples are used for the acoustic model adaptation experiment (Section 4).

2. Phonetic challenge score

We developed the phonetic challenge score as a means to select reading material with native-language-specific phonetic difficulty which then can serve to adapt the AM of a speech recognizer.

The phonetic challenge score combines two disparate sources of information.

The first information source, which we will refer to here as the rule-based score, is derived from information in the linguistics and language pedagogy literature, which records specific difficulties actually encountered by second-language learners (see e.g. [6] [7] [8] [9]).

The second information source, the phoneme language model score, is a measure of how much an utterance in the second language (L2—in this paper this is always English) varies from the phonotactics of the speaker's native language (L1). It is derived from a statistical model of the sound structure of the two languages.

The rule-based score is important because it is tied to actual error types that language learners of a given L1 background commit in the L2. However, the rule-based score is somewhat limited, because the language pedagogy literature does not quantify the frequency of each error type, or tell us the relative importance of multiple sources of difficulty in an utterance. The phoneme language model score, because it is a corpus-based measure, is much more specific in its judgments, and covers many sources of difficulty which may be too infrequent to garner mention in the literature.

The rule-based score is a sum of the weights assigned to specific "difficulty triggers"[2], and the phoneme language model score is composed of the cross-entropy of the utterance with respect to the L2 phoneme language model, inversely weighted by the cross-entropy of the utterance with respect to the L1 phoneme language model (Kullback-Leibler (KL) divergence). The formula is given below, where D is the KL divergence, H is the cross-entropy, u is an utterance in the L2 (English), and $\mathrm{P}_{\mathrm{L1}}[]$ / $\mathrm{P}_{\mathrm{L2}}[]$ are the probabilities assigned to an utterance by the L1/L2 language models (Equation 1).

The statistics for computing the phoneme language model scores for individual languages are derived from phonetically transcribed corpora of spoken language. The corpora used to estimate the phoneme language models are taken from the CALLHOME set of telephone conversation transcripts (available from the Linguistic Data Consortium, LDC), with the exception of Korean, for which telephone transcripts from the Callfriend project were used (also available from LDC).

$$\begin{aligned} \mathrm{D}\left(u, \mathrm{P}_{\mathrm{L1}}[\cdot] \,\|\, \mathrm{P}_{\mathrm{L2}}[\cdot]\right) &= \mathrm{H}\left(u, \mathrm{P}_{\mathrm{L1}}[\cdot]\right) - \mathrm{H}\left(u, \mathrm{P}_{\mathrm{L2}}[\cdot]\right) \\ &= -\log\left(\frac{\mathrm{P}_{\mathrm{L1}}[u]}{\mathrm{P}_{\mathrm{L2}}[u]}\right) \end{aligned}$$

We used approximately 150,000 words from each language to train the model.

Both scores are normalized to have a mean of zero and a standard deviation of 1. The two normalized scores are then added to yield the phonetic challenge score.

It is not entirely straightforward to apply a phone-based language model derived from one language to another language, because certain phones may not be represented in the other language at all, or else the mapping between phones may fail to be one-to-one (for example, if contrastive phones in one language are variants of a single phoneme in the other). We deal with the problem of phones not occurring in the L1 by assigning such phones a small backoff probability proportional to their frequency in English. To deal with unclear mappings between phones, we match them as well as possible, following linguists' judgments and the constraint that every English phone must map to only one phone in the L1 language model (so that the utterance's probability can be calculated unambiguously).

As an illustration of the sources of pronunciation difficulty targeted by these methods, Table 1 shows examples of difficulty triggers used in the rule-based score for different L1s.

Table 1. Examples for weighted difficulty triggers as part of the rule-based score for four native languages.

Native Language (L1)	Difficulty trigger	Weight
Chinese	[r]/[l]	1
	Final stop consonant	2
Japanese	Voiceless stops (shorter voice onset time)	0.2
	Complex syllable coda	5
Korean	[θ] (as in *thing*)	1
	Complex syllable coda	5
Spanish	[θ] (as in *thing*)	1
	Stop consonant between vowels	5

We performed a small-scale evaluation of the phonetic challenge score to determine how well it accorded with the judgments of native speakers of different L1 backgrounds about the relative pronunciation difficulty of specific English sentences.

For each of the languages we modeled as a potential L1 (Spanish, Japanese, Mandarin Chinese, and Korean), we selected 20 sentences by phonetic challenge score from the Lexile corpus, a large corpus of English narrative texts (cf. [10]). For each language, 10 sentences were chosen with a relatively high phonetic challenge score and 10 with a relatively low score. Thus, for each L1 we had two sets of 10 English sentences which we call L1-hard and L1-easy (eg. Spanish-easy). Our evaluation focused on the degree to which our experts' rankings of pronunciation difficulty

[1] "Chinese" refers to "Mandarin Chinese" in this paper.

[2] The weights were determined manually based on expert judgments of the relative importance of different difficulty triggers.

reproduced the division between L1-hard and L1-easy sentences that emerged from the phonetic challenge model.

The result of this experiment was that, with the exception of the Chinese set, the phonetic challenge score was significantly correlated with the native speaker judgments at the 0.05 level for each set of sentences. (The Chinese set came close to significance with $p<0.1$).

3. Building a read sentence corpus

3.1. Sentence selection

For every language group, we composed 10 distinct test forms with 120 sentences each. The sentences were randomly selected, with a mean length of 11 words, from the Lexile corpus.[3] The phonetic challenge score was used to pick 40 sentences each from the difficulty levels "easy," "medium," and "hard." The sentences were filtered through a list of criteria, some content-based (eg. "no mention of violence"), some grammatical (eg. "must have subject and verb"), and some orthographical (eg. "no misspellings"). Sentences not passing the filters were replaced with equivalent randomly selected sentences until all sentences on all test forms met all criteria.

3.2. Data collection

Participants residing in the U.S. were recruited over the Internet and assigned to one of four language groups based on their native language[4]. They had to place a toll-free call to an IVR system that prompted them to read 120 sentences from a randomly assigned test booklet that the participants received via email. After successful completion they received Internet gift certificates of $25 each. Some participants had to be excluded from the analyses post-hoc when it was discovered that they were in fact native speakers of English or that they had participated more than once.

We recruited a total of 200 speakers for this study who read over 23,000 sentences of (accented) English, their native languages being Chinese (85 speakers, all with native dialect Mandarin), Japanese (16), Korean (35), and Spanish (64). We had participants self-rate their English proficiency on a 5-point non-parametric scale; when mapping the levels to 1 (least proficient) to 5 (most proficient), the average proficiency for the five language groups varies from 2.0 to 2.5.

3.3. Cost and time considerations

In order to obtain high quality transcriptions of non-native spontaneous speech, our experience shows that about 10 hours of transcriber time are required per hour of speech. With a typical cost of $50 per hour, we arrive at a ballpark estimate of $500 for transcription of one hour of non-native spontaneous speech.

In contrast, data collected for the read sentence corpus does not have to be transcribed, but unlike in the case for non-native speech where we can take speech from prior test administrations, we had to pay subjects contributing to the read sentence corpus. Typically we collected about 10 minutes of speaking time per candidate with an associated cost of $25. Therefore, an hour of speaking time (read speech) costs in the neighborhood of $150. If one makes the argument that some transcription services may only charge half as much per hour (for a less detailed and less thorough transcription), we could at the same token argue that $10 would be a sufficient compensation for most students residing in the U.S. for half an hour of their time.

In summary, the collection of the read sentence corpus was by a factor of 2–4 cheaper than a transcription of an equivalent-sized non-native spontaneous speech corpus would have been. For our adaptation corpus of 24 hours, the cost could be reduced from an estimated $12,000 (if we had transcribed non-native spontaneous speech) to $5,000.

4. Experiment

4.1. Data

We use a corpus of more than 1,000 spontaneous speech responses of 179 candidates of the TAST s(TOEFL Academic Speaking Test), collected over the phone. The speakers come from over 40 native language backgrounds. Each speaker provided 6 responses of 45-60 seconds each with a total speaking time per speaker of around 5.5 minutes.

The audio was sampled at 8kHz, mono, 8 bit resolution, mu-law encoding. The complete corpus was transcribed for ASR training and evaluation. About 70% of the data was used for ASR training (124 speakers) and the rest as an independent held-out test set (55 speakers). The total corpus comprises about 16 hours of speech. From the TAST test set we extracted all speakers of the four languages Chinese, Japanese, Korean and Spanish. Table 2 provides the number of speakers and hours of speech in this test data which we use for our experiment.

Table 2. Number of speakers and hours of speech for each language in the TAST test set used for the AM experiment.

Language	Speakers	Hours of speech
Chinese	10	0.9
Japanese	5	0.5
Korean	6	0.6
Spanish	4	0.4
TOTAL	25	2.3

For AM adaptation, we use the Read Sentence Corpus described above. with a total of 200 speakers and over 23,000 read sentences. The distribution of speakers and hours of speech across languages is shown in Table 3.

4.2. Speech recognizer

The ASR system, a standard gender-independent fully continuous Hidden Markov Model system, was bootstrapped from a recognizer based on Linguistic Data Consortium (LDC) SwitchBoard and CallHome data using the TAST train set for AM training and a combination of this set with LDC Broadcast News data for LM building. The WER on the held-out test set was 65.9% which is high but typical in a setting with many different native languages and a wide spectrum of English proficiency.

4.3. Acoustic model experiment

In a first step, we built MAP adapted acoustic models for each of the four languages, using the data from the read sentence corpus (see Table 3).

[3] See http://www.lexile.com/

[4] Candidates with a native language other than the four we needed for this study were placed in a separate group for future use.

Then, we ran the recognizer for each language set twice, one time in baseline mode with the original unadapted AM, then with the language-specific adapted AM.

The last two columns of Table 3 report the results of this experiment. The WER is computed by first adding up all counts for insertions (I), deletions (D), substitutions (S) and correct words (C) and then using the standard formula of WER=(S+D+I)/(C+S+D).

Table 3. AM adaptation experiment: corpus sizes of Read Sentence Corpus and WER results for baseline and after AM adaptation on TAST test set. (* indicates significant difference in WER at p<0.05, performing a 2-tailed matched-pair t-test on each utterance WER[5])

	Read Sentence Corpus		TAST test data set	
Native Language	Speakers	Hours of Speech	Baseline	Using AM adaptation
Chinese	85	11.0	0.668	0.638*
Japanese	16	2.0	0.694	0.653*
Korean	35	3.8	0.746	0.744
Spanish	64	7.2	0.573	0.588
TOTAL	200	24.0	0.672	0.655*

4.4. Discussion

From Table 3 we can see that the overall WER was reduced significantly (by 1.7% absolute), and among the four languages we see Chinese and Japanese with significantly reduced WER (by 3.0% and 4.1% absolute, respectively), whereas there was no significant change observed for Korean and Spanish. While the WER for Korean is stays about the same pre- and post-adaptation, the WER for Spanish rises slightly, albeit not significantly. Our conjecture for the worse performance for Spanish is that the average proficiency score for Spanish was the highest among all four languages in the test set (which may also explain the markedly lower absolute WER) and so our adaptation sets geared towards frequent mispronunciations may have distorted the acoustic models too much in the face of the higher proficiency candidate population.[6]

While the better performance of Chinese AM adaptation could also be explained by the fact that it had the largest adaptation set, the performance of Japanese is the most interesting as we had only 2 hours of speech available for AM adaptation and still obtained the highest WER reduction of all sets.

While the overall results are promising with significant WER reductions both overall and for two of four languages, the absolute WER decrease is rather small, particularly given the high initial WER in the baseline. We conjecture that the more traditional approach of collecting more spontaneous speech data, transcription and then AM adaptation may likely have reduced the WER more substantially, but on the other hand would also have been significantly more costly in terms of time and labor of the human transcription process.

In summary, we argue that despite the mismatch between read speech and spontaneous speech, we were able to demonstrate an initial proof-of-concept of the idea to use read speech with L1-specific targeted phonetic difficulty for AM adaptation of a non-native ASR system for spontaneous highly accented speech.

We also deem it likely that a combination of both data sources – spontaneous speech and language-specific targeted read speech - might yield the highest reduction in WER overall.

5. Conclusions and future work

We presented an approach to acoustic model adaptation of a spontaneous speech recognizer for heavily accented non-native speech using a corpus of read speech, where the sentences were selected on a phonetic difficulty criterion, the native-language-specific phonetic-challenge score. We collected over 23,000 sentences, read by 200 non-native speakers of English and then used this corpus to do MAP acoustic model adaptation of a recognizer trained for spontaneous speech.

Our results show that we can significantly reduce the WER for two of four language groups by using recordings of read sentences of unrelated textual material, but with native language-specific phonetic make-up. Also, the overall WER for the entire test set was reduced significantly, as well.

Future work includes experiments with combining different sets of AM data, in particular spontaneous speech data and read speech data.

6. References

[1] Zechner, K., Higgins, D., and Xi, X., "SpeechRater^{SM}: A construct-driven approach to score spontaneous non-native speech", Proceedings of the First SLaTE Workshop (Spoken Language Technology in Education), Farmington, PA, 2007.

[2] Bouselmi, G., Fohr, D., Illina, I., and Haton, J.P., "Fully automated non-native speech recognition using confusion-based acoustic model integration", Proceedings of Interspeech-2005, 1369-1372, 2005.

[3] Deng, Y., Li, X., Kwan, C., Raj, B., and Stern, R., "Continuous feature adaptation for non-native speech recognition", International Journal of Signal Processing, 3(1), 2006.

[4] Wang, Z., and Schultz, T., "Nonnative spontaneous speech recognition through polyphone decision tree specialization", Proceedings of the 8th European Conference on Speech Communication and Technology (Eurospeech-2003), 1449-1452, 2003.

[5] Sturm, J., Kamperman, H., Boves, L., and Os, E.D., "Impact of speaking style and speaking tasks on acoustic models", Proceedings of the ICSLP-2000, 2000.

[6] Arslan, L.M., and Hansen, J.H.L., "Language accent classification in American English", Speech Communication, 18(4):353-367, 1996.

[7] Arslan, L.M., and Hansen, J.H.L., "A study of temporal features and frequency characteristics in American English foreign accent", Journal of the Acoustical Society of America, 102(1):28-40, 1997.

[8] Eckman, F.R., Elreyes, A., and Iverson, G.K.., "Some principles of second language phonology", Journal of Second Language Research, 19(3):169-208, 2003.

[9] Magen, H.S., "The perception of foreign-accented speech", Journal of Phonetics, 26(4):381-400, 1998.

[10] Deane, P., "A nonparametric method for extracting candidate phrasal terms", Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics, 2005.

[5] A speaker's response has on average about 2 utterances.

[6] The average proficiency scores on the TAST test set (discrete 4-point scale from 1 to 4, 4 is most proficient) were: 2.5 for Chinese, 2.0 for Japanese, 1.8 for Korean and 3.7 for Spanish.