THE PERCEPTION OF NATIVENESS: VARIABLE SPEAKERS AND FLEXIBLE LISTENERS

Duncan Markham

Linguistics, School of Literary and Communication Studies
Deakin University, Australia

ABSTRACT

Tests of foreign accent usually treat native listeners as reliable providers of accentedness ratings, and pay too little heed to task-specific effects on non-native speakers' performance. This paper details a number of factors which in fact influence native listeners' perceptions, and the native-like behaviour of non-native speakers' productions, based on the results of a large study of phonetic performance in second language learners. Listeners were observed to vary, at times considerably, in their perception of accent, depending on context, and type of stimulus, and at times showed distinctly idiosyncratic scoring patterns. Listeners' reactions to speaker voice pathology, mixed dialect pronunciation, and artefacts of read speech are discussed, and the effects of using different types of scoring system are examined.

1. INTRODUCTION

Studies of second or foreign language learners' degree of learning achievement frequently test the degree of non-native accent present in the learners' speech. The two approaches to accent are either to measure acoustic differences between specific characteristics of native pronunciation and the pronunciation of the learners (eg, [1, 2]), or to measure the degree of non-native accent perceived by native speakers of the target languages (eg, [3-5]). The latter approach relies on groups of listeners to provide consistent, unbiased, and homogeneous accentedness ratings in an experimental setting. This paper reports on a number of factors which were observed to influence listener judgements in the course of a large study of phonetic performance in second language learners ([6]).

1.1 Accent and perception

Perception of accent in social environments. Sociolinguistic research has demonstrated that listeners develop strong opinions regarding the personality of speakers on the basis of the speakers' accents, overt choice of language in a bilingual setting, and the content of their utterances (cf [7-9]). Second language acquisition research, meanwhile, has demonstrated that the ethnic identity of a speaker, when known to or identifiable by the listener, can influence a listener's judgement of strength of non-native accent ([10]).

Phonetic factors affecting accentedness. Although the language acquisition research community is aware of the existence of differences in the perceptual salience of specific types of learner errors or non-native characteristics, there has been little research conducted in this area. Studies have variously shown that not all types of phonological errors are perceived as being of equal gravity ([11]), and that prosodic errors may more strongly impede intelligibility than segmental errors do ([12]).

The perception of accent in an experimental environment. The presentation of speech in an experimental environment necessarily heightens a listener's awareness of characteristics which are either not noticed or are ignored or tolerated in normal communicative interaction. The interpretation of listener reactions must then depend on the researcher's focus in conducting the research. Where the focus is that of whether absolute native-like behaviour is possible, the experimental context may be appropriate, but if the emphasis is on accent perception in a speech *community*, listener sensitivity to accent may be greater than is normally the case.

Scoring of accentedness. Scoring of accent has taken many forms in previous research, ranging from very coarse measures (eg, 3-value scales) to exceedingly fine measures (eg, 256-value scales (potentiometer used as scoring device)). Usually these scales have labels such as "strong" or "weak" associated with a small number of points on these scales. Clearly these labels are open to a large degree of interpretation, while the use of the scales themselves must be examined post hoc to establish whether listeners have used the scales in similar ways.

2. THE STUDY

The phonetic performance of eight L1 Swedish speakers for a number of second language production tasks was investigated and reported on in [6]. The subjects had been chosen on the basis of their considerably better than average L2 performance in non-experimental circumstances (subjective judgement of the author). The subjects' performance was assessed by a group of three native speakers for each language tested.

The subjects participated in three tasks (further details of the elicitation technique can be found in [6]):

- the imitation of word and phrase stimuli from languages which the subjects could not speak or of which they only had very rudimentary knowledge (PhraseWord task),
- the imitation of phrases forming part of a coherent text for languages of which they

already had some knowledge (TextPhrase task), and

- the reading of seen, unseen, and practised texts (Reading task).

In an attempt to reduce the subjectivity of the interpretation of such values, or at the very least to establish clearer criteria for the choice of a value, a scale with well-described labels was devised for the study described below. A further concern was that many experienced language learners have been exposed to more than one dialectal model of their target L2 and that the effects of this needed to be provided for in any scoring device. The scale had seven values, whereof the highest three were to be applied to speech perceived as native of the target dialect (**L**), mixed target and other dialect (**M**), and native to another dialect (**D**). These were the *Native* values. A further four were defined as *Foreign* values: almost completely native (**1**), mild and occasionally perceptible accent (**2**), noticeably and constantly accented (**3**), and strongly accented requiring concentration from the listener (**4**).

The judges completed up to four test batteries for their language. The subjects' imitations from the PhraseWord and TextPhrase tasks were presented in random order for scoring and brief comment on accentedness (Tests A & B). Judges were required to score each utterance on a seven-value scale, as described above.

All subjects' imitations of each stimulus utterance were then presented in blocks (Test C) and judges were asked to indicate the *relative degree* of accent between imitations. Thus, for a hypothetical block, all imitations of the stimulus utterance X were presented in pairs, and the second member of the pair (spoken by, for example, speaker 2) was assessed for accentedness relative to the first member of the pair (spoken by, for example, speaker 1). The presentation followed the following pattern: $X_1X_2...X_2X_3...X_3X_4...X_4X_5....$ Judges could respond with '+' (more native-like), '=' (equally accented), '-' (less native-like). A small number of utterances in Test C were presented for scoring using the labelled scale, in order to check task-effects on judges' perception of accent. For the text readings (Test D), judges were asked to comment on any non-native aspects of the readings they heard, describe the accent (if any), and give each reading a score on the established scale.

Accent-score data were obtained for American English, Australian English, British English (Reading task only), French, German, and Finland-Swedish (the dialect spoken by an ethnic minority in Finland).

3. DISCUSSION

3.1 Scoring patterns

A number of statistical and non-statistical measures of scoring consistency and similarity were calculated, as it was important to understand judges' behaviour before drawing conclusions about the performance of subjects.

	Test A			Test B		
	jdg1	**jdg2**	**jdg3**	**jdg1**	**jdg2**	**jdg3**
L	22	57	26	32	67	39
M	13	21	24	16	21	39
D	16	5	12	13	7	7
1	14	7	13	31	30	28
2	10	—	13	26	3	15
3	8	—	2	10	2	2
4	3	—	—	—	—	—
n	*86*	*90*	*90*	*128*	*130*	*130*
prop. LMD	59%	92%	69%	48%	73%	65%

Table 1: Frequency distributions for scores used by each judge; American English.

Each judge's scoring patterns were examined for a) score frequency, and b) consistency by comparing scores for imitations which had appeared more than once in the tests (for control purposes). An example of scale-use is found in Table 1, which shows the frequency of use of each score category for each judge of American English. At the bottom of the table, the use of the *Native* range of scores (L-M-D) is expressed as a proportion of the total number of scores. Judge2 clearly differs from the other judges for Test A, scoring 92% of utterances as *Native*, as compared to 59% and 69% for the other judges. This observation was borne out by a Mann-Whitney U test of inter-judge difference (z=-6.31 (judge2:judge1) / -5.32 (judge2:judge3), p≤0.0001). Various idiosyncratic skews and differences in scoring patterns were also observable for judges of other languages.

Test A							
jdg1	**L**	**M**	**D**	**1**	**2**	**3**	**4**
L	9	4	6	5	4	-	1
M		-	-	1	3	2	-
D			6	-	-	-	-
jdg2	**L**	**M**	**D**	**1**	**2**	**3**	**4**
L	36	11	2	-	-	-	-
M		1	2	1	-	-	-
D			2	-	-	-	-
jdg3	**L**	**M**	**D**	**1**	**2**	**3**	**4**
L	12	11	1	4	3	-	-
M		7	4	6	5	-	-
D			2	-	-	-	-
	77%	67%	100%				

Table 2: Confusion matrices for Test A, American English judges, showing number of rescored utterances receiving a combination of at least one *Native* range score and any other score.

Tests of individual judges' scoring consistency revealed at times considerable discrepancies between scores assigned to the same utterance. Table 2 shows confusion matrices for judges and rescored utterances for Test A, American English. It can be seen that judge1, for instance, scored nine utterances as *L* both times they were presented, but a further four utterances were scored as *L* on one occasion and *M* on another, while six were scored as *L* and *D*, etc. One might conclude from this that the judges were highly inconsistent,

but it should be noted that a large number of discrepancies (usually more than 50%) fall *within* the *Native* range, so at least in that respect, some consistency is still observable. There is, of course, a caveat associated with a measure of consistency — judge2 appears to be more consistent in rescoring than other judges because she did not in fact use the entire scale (as seen in Table 1), thereby reducing both the potential for discrepancy and the resolution of the score data.

The existence of these differences was reinforced by significant values returned in Friedman and Mann-Whitney U tests of sample distribution.

Tests of score correlations between judges (Kendall's tau) revealed that general scoring patterns usually showed a significant correlation (two exceptions were observed out of fifteen pairs of judges tested (3 comparison pairs x 5 languages)). From the tests noted above one can conclude that, although overall weightings of individual stimuli *were* similar across judges, the use of the score categories *was not*.

Despite the defined use of each score-value, it is obvious from Tables 1 and 2 that the use of the scale reveals either massive differences in accent percepts for individual judges, or difficulty in using the scale. (It is of course the experimenter's hope that the latter possibility was avoided through careful design and provision of information to the judges.)

One is still left to ponder how the same stimulus can be heard as completely native on one listening and strongly accented on another *by the same judge*, or why one judge might hear a text reading as only slightly accented (1), when a fellow judge describes the same reading as strongly accented (4). Even the more detailed *comments* provided by judges after hearing the text readings show occasionally astounding degrees of disagreement: one subject's imitation of a Southern US accent was described as a "ridiculous imitation" by one judge, while another claimed that she would have accepted it as natural if she had heard it in a non-experimental context.

Further evidence which would lead one to believe that idiosyncratic differences in the listener's perception of accent are responsible for differences in scoring include the example of one judge who had to be rejected from the study because, having misread the instructions and assumed all speakers would be foreigners, he scored utterances more critically than any other judge, even scoring control stimuli by native speakers as accented.

Relative scoring of accent. Given that the assessment of accentedness using a labelled score system appears so plagued by variability, the alternative method of assessing accent — relative scoring (detailed above) — should be considered. The data from Test C, which elicited judgements of relative accentedness between stimulus pairs, were found to be a better guide to perceived accentedness than the absolute scores obtained in Tests A & B. Complete agreement across judges was obtained for 27-53% of stimulus comparisons (results from five languages), while 34-55% involved only minor differences between judges (eg, two judges giving a '+' judgement, and the third giving '=' instead). This leaves approximately 20% of cases where judges contradicted each other (some saying '+', some saying '-'). Of further interest is the fact that the differences between absolute scores obtained for specific stimuli in Tests A & B (eg, higher value vs lower value) corresponded well to the relative scores for the same stimuli when directly compared with each other in Test C.

3.2 Interaction between speaker characteristics and the listener

Listeners' perceptions of individual voices are not only important for sociolinguistics (as noted in the Introduction), but also for the understanding of the perception of accentedness. There were a number of instances of overt listener bias with regard to some speakers. Two of the male subjects had fundamental frequency and/or voice quality which led some judges to wonder whether the speakers were male or female. Assuming that the judges had normal expectations of typical male vs female speech patterns in their native language, these speaker-specific characteristics might have adversely affected the judges' scoring tendencies. A case in point is one of these male speakers. Overall, his American English utterances and text readings were rated most highly of any speaker's utterances, yet one of the three American English judges gave him particularly low scores. This judge described the speaker as sounding wrong, not articulating enough consonants, and mumbling, yet these characteristics were either not noticed by the other judges, or were not regarded by them as being a cue to non-nativeness.

Further examples of voice pathology affecting listener judgements were found for a female subject with a high degree of nasality, and a male subject with a voice variously described as 'dark' or strange.

Speaker models. Very few of the numerous studies of accent in second languages have taken a speakers' L2 model into account. Both natural and formal second language learners can arrive in an L2 environment with some prior experience of the L2. The model they may previously have been exposed to can be that of a different L2 dialect, or perhaps a non-standard or even non-native form of the L2. Alternatively, the speakers can settle in a part of the L2 community where an accented form of the L2 is the communicative norm. Naturally, this presents problems for the establishment of an experimental benchmark for labels such as 'native-like' or 'foreign'.

Listeners in the study described here occasionally had to contend with speech material from speakers with L2 models other than the target model. In particular, most of the speakers had British English as their main model for English, and in imitating American English frequently showed transfer from British English and/or interference from Swedish, their L1. Despite claims by the judges that they were good at identifying dialects of English (this information was elicited before their selection for the study), they often heard non-American features in the imitations as being indicators of an L2 English speaker, even though the features were at times

clearly British. Similarly, the judges for Finland-Swedish sometimes found it quite difficult to score imitations by the subjects who were all native speakers of Swedish dialects of Swedish. The unfamiliarity of other-dialect features mixed with Finland-Swedish pronunciation readily resulted in a *Foreign* impression.

3.3 Effects of methodology & materials

Imitation. An imitative technique was used in this study in an attempt to provide model utterances which speakers could use as immediate sources of information. At times this proved more difficult than assumed, with conflicts arising between existing models and the immediate models, resulting in more dialect transfer and L1 interference than anticipated (given the subjects observed performance outside of an experimental setting (author's subjective judgement)). Due to differences in sex between the stimulus voices and the subjects, two subjects sometimes had difficulty resolving the fundamental frequency characteristics of the stimulus voices (despite the fact that they were aware that imitation of absolute F0 was not part of the task).

Text reading. Read material is sometimes regarded as producing temporarily more native-like speech (due to increased monitoring), but can also be a more demanding task than spontaneous speech tasks, as avoidance strategies cannot be invoked, and reading tasks frequently require specific reading styles. It is this aspect which caused some task-based problems for some subjects. One part of the Reading task required subjects to prepare a text of their choice (permitting them to maximise their performance), but in so doing, the subjects frequently chose texts requiring marked reading styles of which they did not have an adequate command (eg, first person narrative, emotional descriptions, dialogues).

The unseen and seen texts (unprepared) caused problems where, for instance, multi-word concepts or compounds required complex knowledge of stress rules. This led to performance errors which would not have been observable in spontaneous speech. Naturally, this affects the perceived nativeness of a subject and highlights the need for the researcher to be aware of the skills that are being tested. Errors such as those mentioned above are not the typical, relatively intractable characteristics of non-native accent and are easily remediated. Certainly, knowledge of stress rules forms part of necessary phonological knowledge, but using elicited speech material of this kind as the object of subsequent accent scoring will result in the listener gaining a more negative impression of the subject's accent than is the case for the subject's normal or readily remediated performance.

Data presentation. Factors related to experiment design which can affect listener behaviour include such considerations as speed of stimulus presentation, task complexity, and the accentedness of adjacent stimuli. It was found that, when comparing the scores of rescored utterances, the absolute scores occasionally assigned in Test C were generally the same or more negative than those assigned in Tests A & B. One must assume that there was a task-effect when imitations of the same stimulus were explicitly compared in Test C, making listeners more critical.

4. CONCLUSION

This paper has illustrated a number of issues which need to be addressed in both experiment design and the assessment of test results when eliciting L2 speech or native listener judgements of accentedness. Listeners cannot be reliable sources of information regarding accent without first establishing an understanding of what affects their perceptions of accent both inside and outside a laboratory.

5. REFERENCES

1. Flege, J.E., M.J. Munro, and I.R.A. MacKay, *Effects of age of second-language learning on the production of English consonants*. Speech Communication, 1995. **16**(1): 1-26.
2. Caramazza, A., *et al.*, *The acquisition of a new phonological contrast: The case of stop consonants in French-English bilinguals*. JASA, 1973. **54**(2): 421-428.
3. Flege, J.E., *The detection of French accent by American listeners*. JASA, 1984. **76**(3): 692-707.
4. Bongaerts, T., B. Planken, and E. Schils, *Can late learners attain a native accent in a foreign language? A test of the critical period hypothesis*, in *The Age Factor in Second Language Acquisition*, D. Singleton and Z. Lengyel, Editors. 1995, Multilingual Matters: Clevedon.
5. Oyama, S., *A sensitive period for the acquisition of non-native phonological system*. Journal of Psycholinguistic Research, 1976. **5**(3): 261-283.
6. Markham, D.J., *Phonetic Imitation, Accent, and the Learner*. Travaux de l'Institut de Linguistique de Lund. Vol. 33. 1997, Lund: Lund University Press.
7. Giles, H., D.M. Taylor, and R. Bourhis, *Towards a theory of interpersonal accomodation through language: some Canadian data*. Language in Society, 1973. **2**: 177-92.
8. Giles, H. and P. Smith, *Accomodation theory: Optimal levels of convergence*, in *Language and Social Psychology*, H. Giles and R. St. Clair, Editors. 1979, Blackwell: Oxford. Pp 45-65.
9. Giles, H., *Accent mobility: a model and some data*. Anthropological Linguistics, 1973. **15**(2): 87-105.
10. Cunningham-Andersson, U. and O. Engstrand, *Perceived strength and identity of foreign accent in Swedish*. Phonetica, 1989. **46**: 138-154.
11. Cunningham-Andersson, U., *Native speaker reactions to non-native speech*, in *New Sounds 90, Proceedings of the 1990 Amsterdam Symposium on the Acquisition of Second-language Speech*, J. Leather and A. James, Editors. 1990, University of Amsterdam: Amsterdam. Pp 1-13.
12. Bannert, R., *Intelligibility and acceptability in foreign accented Swedish: the effects of rhythmical and tonal features*. PHONUM, 1995. **3**: 7-29.