Speaker-Ethnicity: Attributions Based on the Use of Prosodic Cues

Richard Todd

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

Abstract

The earlier work of [19] showed that listeners could discern speaker-ethnicity from both full and shortened stretches of casually-produced speech. The type of cues listeners used to accomplish this remained unclear, however. In this study the phonetic detail of speech was removed by filtering; both mono- and multilingual subjects were then required to determine the ethnicity of speakers using the prosodic cues that remained. Results show that attributions can still be made using such cues. Also seen was that monolingual listeners tended to be the better performers. Overall however, the task was less trivial than the majority of subjects had anticipated.

1. Introduction

It is well-understood that differences between the suprasegmental elements of one language variety and another may severely effect speaker intelligibility or listener comprehension [2], [15]. In (forensic) speaker identification tasks language topology has also been shown to influence listener accuracy [12], [14].

Essentially, these and other factors have an impact on any auditory impression of speaker-nativeness or -ethnicity that we may have. Along these lines [19] showed that listeners are indeed able to determine speaker-ethnicity if given the full phonetic detail of a speech signal. This also appeared to be the case for speech heard in reduced linguistic contexts, which implies that not only phonological or phonetic information invokes speaker identification, or more particularly, ethnic group attribution (EGA). In studies of lexical access for example, others have illustrated that when phonologically-ambiguous target word onsets are presented in more limited contexts, prosodic information — in this case, stress and accent — begins to facilitate recognition [1], [23]. Similarly, [17] observe that words having strong prosodic ties are recognised more readily than those without.

To clarify whether the speaker-ethnicity would be similarly perceptible by prosodic cues alone, a further empirical study was conducted.

2. Method

2.1. Stimuli

A number of spontaneous speech files were taken from an existing corpus of (non-native) varieties of English speech that was recorded using high quality field recording equipment. The stimuli essentially comprised 5 informally-spoken phrases; each being produced by 9 adult females (mean age = 35.77 years old; s.d. 14.27) thereby giving 45 attribution trails.

As bandpass-filtered versions (80-300 Hz) of this speech was used, listeners could not capitalise on any phonological (or underlying phonetic and coarticulatory) differences present in the original recordings. So that subjects could at least anticipate what was actually being said, a written version of each phrase was provided.

2.2. Subjects

16 adult subjects (14 males; 2 females) participated in this study. All multilingual subjects (7 males; 1 female) reported both daily use of, and proficiency in, the English language (see Table 1, below, for further subject information). The remaining 8 subjects were monolingual and of British Anglo Saxon descent. To ease administration, the subjects were divided into two groups; each of which performed the listening task in separate sessions. To counter any biasing affects, the presentation order of the stimuli was reversed in the second session. All subjects were administered a preliminary audiometric screening; none of them reported speech or hearing problems.

SUBJECT	NATIONALITY	GENDER	AGE
1	Bulgarian	Female	27
2	Greek	Male	22
3	Indian	Male	46
4	Spanish	Male	30
5	Macedonian	Male	31
6	British	Male	23
7	British	Male	36
8	British	Female	21
9	British	Male	53
10	British	Male	33
11	British	Male	21
12	British	Male	26
13	British	Male	30
14	Greek	Male	25
15	Greek	Male	29
16	British	Male	23
Mean Subject Age = 30 years old (s.d. = 9)			

Table 1. All monolingual subjects were of British nationality. Listeners 1-7 and 8-16 attended the first and second listening sessions, respectively.

2.3. Task

Listening sessions (each lasting about 1 hour) were held in a quiet research laboratory situated within the University of

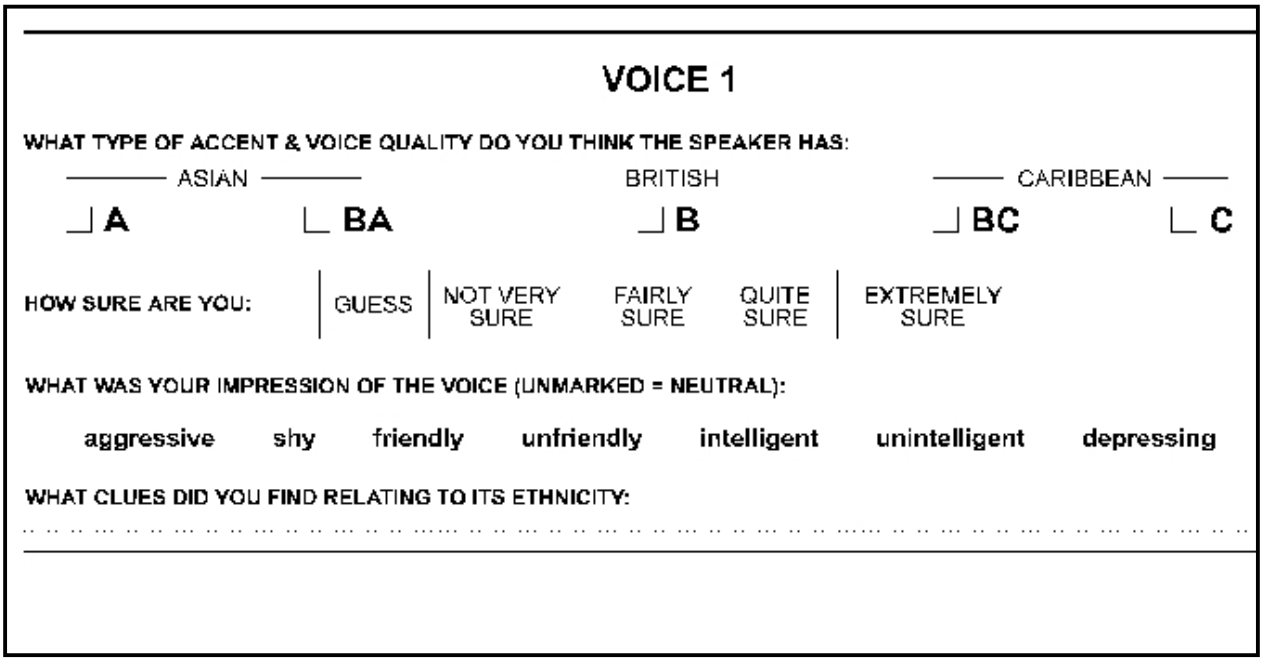

Figure 1. Response forms also recorded confidence ratings and other impressionistic information relating to individual ethnic group attributions of the subjects.

Sheffield's Department of Computer Science.

After hearing pre-recorded instructions and exemplary material, the listening task began. Subjects heard each stimulus being presented three times. They were then required to attribute the given voice to one of five ethnic categories:

1. A — Asian (born in Indian sub-continental region, parents of same descent);
2. BA — British-Asian (born in Britain, parents of Indian sub-continental descent);
3. B — British (born in Britain, parents of British Anglo-Saxon descent);
4. BC — British-Caribbean (born in Britain, parents of Caribbean descent);
5. C — Caribbean (born in Caribbean region, parents of same descent).

The presentation of each stimulus was followed by a silent response time (about 15 seconds) in which an attribution was recorded. The process was repeated until all 45 stimuli had been heard. Figure 1, above, details other items on the form and its formatting.

2.4. The hypotheses

The earlier study of [19] reported an overall mean of about 70% of EGAs being correct when utterances were heard in their full contextual form (maximum individual score = approx. 83%; minimum individual score = approx. 49%; s.d. = 8.5). The present task was thought to more difficult than the latter however, since the phonetic content available to listeners was minimal. The majority of individual scores was thus expected to be lower than 50% correct. It was further anticipated that the attribution accuracy of multi- and monolingual subjects would be notably different, if not significantly so.

3. Results

Individual correct identification scores were not particularly homogeneous and ranged from about 15% (subject 14) to 64%. Overall however, the mean attribution accuracy of the listener group was 40%.

Monolingual subjects were generally shown to be the better performers. Their number of correct EGAs across the 45 stimuli was almost 15% higher than for multilinguals. An independent group t-test showed a significant difference in performance accuracy between mono- and multilingual listeners ($p <= 0.001$). Four of the monolingual listeners achieved scores of 50% or above; the scores of the remaining half ranged from about 28 to 37%. The highest multilingual score was 47%.

Note that listeners were also required to self-report their response confidence for each stimulus heard. This was done on a 5-point scale of certainty ('guess' = 1; 'extremely sure' = 5). Little relationship between confidence ratings and attribution accuracy was observed, however. For example, one low-scoring subject was, at the least, 'fairly sure' for 35% of his attributions yet just one proved to be correct; another low-scorer felt this confident just 4% of the time. In general, the higher-scoring subjects were less likely to report their EGAs as being uncertain but still used the full range of the confidence scale.

The subject response forms further revealed that no specific set of prosodic cues appeared to systematically facilitate the attribution of individual ethnic groups. There was a tendency however, for monolingual listeners to comment

more readily on speech perceived to be from an ethnic group other than their own. South Asian speech was at times noted as being comparatively monotone, in terms of intonation.

On completion of the listening task, subjects were asked to report on its difficulty. Ratings were given on a 5-point scale, where '5' indicated the highest level of difficulty (see below, Table 2). Five participants considered the task to be equally demanding prior to, and after its administration; two of which (subjects 8 and 13) were among the highest scorers.

SUBJECT	DIFFICULTY	MOST USEFUL CUE
1	4	Word duration
2	5	Intonation
3	4	Intonation
4	5	Utterance duration
5	5	Intonation
6	5	Intonation
7	5	Intonation
8	3	Intonation
9	5	Intonation
10	5	Word stress
11	4	Intonation
12	4	Word duration
13	4	Intonation
14	5	Utterance duration
15	5	Intonation
16	5	Intonation
1=very easy; 2=a little easy; 3=just right; 4=a little hard; 5=very hard		

Table 2. Intonation was the prosodic cue most often considered to facilitate EGA.

The remaining 11 listeners perceived the task to be somewhat more demanding than expected. In terms of prosodic features used, intonation was most frequently reported as being the most powerful cue EGA-wise.

4. Conclusions

This study investigated whether listeners are able to extrapolate cues relating to speaker-ethnicity from heavily filtered speech. Results suggest that EGA can be performed when the remaining prosodic features form the sole perceptual basis for decision-making. Listener competence, however was shown to be appreciably lower than in similar tasks using phonetically-detailed stimuli.

Although monolingual subjects tended to perform better than others, attribution errors still arose when aspects of speech (e.g., word stress) were perceived to mirror or overlap those of an ethnic counterpart.

In all, the results suggest that the use of intonation and related prosodic cues is thus, limited to simply *pre-filtering* speaker types, in gross ethnic terms.

5. References

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

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

[3] Braun, A., 1995. The Effect of Cigarette Smoking on Perceived Age. *Proceedings of the XIIIth International Congress of Phonetic Sciences*, 2, 294-297.

[4] Braun, A., 1996. Age Estimation by Different Listener Groups, *Forensic Linguistics*, 3, 1, 65-73.

[5] Braun, A. and Cerrato, L., 1999. Estimating Speaker Age Across Languages. *Proceedings of the XIVth International Congress of Phonetic Sciences 99*, 1369-1372.

[6] Cutler, A., 1987. Forbear is a homophone: Lexical Prosody does not Constrain Lexical Access. *Language and Speech*, 29, 201.

[7] Cutler, A. and Clifton C., 1984. The Use of Prosodic Information in Word Recognition. In *Attention and Performance X*. H. Bouma and D. Brownhuis (eds.). New Jersey: Erlbaum Hillsdale.

[8] Cutler, A. and Norris, D., 1986. The Role of Strong Syllables in Segmentation for Lexical Access. *Journal of Experimental Psychology: Human Perception and Performance*, 14, 113-121.

[9] Darwin, C.J. 1975. On the Dynamic use of Prosody in Speech Perception. In *Structure and Process in Speech Perception*, A. Cohen and S.A. Nootebloom (eds.). Berlin: Springer. 187-194.

[10] Foss, D.J. and Gernsbacher, M.A., 1983. Cracking the Dual Code: Toward a Model of Phoneme Identification. *Journal of Verbal Learning and Verbal Behavior*, 22, 609-632.

[11] Fox, R.A., Flege, J.E. and Munro, M.J., 1995. The Perception of English and Spanish Vowels by Native English and Spanish Listeners: a Multidimensional Scaling Analysis. *Journal of the Acoustical Society of America*, 97, 2540-2551.

[12] Goldstein, A.G., Knight, P., Bailis, K. and Conover, J., 1981. Recognition memory for accented and unaccented voices. *Bulletin of the Psychonomic Society*, 17, 217-220.

[13] Halliday, M.K., 1967. *Intonation and Grammar in British English.* The Hague: Mouton.

[14] Köster, O., Schiller, N.O. and Kunzel, H.J., 1995. The influence of native language background on speaker recognition. *Proceedings of the XIIIth International Congress of Phonetic Sciences*, 4, 306-309.

[15] Ladd, D., 1986. Intonational Phrasing: the Case for Recursive Prosodic Structure. *Phonology Yearbook*, 3, 311-340.

[16] Marslen-Wilson, W.D. and Welsh, A., 1978. Processing Interactions and Lexical Access during Word Recognition in Continuous Speech. *Cognitive Psychology*, 10, 29-63.

[17] Shilcock, R., Bard, E. and Spensley, F., 1988. Some Prosodic Effects on Human Word Recognition in Continuous Speech. *Proceedings of the Institute of Acoustics: Speech '88*, 3, 819-826.

[18] Stuart-Smith, J., 1999. Voice Quality in Glaswegian. *Proceedings of the XIVth International Congress of Phonetic Sciences 99*, 2553-2556.

[19] Todd, R., 1998. 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: Speech '98*, 6, 343-350.

[20] ———, 2002a. Understanding the Efficacy of Intonation (Part 1): a Facilitator of Ethnic Group Attribution? Unpublished manuscript.
[21] ———, 2002b. Understanding the Efficacy of Intonation (Part 2): a Facilitator of Ethnic Group Attribution? Unpublished manuscript.
[22] ———, 2002c. Discerning Foreign-Accented and Ethnic Speech: A Survey of Social Factors. Unpublished manuscript.
[23] Van Heuven, V., 1988. 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.
[24] Wretling, P., Sullivan, K. and Schlichtling, F., 1999. 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.
[25] Zwisterlood, P., 1989. The Locus of the Effects of Sentential-Semantic Context in Spoken Word Processing, *Cognition*, 32, 25-64.

6. Acknowledgements

Thanks are given to my employer Fusion Corporation Research & Development, Nottingham, UK for tolerating the absence caused while conducting this work.