Proceedings of the Institute of Acoustics

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

Richard Todd

Speech and Hearing Research Group
Department of Computer Science, University of Sheffield, England, S1 4DP, UK.
E-mail: R.Todd@dcs.shef.ac.uk

1. INTRODUCTION

The subject of auditory perception (and subsequent rating) of speaker-nativeness in a binary — that is, forced-choice *native* versus *non-native* — setting has been studied by many researchers in the past.[1] Much of this work has been conducted with American English speech being the target phonological framework in mind and, in most cases individual vowels or utterances comprising the stimuli have been derived from citation-form speech produced in either isolated word-lists or longer passages [10, 11]. Indeed, many automatic speech recognisers (ASR's) have been trained on a mixture of such data and word-recognition accuracy is, in turn, fairly high in such systems when tested in similar 'real-life' situations, such as news broadcasts. However, it is also accepted that the performance of current ASR's is highly sensitive to differences in the input speech signal, and so the on-going trend in both algorithm and corpus development is to minimise — or at least document — any potential adulteration of a signal's consistency (see [1-3] for examples relating to the extraneous noise, and [4-7] for speech produced in adverse physiological conditions).

Arslen and Hansen [8], working on the premise that speaker-accent differences could also be seen as noise, demonstrate that an ASR's error rate can be reduced when generalised foreign accent models are employed in a system. As far as noise characterisation is concerned then, current approaches — although not yet optimal — seem quite promising. Even so, little attention has been paid to systematically characterising speakers as discrete groups, following Garrod and Doherty [12] for example. More explicitly, there appears to be little research that investigates speaker-ethnicity as a fine-grained (and perhaps more accent-independent) recognition parameter. It is felt that this is an important area since on auditory-phonetic grounds, ethnically non-native, but British-born speakers may possess many mild, but nevertheless unusual qualities as far as native listeners are concerned.[2]

[1] The term 'non-native' as used in the past has, at best, a rather vague meaning. My definition of speaker-nativeness has deliberately been made rather more fine-grained, and therefore does not require readers to induce their own description, as has typically been the case in the past. The introduction of the expression 'ethnically non-native' later in this paper suggests that the parameter space of a given speaker's (foreign) accentedness or voice quality may well extend beyond the boundaries of his or her birthplace. In other words, we can not postulate that matched speaker-dialects and geographical proximity — on auditory grounds — are robust identifiers of truly indigenous-sounding utterances in all empirical settings, even for monolinguals.

[2] It should be noted that in Britain and the USA, many of the so-called native citizens are descendants of people born in the Caribbean, Indian Sub-Continent, or Latin America which, in turn have language systems other than Standard English in use for everyday verbal communication. Since the passing-on of features belonging to any such co-occurring language system may give rise to differences in analytical findings, ethnicity has been marked. The parents of ethnically non-native speakers on the other hand, can be marked as being geographically non-native speakers of English.

Word recognition — regardless of whether undertaken by engineered or cognitive human systems — may suffer as a result since such speakers are often unable to deliver truly native pronunciation in a highly repeatable manner. With similarly proficient, but geographically non-native speakers both prosodic and pronunciation problems must be contended with, and are seldom eradicated altogether. Hence, it would appear then, that a speaker's second language (L2) lexicon develops at a different pace to its related phonological topology (which is accordingly termed P2 here).[3] So, the main questions here are:

1. Do listeners really have an ability to discem any small differences between these varieties of speech and those that are truly native?
2. If the listeners' auditory systems can in fact perform such a task reliably on full-length utterances, do the same features surface on shorter passages?

2. THE EXPERIMENT

The work was intended to obtain data that contributes to our understanding of the parameter space within which speaker-ethnicity and geographical non-nativeness are both embedded. Should the answers to the above questions be positive to some degree, evidence would then be available to show that there are thus problems that come with employing an entirely native phonological schema in current ASR systems. These problems would be further exposed when some element of (involuntary) 'mispronunciation' or creolization takes place during a speech act.

3. METHODOLOGY

3.1 Selecting the listeners

In order to conduct the required analyses of responses typical of experienced listeners, a group of subjects was assembled. The Listener Group comprised 8 adult males and 2 adult females. The bulk of the group members were part of the University of Sheffield's speech community. Eight of the subjects were speakers of (British) English as their first language (English L1 = 6 males, 2 females). The remaining subjects spoke other European languages as L1 and English fluently as L2. All (geographically) non-native subjects reported both good comprehension and fluent speech of the language. Detailed data on the Listener Group members are shown in table 1 below.

Columns 5, 6, and 7 of the table illustrate the languages spoken by each group member; the number of years the listener has been able to assimilate English speech (YAE); and the typical maximum percentage of time spent per week with speakers from ethnic groups other than their own (TSO) respectively. Clearly shown are the individual TSO levels, which collectively have a group mean value of 61%. Overall listener performance was expected to have been reasonably high due to the familiarity with non-indigenous varieties of English speech by default, in addition to the fact that the group's predominant expertise was speech-related.

[3] It is argued that the *way* in which things are said by a given speaker contributes more to any perceived non-nativeness than lexical word choice alone.

LISTENER	SEX	AGE	NATIONALITY	LANGUAGES SPOKEN FLUENTLY	YAE	% TSO (max)
L01	F	31	BRITISH	ENGLISH + GERMAN	31	75
L02	F	20	BRITISH	ENGLISH	20	50
L03	M	37	BRITISH	ENGLISH	37	25
L04	M	52	BRITISH	ENGLISH	52	25
L05	M	25	BRITISH	ENGLISH	25	100
L06	M	30	BRITISH	ENGLISH	30	75
L07	M	21	BRITISH	ENGLISH	21	10
L08	M	31	BRITISH	ENGLISH	31	50
L09	M	30	SPANISH	SPANISH + ENGLISH	18	100
L10	M	30	MACEDONIAN	MACEDONIAN + ENGLISH	22	100
MEAN		30.7			28.7	61

Table 1. Some details of the Listener Group used in both perceptual experiments. Note that significantly more than half of their weekly time is spent with speakers from an ethnic group other than their own (% TSO).

As a preliminary measure, members of the Listener Group were administered an audiometric screening (binaural stimulation via four sweeping sinewaves ranging from 80-500Hz, and 500Hz-3.5kHz). None of the group members considered in this paper failed to perceive the tones presented in the screening, or reported the existence of any known speech or hearing problems.

3.2 The stimuli

A collection of utterances were taken from a corpus being collected specifically for the study of non-native varieties of English speech. The selected material was recorded using high quality field equipment (principally, a Sennheiser MD421-U5 microphone, and a Philips DCC170 18 bit digital audio recorder). For the purpose of this study the utterance subset of interest was determined by the presence of nine phonological features being produced by each of the recorded speakers.[4] The Listener Group heard to two types of stimuli:

1. Full Contextual Form (FCF) — where speakers were recorded whilst saying an entire phrase and were using spontaneously self-chosen suprasegmental features, such as intonation and rhythm;

2: Reduced Contextual Form (RCF) — short excerpts of the above phrases were digitally edited out in order to reduce the intonational and rhythmic cues made available to the Listener Group.

The selected utterances consisted of 7 short statements, and each was produced by 5 female speakers (utterance total = 35). For Experiment 1 the FCF utterances were randomised on a speaker-order basis, classified via their respective phrase-groups, and finally recorded onto a compact disk. The RCF utterances used for Experiment 2 were arranged in a similar way on the same compact disk.

[4] In this case the features are the diphthongal glide /aɪ/; a voiced bilabial plosive /b/; lateral and post-alveolar approximants /l, r/ respectively; the alveolar stops /t, d/; the affricates /dʒ, tʃ / and an inter-dental fricative, /ð/. The choice was motivated by the fact that many speakers of English as L2 encounter problems with producing all of these sounds accurately until late in acquisition, if ever at all.

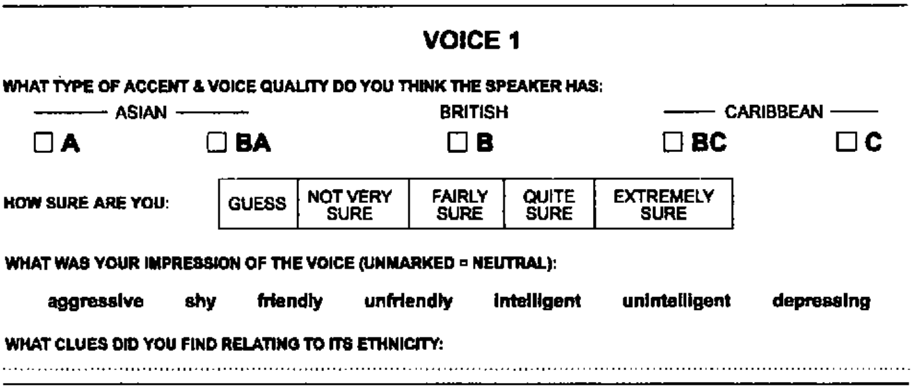
The experiments took place in a quiet area, at what the Listener Group considered to be both an acceptable listening level, and loudspeaker positioning. The group members' task in both experiments was to listen carefully to the FCF and RCF utterances (heard in experiments 1 and 2 respectively) with the specific aim of attributing each discrete stimulus to just one of five available ethnic groups.

The choice of ethnic groups that any one voice could have been attributed to was coded as follows:

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

A representative token of an unknown speaker's voice was presented to the Listener Group three times only, and sufficient time was allowed for responses to be recorded prior to moving on to the next phrase (typically 10-20 seconds after the last repetition of the token in question).

In order to reduce the cognitive load placed upon the listeners, the response forms featured a number of potential keywords that related to some extralinguistic (e.g., attitudinal) qualities of the speaker's voice (see figure 1 below). The group were encouraged to use any applicable features as markers to facilitate group attribution.[5]



VOICE 1

WHAT TYPE OF ACCENT & VOICE QUALITY DO YOU THINK THE SPEAKER HAS:

—— ASIAN —— BRITISH —— CARIBBEAN ——

□ A □ BA □ B □ BC □ C

HOW SURE ARE YOU:

GUESS	NOT VERY SURE	FAIRLY SURE	QUITE SURE	EXTREMELY SURE

WHAT WAS YOUR IMPRESSION OF THE VOICE (UNMARKED = NEUTRAL):

aggressive shy friendly unfriendly intelligent unintelligent depressing

WHAT CLUES DID YOU FIND RELATING TO ITS ETHNICITY:

Figure 1. The response forms used by the Listener Group during Experiment 1 provided intuitive means of documenting extralinguistic qualities of voices when heard in their full contextual form.

[5] The seven keywords used pertained to impressionistic — albeit, generally accepted — aspects of a speaker's attitude/character: aggressive, shy; friendly, unfriendly, intelligent, unintelligent; depressing. Unmarked voices were considered neutral.

Experiment 2 used the RCF utterances as stimuli, and the listeners' instructions were to attribute a given focus word to one the available ethnic groups and highlight any portion of the word that served as a cue to the grouping decision. As with Experiment 1, the response form contained additional space for written comments or transcriptions to be made. Each of the experiments took approximately 45 minutes to complete and the Listener Group were given a rest period of 15 minutes prior to the onset of Experiment 2.

4. RESULTS

The group attribution accuracy of the subjects taken as a whole was 69.71% in the first experiment. With the exception of one listener's performance (only 48.6% correct identification during the entire task), individual correct identification scores were fairly homogeneous. The lowest correct identification score was 62.8%, whilst the highest score was 82.8% (where the whole group s.d. for entire task = 8.52%).

As can be seen in table 2, listeners appeared to experience difficulty in grouping speakers in phrases 1 and 6. In these two cases, listeners were frequently confusing ethnically non-native speech with the native variety. As hypothesised, the native-speaking monolingual Listener Group members' confidence ratings tended to be a little higher than those of their bilingual colleagues. For the second experiment, all subjects perceived an increased level of task difficulty and thus, were not quite as certain of their decisions. This being the case, it was expected that the grouping error rate would increase accordingly. However, it materialised that the reduction of available questioned data led to listeners activating a rather more elaborate percept in order to discern ethnic group belongingness. In almost all cases overall performance had increased (see table 3 for group mean scoring data).

EXPERIMENT 1: GROUP ATTRIBUTION FCF-TYPE UTTERANCES						
PHRASE	VOICE 1 ETHNICITY	VOICE 2 ETHNICITY	VOICE 3 ETHNICITY	VOICE 4 ETHNICITY	VOICE 5 ETHNICITY	% CORRECT (GROUP MEAN)
1	A	BC	B	C	BA	60
2	C	B	BA	BC	A	68
3	B	C	BC	A	BA	74
4	BC	A	BA	C	B	74
5	A	B	C	BC	BA	84
6	BC	BA	B	A	C	60
7	B	BC	A	BA	C	68
		GROUP MEAN FOR ENTIRE EXPERIMENT				69.71
		GROUP S.D. FOR ENTIRE EXPERIMENT				8.52
A=ASIAN BA=BRITISH-ASIAN B=BRITISH ANGLO-SAXON BC=BRITISH-CARIBBEAN C=CARIBBEAN						

Table 2. Details of speaker-ethnicity for each auditioned voice are shown on a per-phrase basis. The listeners' correct group attribution across the 35 questioned utterances was close to 70%. Errors were often due to listener confusion in discerning instances of British, British-Asian, or British-Caribbean speech when presented without any intermediate Asian and/or Caribbean tokens.

EXPERIMENT 2: GROUP ATTRIBUTION OF RCF-TYPE UTTERANCES

PHRASE	VOICE 1 ETHNICITY	VOICE 2 ETHNICITY	VOICE 3 ETHNICITY	VOICE 4 ETHNICITY	VOICE 5 ETHNICITY	% CORRECT (GROUP MEAN)
1	C	B	BA	BC	A	62
2	B	C	BC	BA	A	72
3	A	BC	BA	C	B	86
4	BA	BC	B	A	C	74
5	B	C	BC	A	BA	76
6	B	BC	A	BA	C	70
7	A	BC	B	C	BA	78
					GROUP MEAN FOR ENTIRE EXPERIMENT	74
					GROUP S.D. FOR ENTIRE EXPERIMENT	7.39

A=ASIAN BA=BRITISH-ASIAN B=BRITISH ANGLO-SAXON
BC=BRITISH-CARIBBEAN C=CARIBBEAN

Table 3. Utterances scrutinised in their Reduced Contextual Form (RCF) produced a higher perceived level of task difficulty for the Listener Group, but nevertheless enabled some low-level features to be utilised during the grouping process. In all but one trial (Phrase 5) gains in grouping performance were realised.

5. DISCUSSION AND CONCLUSIONS

Trained listeners who have been exposed to prosodically unconstrained utterances appear to have difficulty in highlighting the extent to which some auditory qualities of ethnically non-native English speech overlaps its geographical counterpart. Additionally, no comparison with a fully innate variety of English — i.e., one derived from a British Anglo-Saxon speaker — revealed systematically discernible phonetic markers for accent until the contextual form of the questioned speech streams had been reduced.

It would seem then, that listeners undertake the group attribution by process of top-down elimination: sentential-level intonation first, followed by phonological topology, and finally, voice quality. This suggests that intonation provides adequate grouping information when long utterances are being processed. Lower-order phonetic manifestations of ethnocentric cues on the other hand, are apparently only fully utilised the during cognitive processing of sparse data (see the companion paper [15] in this volume).

The experiments reported here reveal group attribution tasks can be undertaken with reasonable competence, since prosodic information alone may provide an adequate cue to a coarse-grained level of speaker-ethnicity. However, a caveat is that the extended utterance length required may inhibit explicit surfacing of any detailed phonetic characterisation [13,14].

ACKNOWLEDGEMENTS

Thanks are given to the reviewers whose comments have helped improve this paper's content, and additionally to the Forensic Audio Unit at Fusion Corporation Research & Development for the unlimited use of their field recording equipment.

REFERENCES

[1] Le Bouquin, R. (1996), 'Enhancement of Noisy Speech Signals: Application to Mobile Radio Communications', Speech Communication, 18, 3-19.

[2] Zhao, Y. (1996), 'Self-learning Speaker and Channel Adaptation Based on Spectral Variation Source Decomposition', Speech Communication, 18, 65-77.

[3] Gierlich, H.W. (1996), 'The Auditory Perceived Quality of Hands-free Telephones: Auditory Judgements, Instrumental Measurements and their Relationships', Speech Communication, 20, 241-254.

[4] Hollien, H. and Hollien, P.A. (1991), 'Speech Intelligibility in Deep Diving', Proceedings of the XIIth international Congress of Phonetic Sciences, 5, 90-93.

[5] Bard, E.G., Sotillo, C., Henderson, A.H., Thompson, H.S. and Taylor, M.M. (1996), 'The DICEM Map Task Corpus: Spontaneous Dialogue under Sleep Deprivation and Drug Treatment', Speech Communication, 20, 71-84.

[6] Junqua, J.-C., (1996), 'The Influence of Acoustics on Speech Production: a Noise Induced Stress Phenomenon known as the Lombard Reflex', Speech Communication, 20, 13-22.

[7] Stanton, B.J., Jamieson, L.H. and Allen, G.D (1988), 'Acoustic-phonetic Analysis of Loud and Lombard Speech in Simulated Cockpit Conditions', IEEE International Conference on Acoustics, Speech and Signal Processing, 331-334.

[8] Arslan, L.M. and Hansen, J.H.L. (1996), 'Foreign Accent Classification in American English', Speech Communication, 18, 353-367.

[9] Hansen, J.H.L. and Arslan, L.M. (1995), 'Foreign Accent Classification using Source Generator Based Prosodic Features', IEEE International Conference on Acoustics, Speech and Signal Processing, 836-839.

[10] Lively, S.E., Psioni, D.B., Yamada, R.A., Tohkura, Y. and Yamada, T. (1994), 'Training Japanese Listeners to Identify English /r/ and /l/. III. Long-term Retention of New Phonetic Categories', Journal of the Acoustical Society of America, 96, 2076-2086.

[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 Mutlidimensional Scaling Analysis', Journal of the Acoustical Society of America, 97, 2540-2551.

[12] Garrod, S. and Doherty, G. (1994), 'Conversation, Co-ordination and Convention: an Empirical Investigation of how Groups Establish Linguistic Conventions', Cognition, 53, 181-215.

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

[14] Lublinskaja, V.V. and Sappok, C. (1991), 'The Role of F0 and Spectrum in the Perception of Voice Belongingess', in: C. Sappok and L.V. Bondarko (eds.), 'Bjuleten' Fonetcheskogo Fonda Russkogo Jazyka', Bochum: St. Petersburg, 80-99.

[15] Todd, R. (1998), 'Acoustic-phonetic Qualities of Asian- and Caribbean-English Consonant Clusters', Proceedings of The Institute of Acoustics 1998 Autumn Conference: Speech and Hearing 98.

[16] Elman, J.L., Deihl, R.L., Buchwald, S.E. (1977) 'Perceptual Switching in Bilinguals', Journal of the Acoustical Society of America, 62, 971-974.

[17] Flege, J.E. (1988), 'Factors Affecting Degree of Perceived Foreign Accent in English Sentences', Journal of the Acoustical Society of America, 84, 70-79.