Real Voice and TTS Accent Effects on Intelligibility and Comprehension for Indian Speakers of English as a Second Language

Frederick Weber [1], *Kalika Bali*[2]

[1] Earth Institute, Columbia University, New York, USA
[2] Microsoft Research India, Bangalore, India
fw2174@columbia.edu.edu, kalikab@microsoft.com

Abstract

We investigate the effect of accent on comprehension of English for speakers of English as a second language in southern India. Subjects were exposed to real and TTS voices with US and several Indian accents, and were tested for intelligibility and comprehension. Performance trends indicate a measurable advantage for familiar accents, and are broken down by various demographic factors.

Index Terms: speech synthesis, accent, intelligibility, comprehension, English as second language

1. Introduction

Accented speech affects intelligibility and comprehension through a combination of linguistic and demographic factors that depend on both speaker and listener. The impact of accent of second language (L2) speakers as perceived by native (L1) speakers has been investigated in the context of English as Second Language (ESL) teaching and testing. There has also been some interest in exploring the role accent plays in the perceived intelligibility and comprehension of English by L2 speakers [1, 2, 7]. While these studies are inconclusive on the relation between accent, intelligibility and comprehension in ESL speakers, they do illustrate the role of proficiency, attitude and contextual linguistic information in addition to accent in perceived comprehension by L2 speakers.

Studies have also shown that a greater cognitive effort is required for the processing of synthetic speech vis-a-vis natural speech [3], and there has been recent effort to introduce [4] or modify [5] foreign accents in synthetic speech.

In India with 18 official languages and several hundred dialects, English remains the preferred language of business and commerce, science and technology as well as higher education. Each region may have its own normative Indian English accent which makes it difficult to define a single "Indian English" standard. We have investigated the interaction between accent and comprehension of ESL by visually impaired L2 speakers in southern India. The subjects of this study were students of the basic and advanced computer courses conducted by National Association of Blind, a not-for-profit NGO that works on economic rehabilitation of the visually impaired. Most of the training is imparted using the screen reader JAWS (version 9) with its default US English TTS voice.

The goals are to examine how the characteristics of fluent talker and L2 listener interact in an Indian context, for TTS and natural ("real") voices. Besides the traditional intelligibility test, we wished to quantify the effect of accent on task performance. For the use of a screen reader, we chose a comprehension test that included performance metrics for ease and accuracy.

In this study, we have treated our blind and low-vision subjects as we would any group in the population of 2nd language English speakers. None of the subjects had any detectable hearing impairment. Some considerations that have little to do with performance have been made and are mentioned as they arise.

2. Methodology

The intelligibility and comprehension tests each consisted of a series of recordings made with real or TTS voices of a specific accent, followed by multiple choice questions. Five voices were recorded in all.

2.1 Intelligibility test

Intelligibility was determined by the Diagnostic Rhyme Test (DRT) test, similar to the approach used in the Blizzard evaluations [6]. A series of sentences of the form

Now we will say _____ again.

were recorded with each voice to be compared, where the word in the blank was taken from one of the confusability pairs specified in the DRT tables (dense,tense) voicing, (mend,bend) nasality, etc. The subject was then asked to identify which of the pair they just heard. The two choices were provided in Braille, and the subject would respond with A or B.

Fifty sentences were played in all, with about 6 examples taken from each DRT confusability group. No two sequential sentences were from the same confusability type. Each of the 5 voices was heard for a group of 10 sentences, with the configuration changed for each subject so that all voices were heard speaking all sentences by some subject.

2.2 Comprehension test

In the fall, a comprehension test was added, which measured understanding by playing passages followed by multiple-choice questions. Passages were taken from standardized Indian textbooks for English-medium instruction, and were 20-60 seconds long. Three questions were asked/passage, and three choices were given for each answer. The questions were also read by the same voices (rather than using Braille), increasing the number of subjects that could take this test. Again, each voice was used for a group of three passages and questions, with the configuration rotating for each subject.

2.3 Recordings

The real voices were recorded with one of two headset microphones in a quiet office environment. The speakers were a native Kannada speaker (the main language spoken in the Bangalore region), a speaker from the Delhi region (referred to herein as the North Indian voice), and a native US

speaker (from the Midwest). All were fluent English male speakers whose education was in English, but with recognizable accents. To these voices were added two TTS voices, the RealSpeak US English (male, Tom) and Indian English (female, Sangeeta) voices available from the JAWS 9 release, where the TTS samples were approximately matched in rate and volume with the real voices. However, both the human and TTS voices retained stylistic differences in terms of characteristic prosody and articulation.

3. Test Protocol

The tests were administered at the National Association for the Blind in Bangalore, India, to two batches of computer training students in the summer and fall of 2008. The students ranged in age from 19-42, and were required to have at least 10^{th} standard (grade) education and basic English skills. No subject was a native speaker, though roughly 30% of the subjects received English-medium education through 10^{th} standard; the others learned in an Indian language such as Kannada, with English as a second language. Many subjects had or were progressing to university degrees, with most of the advanced education conducted in English. There were 13 subjects tested in the summer, and 21 subjects in the fall. Each subject was subjectively assessed by NAB (summer) or the researcher (fall) for command of English ("good" or "fair"), and asked a number of demographic questions such as age, level of vision, mother tongue, and education.

The tests were administered by a researcher in a quiet room with the subjects listening to the recordings on padded earphones and having control over volume. The intelligibility test required typically 10 minutes to complete, and was performed first (if the subject knew Braille). Each passage was played only once. For the comprehension test, subjects were encouraged to request repeats of any passage or question they liked; it required 45-60 minutes to complete.

4. Results

4.1 Intelligibility

The intelligibility test data was collected from two sets of subjects in the summer and fall of 2008. Despite some demographic variation between the subject pools, the performance results from the two groups were roughly consistent, as can be seen in Tab 1. The real Indian voices had generally better performance, while both TTS voices yielded better performance than the real US voice. In the combined data set, these variations were statistically significant.

Table 1*: Average performance (% accuracy) on intelligibility test for summer and fall subject groups*

Voice	Summer	Fall	Combined
Real NI	96±2	95±2	95±1
Real Kan	96±2	89±3	93±2
TTS Ind	90±3	87±3	89±2
TTS US	88±3	90±2	89±2
Real US	86±3	82±3	83±2

The sentences were grouped in sets of 10. Fig. 1 shows the variation of performance by question group for each voice. Some individual variation is revealed, with the first group evidently the most difficult, especially for the real US voice (6 subjs). In contrast, for groups 4 and 5 the real N Indian voice scored perfectly (5 and 6 subjects). The differences correlate with our expectations due to accent, both for speaker and subject: the TTS and real US voices scored 0.5 and 0.33 on voicing pair {dense,tense}, and TTS US and real NI scored 0.33 and 0.25 on sustenation pair {den,then}. The English "t" and "d" (alveolar) get mapped to the retroflex 't" and 'd' in most Indian languages, and the fricatives "θ" (as in "thin") and "ð" (as in "the") get mapped to the corresponding aspirated dental stops. Similarly, the real US voice scored poorly (0.5) on the compactness pair {yield,wield}, where the *y* is often assimilated in a Kannada English accent which does not allow a glide-high vowel hiatus. Of course, some failures seemed more general, like sustenation pairs *sh/ch* which appear in all 3 languages.

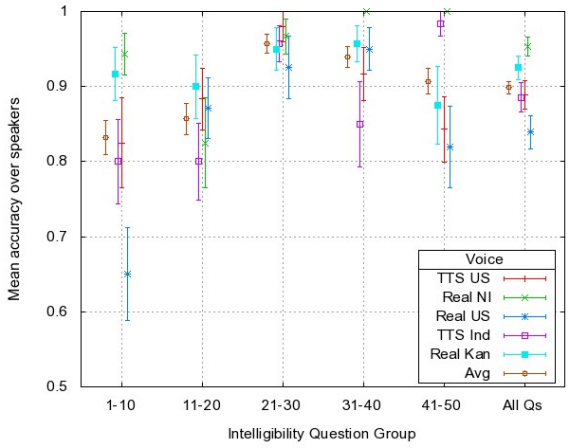

Figure 1: *Intelligibility test performance vs. question group and voice, combined summer and fall data, shortened y axis*

Of the demographic information collected, the strongest predictor of performance was the subject's command of English. The breakdown is shown in Fig 2 for the combined data, with a marked decrease in overall performance for the 8 "fair" English speakers with respect to the 19 "good" speakers. Besides the overall decrease in score, the advantage of a familiar accent seems to be enhanced when English is less than fluent—for "fair" English, the gap between best (real NI) and worst (real US) is 20% absolute, while it is 8% absolute for "good" English.

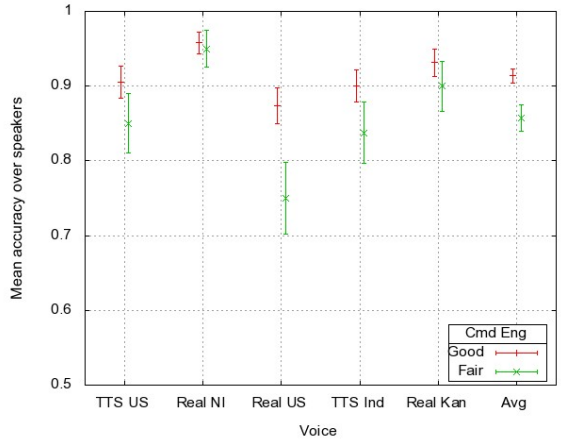

Figure 2: *Intelligibility test performance by voice for all data, vs command of English, with shortened y axis*

However, the specific accent performance advantage showed no real correlation to the mother tongue of the subject, contrary to our expectations. If we split the intelligibility data according to native Kannada speakers and others, we have 15 native Kannada speakers, 13 'other', of which 6 speak Telugu. The average performance by voice appears in Fig 3 and shows essentially no difference between the performance distributions due to mother tongue. This may be because all the subjects were exposed to Kannada-accented English: several of the 'other' native language speakers are still from Karnataka, and all are resident in Bangalore while attending the training classes.

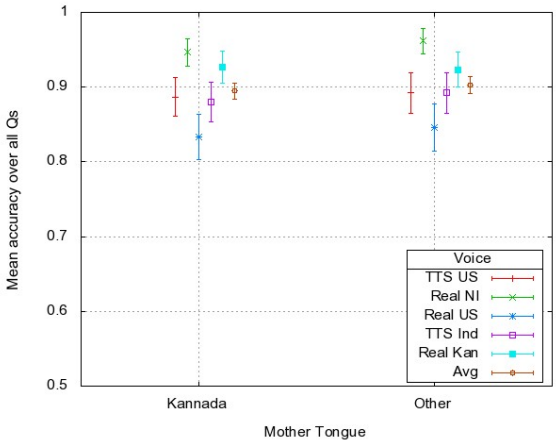

Figure 3: *Intelligibility test performance vs. mother tongue, full data set, and shortened y axis.*

4.2 Comprehension

The comprehension test was only administered to the second group of subjects. An earlier version was tried for the summer group, but the level of questions turned out to be too difficult. The passages were simplified, discarding the most difficult altogether and adding several readings from 3rd-6th grade English medium reading level. The questions were also made much more straightforward, and the number of choices was reduced from 4 to 3. The original approach was to have the questions posed in Braille; this was very slow and excluded several subjects that knew no Braille. For the fall subjects, the questions and multiple-choice answers were recorded by the same voices reading the passages, making this part of the task much faster and including all possible subjects. We excluded two speakers from consideration who found the test too difficult. Despite these simplifications, overall performance on the comprehension test averaged only 54% accuracy (random chance would yield 33%). Fig. 4 displays the performance for each voice by question group (now with 3 passages and 9 questions per group), combining all 21 fall subjects' results. The relative performance of each voice displays the same basic trends as in the intelligibility test: the real NI and real Kan voices score best, followed by the two TTS voices, with the real US voice scoring worst. The difference between best and worst is 13% absolute, or 21% relative, with p-value 0.012.

The performance effects of the subject's command of English appear in Fig 5, where as before, the "good" speakers of English significantly outperform the "fair" speakers.

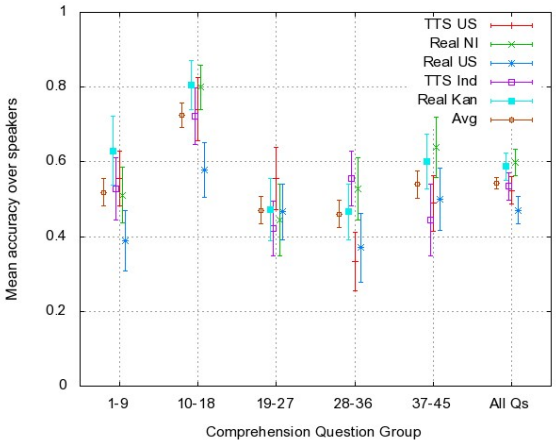

Figure 4: *Comprehension test performance by question group, average by voice*

However, the spread in scores vs accent has changed from the intelligibility test—here the 14 "good" English speakers seem to have found familiar accents significantly easier to comprehend, leading to a spread of 21% absolute between best and worst voice, while the 6 "fair" English speakers found all voices hard to follow (no significant difference). The low number of shared subjects between the intelligibility and comprehension tests prevent a direct comparison on the same speakers. We will return to this issue below.

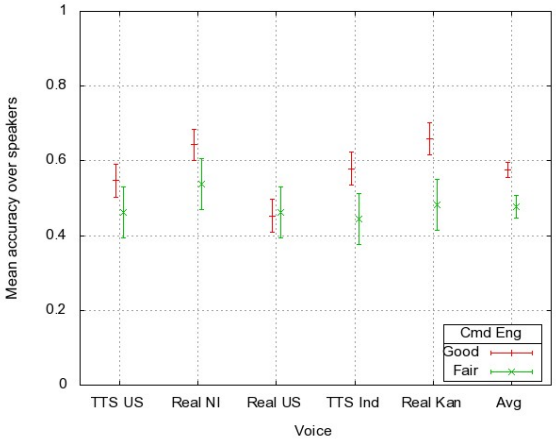

Figure 5 *Comprehension test performance by voice and subject's command of English*

Another way to assess difficulty in comprehension is the number of times a subject asked for a question or passage to be repeated. Clearly, the total number of repetitions requested varied considerably for the different subjects; however, certain passages and accents triggered more repetition requests than others, regardless of the subject's baseline. This is reflected in the average number of replays (summing over passage and question replays) for each group of questions, appearing in Fig 6. It reinforces the conclusion obtained from the performance scores of Fig 4: real Indian accents are easier to follow than TTS voices, but the real US voice is the hardest.

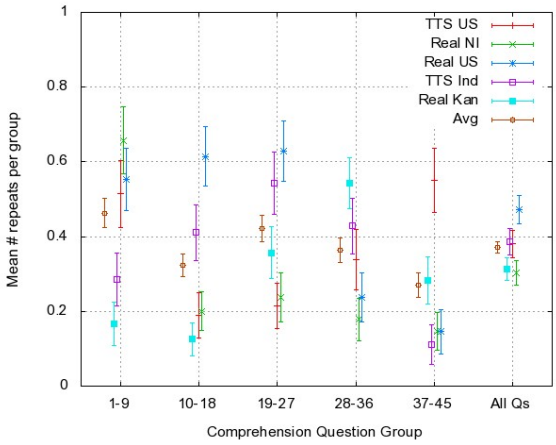

Figure 6 *Average number of repeats for each question group and voice, comprehension test*

4.3 Subjective preferences

After the tests were completed, the fall subjects were asked which voice they found easiest to understand. Not surprisingly, all but one chose either the real Kannada or real N Indian accented voice. Here the effect of mother tongue *did* enter: native Kannada speakers chose the Kannada-accented voice by a factor of roughly 2:1, whereas the other speakers were roughly split between the two real Indian accents. The remaining subject chose the Indian TTS voice, which is an achievement in itself for a TTS system.

The response did depend on how the question was asked: when asked what voice they "liked best", the first two subjects chose the real US voice, although for the first subject it was the lowest-scoring voice by a factor of four.

5. Conclusions

Returning to our stated objectives, we find for our population that accent has a complex but measurable impact on the intelligibility and comprehension of non-native Indian speakers of English.

In terms of intelligibility, we found that the voices with familiar Indian accents showed up to a 20% relative performance advantage over the least familiar accent, a native US speaker. However, the equivalent performance on intelligibility of the two TTS voices—one Indian-accented, one US-accented—shows that accent alone cannot explain all variation. This reservation is strengthened by the lack of correlation between performance by voice and the mother tongue of the subjects, though greater variation in subject demographics would be required to be definitive.

This ambiguity is consistent with results from earlier ESL studies indicating that accent match is not always helpful for intelligibility. The larger advantage of accent match for intelligibility for "fair" English speakers is similar to results in [7], for example.

The comprehension test showed a roughly 20% relative advantage for voices with familiar accents versus the unfamiliar real US voice, despite lower absolute performance. The data on requested repetitions reinforces the conclusion that poor performance reflects perceived difficulty in following a particular voice. Given the strong effect of mother tongue on subjective voice preference, we can attribute at least part of the effect to accent familiarity.

The subject's command of English seemed to affect performance on the intelligibility and comprehension tests differently. On the intelligibility test, "fair" speakers of English had a wide variation in performance by voice, and the "good" speakers showed much less of an effect. On the comprehension task, the spreads were reversed. Since only 13 speakers are shared between the two tasks, it is hard to draw strong conclusions. We hypothesize that the change in performance spread is related to task complexity. The intelligibility test is easy for the subjects with high English fluency, so they can focus on the phonetic disambiguation task, but those with a limited command of English are sufficiently slowed by unfamiliar accents that they do measurably worse on them. The comprehension test is more challenging, thus the more fluent English speakers are now slowed by the accent differences, while the task is simply so hard for those with limited English that accent no longer matters. A more extensive study would be required to test this rationale definitively.

From a speech synthesis standpoint, state of the art TTS voices are now outperforming a real voice on all tests, and have won over at least one convert subjectively. Though the TTS performance on the quantitative tests did not show a dependence on accent, the subjective preferences indicate that accent familiarity is important for perceived difficulty of comprehension. As anticipated, no single objective or subjective measure captured the effects of this complex issue. Taken together, our results suggest the choice of accent should be a significant consideration for user interfaces that address multilingual populations.

6. Acknowledgements

The authors would like to thank the recording team, Ms. Megha, and Microsoft Research Labs India for their participation and support. We would especially like to thank NAB for their help. FW thanks MSR India for support as a Visiting Researcher while much of this work was done.

7. References

[1] Major, R., *et al.,* "The Effects of Nonnative Accents on Listening Comprehension: Implications for ESL Assessment", TESOL Quarterly, 36(2): 173-190, Summer 2002.

[2] Munro, M., *et al.,* "The Mutual Intelligibility of L2 Speech", SSLA 28, 111-131, 2006.

[3] Pisoni, D., *et al.,* "Comprehension of natural and synthetic speech, effects of predictability on the verification of sentences controlled for intelligibility", Computer Speech and Language (1987) 2, 303-320.

[4] Yanagisawa,K., Huckvale,M. "Accent morphing as a technique to improve the intelligibility of foreign-accented speech.", *International Congress of Phonetics Sciences*, 2007, Saarbrücken, Germany

[5] Tomokiyo, L., *et al.,* "Foreign Accents in Synthetic Speech: Development and Evaluation", proc. Interspeech 2005, Lisbon, Portugal.

[6] Black, A., and Tokuda, K., "The Blizzard Challenge – 2005: Evaluating corpus-based speech synthesis on common datasets", proc Interspeech 2005, Lisbon, Portugal.

[7] Hayes-Harb, R., *et al.,* "The interlanguage speech intelligibility benefit for native speakers of Mandarin: Production and perception of English word-final voicing contrasts", Journal of Phonetics 36 (2008) 664–679.