On Non-Native Speaker Prosody: Identifying 'Just-Noticeable-Differences' of Speaker-Ethnicity

Richard Todd

Speech and Hearing Research Group
Department of Computer Science, University of Sheffield, England, UK
e-mail: R.Todd@dcs.shef.ac.uk

Abstract

This study tackles the issue of speaker-nativeness and the minimal cues required for determining particular ethnic groups beyond chance level. The work highlights that the heavily marked nature of some L_n speakers' voices make them fairly easy to single out as being foreign, generally. The investigation shows this is not the case for all non-native speakers, as some characteristics may sound quite like those of the native variety. By taking a finely nuanced approach to speaker-nativeness/-ethnicity, the empirical study uncovers which prosodic aspects remain native-like or become most differentiated during emphatic utterances.

Index Terms: cues, ethnicity, attribution, identification

1. Introduction

1.1. Explorations of accentedness

For some years there has been an interest in, and growing awareness of, how international differences in same-language speech features impact on intelligibility, comprehension, or even identifiability (e.g., [1], [2], [3], [4], and [5]). Much of such works arguably flow from the question of what causes some L_n users to be perceived as 'better', if not more or less marked than others, in terms of foreign-accentedness ([6], [7], and [8]).

With regard to English speech, some of the aforementioned attention has been more narrowly focused on phonetic production and perception. At times, like in [9] and [10], such examinations remain confined to one national variety. In other cases, researchers have compared multiple ethnic or national forms to another, regionally-dominant or target variety, as in [11], [12], and [13].

1.2. Unresolved issues

Despite all the approaches and motives illustrated above, a number of issues falling within the research areas of identification and foreign-accented speech have remained largely uninvestigated. One such issue relates to determining the point at which *actual* differences in speaker-nativeness or -ethnicity are likely to be *perceived* as barely palpable, or otherwise indistinguishable. In other words, there may be occasions when the utterances of a non-native speaker — or several members of his/her group — either approach or fall within the arguable limits of native inter-speaker variation.

It is unclear whether speech features that are perceived in this way generally remain 'unmarked' (i.e., native-sounding) or, once again, become 'marked' (foreign-sounding) as other i.e., personal pressures arise. Disentangling this matter is further complicated by the knowledge that reliance on listeners' evaluations alone would only takes us so far. Indeed, while [13] informs us that fallibility in human identification may be influenced by inter-stimulus commonalities, others contend that there is an effective limit on an individual's ability to faithfully categorise or recall different datasets in any case [14]. Furthermore, notions of comprehensibility and accentedness (i.e., ease of understanding the meaning of spoken content, and its (non-)conformity to a given/expected pronunciation standard) can become somewhat conflated, for some listeners [2].

2. Research aims

The initial goals of this study are to identify:

1. Which prosodic aspects of non-native speech appear most native-like, when speakers produce utterances free of any overt time, environmental, or social pressures.
2. Whether any of those features become stabilised, to the extent that they remain within the limits of native-like variability, when a follow-on utterance occurs.

3. Method

3.1. Speech Characteristics and style

With respect aim (1), above, it was decided to concentrate this study on the examples of speech produced solely when talkers do not feel any conscious need to accommodate or adapt to the speed, style, or interlocutor(s) of an ongoing conversation. Likewise, this would be the case with respect any audience/listener(s) or unwanted noise (such as, an extraneous/competing signal). Regarding aim (2), above, it was considered most appropriate to continue with phrases of one kind, thus ensuring lexical items and syntax used did not evoke needless semantic ambiguities. In this sense, speakers had a shared understanding of the message they would put across, and were comfortable their respective utterances would be comprehended as intended.

Essentially, the foregoing circumstances meant speakers were best allowed to talk in a way they considered representative of their most relaxed manner. This was believed to be typical of that used among close friends or peers.

3.2. Stimulus collection

A suitable range of stimuli for these purposes were found within the Non-Native English Speech Corpora (NESC). This is a collection of recordings used for forensic reference purposes [14]. The NESC contains a variety free- and fixed-form utterances; shadowed and imitated speech styles; plus,

expletives, etc., particular to several varieties of English, using speakers who were marked at differing levels. It therefore, provided a ready-made alternative to creating the source material needed.

In light of the impending difficulty of the tasks, and complications that could arise from bias or pre-conditioning/familiarity (see [13], [15], and [17]), twenty adult female only voices comprised the stimuli (mean age = 33.75 years old; s.d. 11.01). Four of the females were from the target, native, speech group (English-born; of British Anglo-Saxon descent). The remaining, non-native, female speakers were variously of either South Asian or Caribbean ethnic descent. In those two latter groups, were speakers who were British-born (for these British Asian and British Caribbean speaker-types n = 4, in each case). Given an awareness of contemporary shortcomings regarding the traditional definition of nativeness, this was determined in a more fine-grained manner than usual, following that of [16]. In that regard, Figure 1 (below) details how the native|non-native issue is actually nuanced in any diverse modern society. Both British Asian and British Caribbean speaker-types, above, were therefore considered *ethnically* non-native. Notwithstanding geographical origin, ethnic descent, or other languages spoken, all speakers were fluent in English and confirmed long-term residents of the same city.

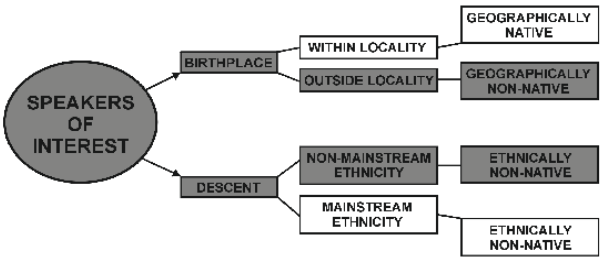

Figure 1: *A nuanced view of speaker-nativeness, applicable to any modern multi-ethnic society*

Utterances were randomised on a speaker-order basis, for this task. The words spoken by each of the speakers were:

"I'm too used to it. I just can't eat without it."

The phrase was chosen due to the fact the overall message allowed scope for personal emphasis (and thus, creating strong prosodic ties) on each speaker's part, without causing detriment to listener's obtaining overall meaning.

3.3. The listeners

The Listener Group comprised of 46 adult males and females who reported both daily use of, and proficiency in, the English language (overall mean age = 38.83 years old; male n = 26; female n = 20). Each passed a preliminary audiometric screening, and none of them reported the existence of any speech or hearing problems known to affect everyday conversation.

These listeners were taken from a pool of 120 participants in a previous ethnic group attribution (EGA) study of Todd [16] which investigated robustness in both good and impaired auditory settings. Thus, their potential for accurate identification was known to be frequent, at worst. Equally important, was the fact all listeners were from the same English city (Nottingham) as the speakers they would later hear. As such, there were no undue locale effects for these participants to overcome.

The Listener Group was divided into two sub-groups. Each performed the experiment in a separate session (n = 25 and 21, for sessions 1 and 2, respectively). In both cases listeners were situated in a quiet room, in what they agreed to be acceptable listening conditions.

3.4. The tasks

Identical instructions were given to all listeners before making their auditory judgments. This ensured that only voices which were considered native were attributed to one ethnic group — i.e., born in England, with parents of British Anglo-Saxon descent. Listeners were allowed to use their response forms to illustrate any features which assisted decision-making, if it helped them.

There is evidence to variously support the argument that repeated general exposure does not assure heightened EGA performance ([18], [19]) or one speaker's identifiability [20]. However, because the listeners had heard the speakers' voices previously (albeit, not for this study), the mode of presentation was changed. Hence, rather than providing complete speech signals as stimuli, they were bandpass-filtered (80-300 Hz). This meant listeners only had speech prosody available for decision-making. Thus, they were unable to make use of any phonetic and coarticulatory features that would otherwise have been audible. Since an orthographic transcript of the phrase was provided, listeners at least knew, lexically and semantically, what was actually being said.

3.5. Hypothesis

This task was thought to be even more challenging than that of [12], [13], [16] and [20]. This is, firstly, because phonetic detail is stripped from the speech; and secondly, in this study the speakers were anecdotally prone to having their voices misattributed. As such, it was expected that the majority of individual EGA scores for this study were unlikely to exceed 40% accuracy, even if most were above chance level (i.e., 1-in-5 attributions correct).

4. Results

4.1. Viability of speech prosody

From past studies we see that only allowing listeners to access prosodic information (alongside a relevant transcript), would still permit them to align, segment, and otherwise comprehend stimuli such as that presented in this investigation [21], [22]. This is since, in more limited contexts, prosodic information (stress and accent, in this case) begins to facilitate recognition. Additionally, Shilcock, Bard, and Spensley [23] have provided evidence to show that words having strong prosodic ties are recognised more readily than those without.

Clearly, the stimuli and auxiliary texts used for this study meet the above constraints. Additionally, in what is thought to be the first ever investigation into prosody, alone, being used to identify speaker-ethnicity, Todd [24], has illustrated both its overall efficacy and difficulty.

4.2. EGA task scores

Overall, the mean accuracy was about 38.6% for the task. Table 1, overleaf, shows how individual listeners performed, EGA-wise, in addition to providing details of inter-ethnic socialisation, using measures $Tdiff_1$ and $Tdiff_2$, further to [19]. With an attribution accuracy of almost 39% it was found that listener performance did not outstrip but, more or less, equalled that showed from the prosodic study of [24], in which overall mean EGA accuracy = 40%.

Table 1: *Listener Group details for task sessions 1 & 2.*

Listener	Gender M|F	Age	Tdiff$_1$	Tdiff$_2$	EGA Score
1	F	26	4	4	25
2	F	29	5	5	30
3	M	40	3	3	35
4	M	22	1	1	35
5	M	42	3	4	50
6	M	41	4	5	45
7	F	34	5	5	50
8	M	54	5	5	30
9	F	56	4	1	10
10	F	47	5	5	45
11	M	50	1	1	45
12	F	31	1	2	50
13	M	25	5	5	40
14	M	39	1	1	20
15	M	39	1	2	20
16	M	40	1	2	25
17	M	55	4	5	30
18	M	53	5	5	60
19	F	44	2	2	50
20	M	20	1	1	45
21	F	30	3	4	50
22	F	28	2	3	40
23	F	35	2	2	15
24	M	49	3	4	30
25	M	47	2	2	35
Session 1 Means:		**39.04**	**2.92**	**3.16**	**36.4**
26	F	27	4	5	40
27	F	41	5	5	55
28	M	69	4	5	55
29	M	18	4	5	40
30	M	31	2	2	45
31	F	39	1	4	40
32	F	19	4	4	30
33	F	54	4	5	25
34	M	24	5	3	30
35	F	42	1	1	60
36	M	65	1	5	45
37	M	29	1	1	40
38	F	23	5	5	45
39	F	55	3	4	35
40	F	55	1	1	20
41	M	46	2	3	20
42	M	30	4	5	50
43	M	29	2	4	55
44	M	41	4	4	55
45	F	20	5	5	60
46	M	53	4	5	20
Session 2 Means:		**38.57**	**3.14**	**3.86**	**41.19**
Overall Means:		**38.83**	**3.02**	**3.48**	**38.59**
Overall s.d.		12.93	1.54	1.56	13.07

Clearly, there was a larger number of participants in this research compared with [24]. Being almost 3-fold in size, these results greatly strengthen the findings of the other study. These results will furthermore, allow us to generalise such performance outcomes to attentive, yet non-expert, listeners.

We can see however, that a number of identifications were at chance level, and others were lower (n = 5 at 20% chance threshold; for below-chance n = 2, at 10 & 15%). It could perhaps be argued that this was largely due to the task's obvious difficulty. A more elaborated rationale for this outcome — which does not simply assume a pure lack of skill on part of the Listener Group — may be given in the terms:

- Individual listeners previously exhibited strong EGA potential in a task featuring degraded/filtered stimuli.
- Listeners only heard (as in a forensic 'voice line-up') a short, albeit somewhat emphatic, speech sample.
- The stimuli were certainly more nuanced, in terms of speaker-nativeness, than other identification studies.
- Non-native/ethnic minority group speakers may find an emphatic speech style easier (if not more important) to accurately reproduce, initially (see [25] and [26]).
- Non-native L_n speakers who acquire native-like prosodic characteristics do, indeed, come to be routinely misattributed as being British Anglo-Saxon.

The last of the above five arguments is perhaps the most complex to pick apart. In that regard, the auditory attribution task could only take us so far. Acoustic analyses were therefore performed to further develop an understanding of how traits of speaker-ethnicity could, at times, be 'just-noticeable' when not confused with other groups' forms.

Analyses were largely facilitated by listeners' range of marks and expressions to variously indicate characteristics which directed (rightly or wrongly) their decision-making. This information roughly translated into two feature-sets, being (1) Timing (measure = *Articulation Rate*); and (2) Pitch (measures = f_0 *Range*, *Minimum* f_0, *Maximum* f_0, and *Mean* f_0).

Taking the *Articulation Rate* measure first, some listeners perceived this to be a salient EGA cue, but only with respect speaker-types who were *geographically* non-native (further to §3.2 and Figure 1, above). A Dunnett's test confirmed a rate difference did exist between the control speakers (British Anglo-Saxon) and the others ($F(4,15) = 13.4$; $p < 0.001$). Tukey multiple comparisons further identified that it was both South Asian and Caribbean speaker-types who had significantly slower utterances (respective means = 4.42 and 4.25) than the British Anglo-Saxons (mean = 4.88), in terms of syllables per second. Still indistinguishable to all listeners however, was the finer-grained, yet significant, difference in *Articulation Rate* between the British Asian and British Caribbean voices which respectively sat closely below and above the British Anglo-Saxon range (where $t(6) = 11.3$; $p < 0.001$).

For f_0 *Range* listeners typically flagged South Asian and British Asian varieties as less expansive. Their intuition was borne out statistically ($F(4,15) = 46.04$; $p < 0.001$). However, the wide British Anglo-Saxon span may account for why the (also significantly different) British Caribbean shifts were misattributed, and so considered the same, by some listeners.

Minimum f_0 values were also a likely contribution to both kinds of South Asian speech seeming marked, pitch-wise. Their heights (maximally, 177.63 Hz) far exceeded the mean British Anglo-Saxon value (136.29 Hz, with s.d. 19.16); likewise, for all Caribbean-based forms, which were the lowest ($F(4,15) = 26.61$; $p < 0.001$). For the *Maximum* f_0 measure there were concerns regarding its extent for Caribbean-based utterances. Some listeners perceived tokens to be 'exaggerated' (but still correctly identified) or 'unnecessary' (and misattributed as South Asian). This was despite the data showing the only significance was that British Caribbean forms differed from British Asian and South Asian varieties ($F(4,15) = 7.26$; $p = 0.0018$), though it did become positioned higher than others.

An ANOVA followed by Tukey multiple comparisons of the *Mean* f_0 measure across the five speaker-types revealed a clear separation between the British-born, yet ethnically non-native speakers ($F(4,15) = 10.56$; $p < 0.001$). While British Asian speakers claimed the highest mean value (217.37 Hz) and their British Caribbean peers held the lowest (180.25 Hz; where a paired t-test showed $t(6) = 9.34$, with $p < 0.001$), neither was significantly differentiated from the rather intermediate position of the standard variety (British Anglo-Saxon mean = 198.84 Hz).

Considering that there were comments about 'high' or 'exaggerated' pitch, it is believed that overall utterance register — most particularly for ethnically non-native speakers — plays an important role during attributions of speech prosody. The fact *Mean* f_0 is essentially guided by overall f_0 minima and maxima suggests some speaker-types are disambiguated in this way by listeners. Even though this strategy may prove useful in some stages of perceptual categorisation, its effectiveness soon becomes limited when used to attribute speech that is not significantly differentiated from other (similar or related) ethnic varieties.

As a consequence to the foregoing dilemma, it would appear that the very most skilled (i.e., high-scoring) listeners employ a more sophisticated, adaptive, approach to performing EGAs. In this sense, they could seek to rely on a range of cues. From within the two feature-sets (Timing and Pitch) mentioned in the previous page it has been shown that speech prosody variously makes available at least five characteristics. In auditory tasks as difficult as the present, the most competent listeners may adaptively attach importance to each or any of these items as the auditory stimuli allows. In addition to this, other Timing-related features (e.g., pause length and stress) and Amplitude characteristics — which have not been considered in this study — may also play a part in decision-making.

5. Conclusions

This study confirms EGA can be performed when prosodic features form the sole basis of stimulus comparison and categorisation. Competence is lower than if judging speech normally, with all respective phonetic details audible. Though the mean score was just below 40%, only 4.3% of the Listener Group achieved below-chance accuracy. Unlike hypothesised however, 13 listeners had a 40-45% score; a further 28% reached or surpassed the expected ceiling level of 50%.

The score data from this study's EGA task support the suggestion that prosodic cues are *"limited to simply pre-filtering speaker types, in gross ethnic terms"* [24; p. 665]. However, it is clear there were several perceptual insights gained by conducting the task on a larger, more nuanced, scale. They have, in turn, been beneficial in directing both acoustic and statistical analyses, thus, helping to pick the 'just-noticeable' differences apart from those which remained much less salient, or more native-like.

In all, the results confirm that, when listeners' intuitions can be elicited, they should be heeded whenever mappable to an acoustic feature. However, even for well-attuned listeners, making attributions of speaker-ethnicity using speech prosody alone is by no means trivial [24]. Furthermore, attribution accuracy of partial/incomplete utterances is uplifted when even brief phonetic detail is afforded [12].

6. Acknowledgments

Thanks to Fusion Corporation R&D in Nottingham, UK for assisting with speech corpora and listening facilities.

7. References

[1] Bansal, R. "The Intelligibility of Indian English", PhD Thesis, University of London, 1966.

[2] Jun, H. G. & Li, J. Factors in Raters' Perceptions of Comprehensibility and Accentedness, in J. Levis & K. LeVelle [Eds.], Proceedings of the 1st Pronunciation in Second Language Learning and Teaching Conference, 53-66, 2010.

[3] Rickford, A.E. "Cognition, Comprehension and Critical Evaluation in a Multicultural Classroom: a Study in Literary Analysis and Appreciation", PhD Thesis, Stanford University, 1996.

[4] Benrabah, M. "Word-stress: a Source of Unintelligibility in English", International Review of Applied Linguistics, 35, 157-165, 1997.

[5] Köster, O., Schiller, N.O. and Kunzel, H.J. "The Influence of Native Language Background on Speaker Recognition", Proceedings of the XIIIth International Congress of Phonetic Sciences, 306-309, 1995.

[6] Stern, H. H. "What Can We Learn From the Good Language Learner?", The Canadian Modern Language Review, 31, 304-318, 1975.

[7] Vann, R. J., & Abraham, R. G. "Strategies of Unsuccessful Language Learners", TESOL Quarterly, 24, 177-198, 1990.

[8] Bamgbose, A. "Standard Nigerian English: Issues of Identification" in B. Kachru [Ed.], The Other Tongue. English Across Cultures, 99-111. Oxford: Pergamon Press, 1982.

[9] Wells, J. Jamaican Pronunciation in London. London: Basil Blackwell, 1973.

[10] Ingels, S. "The Effects of Self-Monitoring Strategy Use on the Pronunciation of Learners of English", in J. Levis & K. LeVelle [Eds.], Proceedings of the 1st Pronunciation in Second Language Learning and Teaching Conference, 67-89, 2010.

[11] Todd, R. "Acoustic-Phonetic Qualities of Asian- and Caribbean-English Consonant Clusters", Proceedings of the Institute of Acoustics, 20, 6, 351-355, 1998.

[12] Todd, R. "Auditory Perception and Ethnic Group Attribution of Unknown Voices: Assessing the Robustness of Experienced Listeners' Ratings when Confronted with Non-Native but Proficient English Speech", Proceedings of The Institute of Acoustics, 20, 6, 343-350, 1998.

[13] Goldstein, A.G., Knight, P., Bailis, K. and Conover, J. "Recognition Memory for Accented and Unaccented Voices", Bulletin of the Psychonomic Society, 17, 217-220, 1981.

[14] The Non-Native English Speech Corpora (NESC) comprises data from the ongoing 'Foreign English Speech' project and other content for forensic speech analyses and applications. See http://www.fusioncorporationrd.co.uk/project_overview.html.

[15] Miller, G. A. "The Magical Number Seven, Plus or Minus Two: some Limits on Our Capacity for Processing Information", Psychological Review 63, 2, 81-97, 1956.

[16] Todd, R. "Identifications of Speaker-Ethnicity: Attribution Accuracy in Changeable Settings", Proceedings of the IVth ISCA Workshop on Experimental Linguistics, ExLing 2011, 135-138.

[17] Pennington, M. C., & Ellis, N. C. "Cantonese Speakers' Memory for English Sentences with Prosodic Cues" The Modern Language Journal, 84, 3, 372-389, 2000.

[18] Gass, S., & Varonis, E.M. "The Effect of Familiarity on Nonnative Speech", Language Learning, 34, 65-89, 1984.

[19] Todd, R. "Ethnic Group Attribution: is Our Accuracy Constrained by Time Spent with Others?", Proceedings of the XVIIth International Congress of Phonetic Sciences, 1999-2001, 2011.

[20] Wretling, P., Sullivan, K. and Schlichtling, F. "Does Repeated Exposure to a Target Voice Reduce the Impact of a Similar Voice", Proceedings of the XIVth International Congress of Phonetics Sciences 99, 1385-1388, 1999.

[21] Ainsworth, W. "Pitch Change as a Cue to Syllabification', Journal of Phonetics", 14, 257-264, 1986.

[22] Van Heuven, V. "Effects of Stress and Accent on the Human Recognition of Word Fragments Spoken in Context: Gating and Shadowing", Proceedings of the Institute of Acoustics: Speech '88, 3, 811-818, 1988.

[23] Shilcock, R., Bard, E. and Spensley, F. "Some Prosodic Effects on Human Word Recognition in Continuous Speech", Proceedings of the Institute of Acoustics: Speech '88, 3, 819-826, 1998.

[24] Todd, R. 2002. "Speaker-Ethnicity: Attributions Based on the Use of Prosodic Cues", Proceedings of the 1st International Conference on Speech Prosody, Speech Prosody 2002, 663-666.

[25] Shepherd, M. "Effects of Ethnicity and Gender on Teachers' Evaluation of Students' Spoken Responses", Urban Education, 20, 10, 1-18, 2011.

[26] McKiman, D.J. and Hamayan, E.V. "Speech Norms and Attitudes Toward Outgroup Members: a Test of a Model in a Bicultural Context", Journal of Language and Social Psychology, 3, 1, 21-38, 1984.

[27] Munro, M.J. & Derwing, T.M. "Foreign Accent, Comprehensibility, and Intelligibility in the Speech of Second Language Learners", Language Learning, 45, 1, 73-97, 1995.

[28] Riney, T. J., & Flege, J. E. "Changes Over Time in Global Foreign Accent and Liquid Identifiability and Accuracy", Studies in Second Language Acquisition, 20, 213–243, 1998.