

LLAS Subject Centre Mini-Project Report

Assessing and marking students' productions of IPA consonants and vowels.

Project Leader - Barry Heselwood, Dept of Linguistics and Phonetics, University of Leeds

Research Assistant – Fiona Skilling

1.0 Aim of the project

Students in the Dept of Linguistics & Phonetics at Leeds University take, as part of their diet of assessment in phonetics, a short individual Oral test in which their ability to produce six consonants and four vowels is assessed. Students will have had approximately 6 hours of classroom instruction and practice in the production and perception of IPA consonants and vowels, as well as having in-house recorded material to practise with in conjunction with their Module Workbook. The Oral test is worth 10% of their overall module mark, the other components being a coursework essay, a Dictation test and a written unseen examination. The purpose of the project is to see if the mark awarded for a production of a consonant or vowel might be influenced by the examiner's knowing what the target sound is, and whether any such influences might be greater or lesser depending on the kind of sound being produced, e.g. plosive, nasal, vowel, etc.

The hypothesis to be tested is that an examiner who knows which sound a student is attempting to produce will more often transcribe the production using a symbol which is closer to the symbol representing the target sound than will an examiner who does not know the target sound. If the hypothesis is confirmed, it would suggest that knowing the target is likely to result in an examiner awarding a higher mark to a student than s/he would give if the target were not known until after the examiner has made a transcription. This would have clear implications for the way phonetics Oral tests should be conducted and examined.

2.0 Oral test material

The material comprises thirty six laminated cards, each with symbols representing six consonants and four vowels, i.e. 10 sounds per card. There are altogether 36 cards, numbered 1-36, which between them have all the consonants and vowels on the practical phonetics syllabus (see list in Appendix A). The consonants are selected to be a balanced sample containing one non-pulmonic sound, and sounds of different manners and places of articulation. No card has both a trill and a voiced implosive as these are known to be particularly difficult for a significant number of students. Vowels are selected so that there are three primary cardinal vowels spread round the vowel quadrilateral and one secondary cardinal vowel. One of the primary vowels has a diacritic representing either creaky voice, breathy voice or nasalization.

2.1 Format of the Oral test

Students are tested individually in a session lasting five or six minutes. The cards are shuffled face-down at the start of each batch of six students and each student is given a card from the top of the pile which is afterwards placed at the bottom. Which card a student gets is therefore entirely due to chance. An IPA chart is provided which they can consult if they wish. Information about the format and the procedure for the test are published beforehand (see Appendix B). Each consonant must be produced in an open vowel frame – either [a__a] or [a__a] – with marks awarded for the fluency of

the transitions into and out of the consonant as well as for the consonantal articulation itself. Vowels have to be produced as sustained monophthongs, with marks awarded for the stability with which the correct quality is maintained.

3.0 Marking the students' productions

Two examiners are present, both of whom will have taught the practical phonetics classes. The first examiner has a copy of the student's card and therefore knows which sounds the student is required to produce. The second examiner does not know which card the student has. Each examiner has a score sheet (example in Appendix C) on which s/he notes the number of the student's card and transcribes the student's productions, supplementing the transcriptions with brief notes if necessary. When the student has left the room, the examiners compare their transcriptions, discuss the student's performance, and agree a mark for each sound. Marks are awarded in half-mark intervals – 0.5, 1.0, 1.5 etc. up to 4. The highest mark a student can get in the Oral test is therefore 40 (10x4).

3.1 Lack of anonymity in marking

Unlike with other forms of assessment, it is not possible to have anonymous marking for phonetics orals because the student has to produce the sounds in front of the examiner/s in real time. Any arrangements to make anonymous marking possible would either incur considerable extra expense by bringing in outside examiners who do not know the students, or, if marking were to be done on the basis of students' pre-recorded productions, would mean losing visual information which has always been valued in transcribing speech (e.g. Abercrombie, 1958: 232; Kelly & Local, 1989: 35; Heselwood & Howard, 2008: 385). This could in some cases jeopardise the student's mark. In our experience, students aiming to produce a retroflex consonant have been given some credit for observed tongue-tip retraction in productions that did not really sound convincingly retroflex. Similarly, observed lip-shape has gained students marks in vowel productions that were not auditorily spot-on, and being able to see vertical movements of the larynx in attempts at glottalic consonants has enabled some marks to be awarded where the auditory evidence was lacking. Being able to see the student's mouth therefore helps examiners to apply the long-held and general principle in academic examinations that the candidate always has the benefit of the doubt.

4.0 Data for this project

The data for this project are the transcriptions and notes on the score sheets of the first and second examiners relating to the productions of 131 students. In all, 1303 sounds were produced – 783 consonants and 520 vowels – yielding 2606 examiner transcriptions. In seven instances, a student made no attempt at a consonant.

5.0 Data analysis

The target sound for each production was ascertained by reference to the card number and then the transcriptions of the two examiners were compared to it and to each other. For each type of sound, the percentage agreement between the examiners was calculated. Agreement was either using exactly the same transcription, or using equivalent transcriptions (see section 5.1). Where there was disagreement, it was noted which examiner's transcription was closer to the target symbol. Deciding which transcription is closer to a target symbol is straightforward if one transcription matches the target either absolutely or closely and the other doesn't, but potentially

problematic if they differ from the target along different dimensions of classification. A 'distance metric' has been proposed by Cucchiarini (1996) to try to measure the degree of similarity between two sounds on the basis of their component features. A critique of this is offered below (section 5.2) with reasons why we did not find it very suitable for our purposes.

Where it was not clear which transcription was closer to the target, the transcriptions were assigned to a 'difficult cases' category (see section 5.2). Where it was clear, a one-tailed Chi-Square test was carried out to see if one examiner's transcriptions were closer by more than just chance levels. Chance dictates that the first examiner would not be closer significantly more often than the second examiner.

It needs to be appreciated that the closeness of a transcription to the target symbol is NOT an indication of how accurately the transcription represents what the student actually produced. The purpose of this project is not to assess the accuracy of examiners' judgements but to see if there is any evidence that knowledge of the target influences an examiner's transcription of a student's production.

5.1 Equivalent transcriptions

Two different transcriptions were deemed equivalent if there was no clear reason to interpret them as denoting different sounds. For example, [j^w] and [w^j] were taken as equivalent and therefore equally close to the target symbol [ɥ]; [ɛ̣] and [ɛ̤], however, were not deemed equivalent because we were interested to see if the examiner knowing the target would be more likely to interpret the production as a modification of the target vowel rather than of a different vowel. Arguably, they are not quite equivalent anyway.

5.2 'Difficult cases' and Cucchiarini's 'feature-based distance matrices'

In 183 (23.3%) of the 787 cases where the two examiners' transcriptions did not agree, it was problematic to decide if one was closer to the target than the other. For example, if the target is velar ejective [k'] and one examiner transcribes the student's attempt as velar plosive [k] and the other as uvular ejective [q'], which is closer? In one case the main symbol matches, in the other the diacritic matches. Not all examples involve diacritics. Can we decide whether alveolar [l] or palatal [ʎ] is a closer match to retroflex [ɭ]? The former matches more closely in terms of the active articulator, the latter in terms of the passive articulator.

Cucchiarini (1996) describes a method for assessing transcription symbol agreement which includes, as well as ways of trying to ensure a sensible and meaningful alignment of transcription texts, a way of measuring how similar one sound is to another. She rightly points out the inadequacy of all-or-nothing judgments of transcription agreement, emphasizing the need to recognize that similarities and differences are gradient. Her solution is to use the component articulatory features of the sounds, a solution which she justifies with the claim that 'transcribers analyse speech sounds in a proprioceptive way' (p.146). No doubt this is often true, but in assessing students' productions of sounds examiners may also make judgements about auditory quality, particularly of vowels, which they then reflect in their transcriptions. If judgements are partly auditory as well as proprioceptive/articulatory, then a distance metric based solely on articulatory features is not going to be entirely adequate.

A further problem is the decision of which articulatory features to use. One of the features Cucchiarini uses is [high]. Uvular and glottal sounds are specified as [-high] whereas velars are [+high]. If one examiner transcribes glottal [h] for a student's attempt at uvular [χ] and the other transcribes velar [x], the distance metric shows both transcriptions to be the same distance (1.0) from the target by the following calculations (see Cucchiarini, 1996: 146 and 154-5):

	place	voice	nasal	stop	glide	lateral	fric	trill	high	distributed
χ	4.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0
x	4.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0
difference	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
distance = 1.0										
χ	4.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0
h	5.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0
difference	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
distance = 1.0										

However, to a phonetician [x] is a closer attempt at the target than [h] and would be awarded a higher mark. A different set of features, e.g. one similar to the acoustically-motivated features of Jakobson, Fant & Halle (1952), could be constructed to come up with a different distance measure. If pharyngeal was included as a value of the place feature in Cucchiarini's matrices then the metric could be made to show that [h] is further from the target, but this underlines the somewhat arbitrary decision of how many values to set up for a feature. Whatever features are used, calculating distances by feature-matching and assigning numerical values is sometimes going to give the same numerical value for different differences, thus leaving us with the kinds of problems alluded to above – whether manner differences should be regarded as further from the target than place differences, or a place further forward should be regarded as the same distance as a place further back, etc. And numerical values tend to quantize features that are really continua, e.g. degrees of voicing, aspiration, lip-rounding, nasality, etc. thus failing to fully avoid an all-or-nothing approach.

Sometimes examiners give a sequence of symbols, e.g. [ʔnʔ] compared to [nj] where the target is [nj]. One transcription contains the target symbol and two extraneous sounds, the other does not contain the target symbol but two symbols which between them contain the target features – voicing, nasality and palatality. Which one should be judged closer, bearing in mind that fluency of transition is a marking criterion? It isn't clear how the feature-based distance matrices, which implicitly assume the reality of segments, would calculate this.

Fortunately in our data we have 604 transcription-pairs where it is clear that one transcription is closer than the other to the symbol representing the target sound. This represents 76.7% of the pairs where the two transcribers did not provide identical transcriptions, or transcriptions taken to be equivalent.

6.0 Results

The analysis revealed the number of agreements and disagreements given in Table 1.

Table 1. Summary of overall results.

Examiners give same transcription	Examiner 1 closer	Examiner 2 closer	'Difficult cases'	Total
516 (39.6%)	361 (27.7%)	243 (18.7%)	183 (14.0%)	1303 targets 2606 transcriptions

The two examiners provided the same transcription in 39.6% of all cases (516 out of 1303). Where the two examiners provided different transcriptions, a judgement was made about which was closer to the target symbol. In 183 cases this was problematic (see section 5.2) but in the other 604 cases, where it was unproblematic to decide which was closer, the first examiner was closer in 361/604 (59.8%) instances, and the second examiner was closer in the remaining 243/604 (40.2%) instances. A one-tailed Chi-Square test shows this to be significant at the 0.0005 level (with Yate's correction because only one degree of freedom). Within this overall result, there is however considerable variation across phonetic classes (see Table 2 and Fig.1). In some phonetic classes there was no trend for one examiner to be closer to the target more often than the other, while for other phonetic classes such a trend was found to be statistically highly significant. Classes where no significant difference was found were:

- STOPS - plosives, voiced implosives, ejectives, clicks
- NASALS – voiceless nasals
- FRICATIVES – lateral fricatives
- APPROXIMANTS – lateral approximants
- VOICE QUALITIES

Classes where a significant trend was found of one examiner being closer to the target symbol more often than the other were:

1st examiner closer more often at significance level of 0.025:

- VOWELS – those without added voice quality
- APPROXIMANTS – non-lateral approximants
- TRILLS

1st examiner closer more often at significance level of 0.005:

- NASALS – voiced nasals
- FRICATIVES – non-lateral fricatives

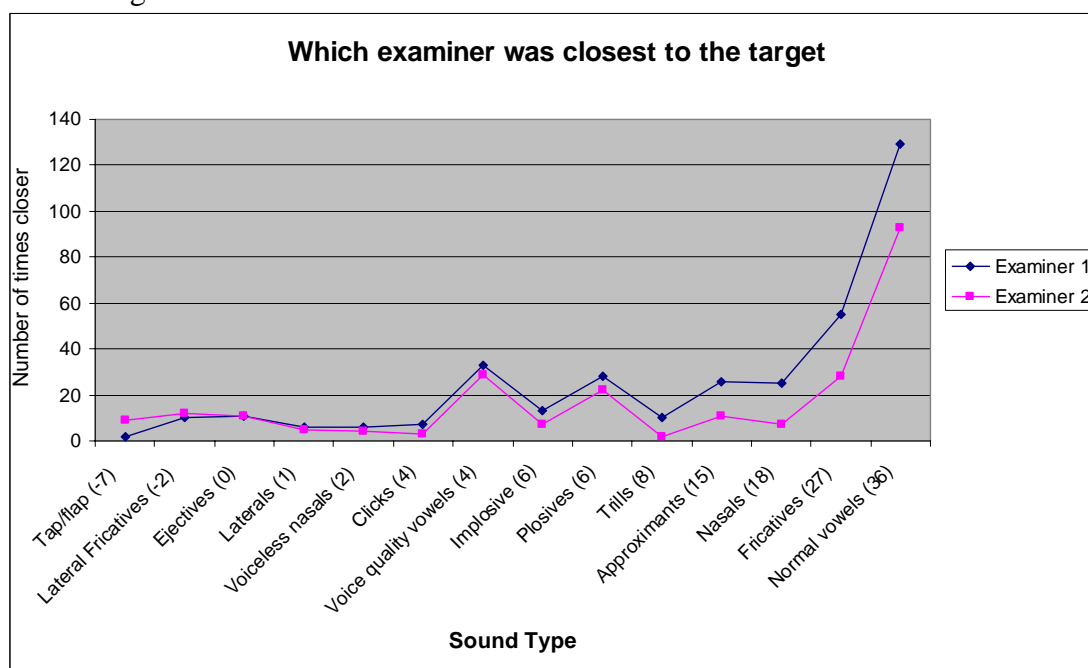
2nd examiner closer more often at significance level of 0.01:

- TAPS and FLAPS

Table 2. Clear cases of one examiner providing a closer transcription than the other analysed by sound type.

Sound type	Examiner 1 closer	Examiner 2 closer	Difference between 1 and 2
Tap/flap	2	9	-7
Lateral Fricatives	10	12	-2
Ejectives	11	11	0
Lateral Approximants	6	5	1
Voiceless nasals	6	4	2
Clicks	7	3	4
Voice qualities	33	29	4
Implosive	13	7	6
Plosives	28	22	6
Trills	10	2	8
Median Approximants	26	11	15
Nasals	25	7	18
Median Fricatives	55	28	27
Vowels	129	93	36

Figure 1. Summary of Table 2 in graph form. Difference between scores for each examiner given in brackets.



7.0 Discussion of results

The unexpected result of the second examiner being closer more often than the first in transcribing productions of taps and flaps may not be a robust finding. There were only 20 productions all told (not many cards have a tap or flap on them, there being only one of each on the syllabus). The examiners provided the same transcriptions for seven of these, two were ‘difficult cases’, so there were only eleven where one examiner was closer than the other. This may simply not be enough for a robust result to emerge. The same is true of clicks of voiceless nasals and clicks (10 of each where one

examiner was closer than the other), lateral approximants (11) and trills (12). Table 3 gives the numbers for all sound types.

Other sounds where it was found that one examiner was closer than the other with greater than chance frequency were vowels, median approximants, voiced nasals and median fricatives (highlighted in Table 3). In all these cases it was the first examiner's transcriptions that were more often closer to the target. Perhaps the most surprising result is that for all classes of oral stops other than taps, flaps and trills, there was no significant difference in which examiner was closer. At the moment we can offer no explanation for this difference between these oral stops and the classes highlighted in Table 3. Voice qualities also showed no propensity for one examiner to be closer more often than the other. This may be because there were only three voice qualities in addition to modal voice (creaky, breathy, nasalized) and it is usually clear from students' attempts which one they are trying to do. Both examiners know that one of the four vowels has a voice quality diacritic and that it isn't the secondary cardinal vowel. This means that the second examiner is perhaps no less aware of the target voice quality than the first examiner.

Table 3.

Sound type	Number where one examiner is closer
Plosives	50
Ejectives	22
Implosive	20
Clicks	10
Voiced Nasals	32
Voiceless Nasals	10
Median Fricatives	83
Lateral Fricatives	22
Trills	12
Tap/flap	11
Median Approximants	37
Lateral Approximants	11
Vowels	222
Voice qualities	62

8.0 Conclusion

The results of this study have confirmed the hypothesis that if an examiner knows the sound a student is trying to produce, his/her transcription of the student's actual production is more likely to be closer to the target than if s/he does not know what the student is trying to produce. This was found to be true for median fricatives, trills, voiced nasals, median approximants and vowels produced with modal voice and without nasalization. For other classes of sounds it was not found to be true. For taps and flaps, the opposite was found: the transcriptions of the examiner who did not know the target were more often closer than those of the one who did.

The main implication of the results for good practice in examining students' productions of consonants and vowels is that there should be at least one examiner present who does not know what the students are attempting to produce. This condition can of course be satisfied by having just a single examiner, but in this study the examiners did not completely agree on their transcriptions about 60% of the time which suggests that it is better if two examiners are present. The question then is

whether neither examiner should know the target. Our results indicate that neither of them should if we wish to avoid the biases revealed in transcriptions done with knowledge of the target. It is quite easy to arrange for both examiners to not know the target – a student's card number can be left undisclosed until after the student has completed the test.

9.0 Some further issues in marking phonetics tests

A number of other issues can be identified in the marking of phonetics orals which are beyond the scope of the present project but which we feel warrant attention. These are briefly sketched below.

Examiners' experience – Where there are two examiners, it will often be the case that one will be more experienced than the other, and/or more senior in the department, or perhaps just more confident or forceful about expressing a judgement. There may be a tendency for the less experienced, or more junior, or less confident, colleague to defer to the judgement of the other examiner. It is not clear how this can be addressed.

L1 of examiners and students – Although IPA sounds and Cardinal Vowels are supposed to be defined in terms of universal phonetic properties and therefore to be language-independent, it is widely recognized that in practice this is not entirely so. Ladefoged (1990: 342-3) raises serious doubts about whether sounds are ever the 'same' in different languages, and Laver (1994: 556-7) draws attention to the effects on perception of the phonological categories of one's own first language. If the two examiners have different first languages, even different accents of the same first language, this might bias their perceptions of a student's production in different directions. There may also be first language effects on production of IPA sounds possibly leading to the marking down of attempts by students whose first language is different from that of the examiners.

Ranking of errors – Scoring of student productions should not be all-or-nothing. A lot of thought needs to be given to how errors should be ranked in severity. Producing [ɔ] for a target [o] is less of an error than producing [ɔ̃]. These errors can be placed on the same parameter, i.e. vocalic openness, and are quite straightforward to make relative judgments about. But errors involving different parameters are much more problematic. Are manner-of-articulation errors better or worse than place-of-articulation errors, for example? And what about voicing errors? To have principled ways of answering these questions would help enormously in ensuring that errors are penalized consistently within and across candidates' performances. In section 5.2. above we draw attention to problems in using the feature-based matrices advocated by Cuchiarini (1996) to measure how close one sound is to another.

Dictation/perception tests – A distance metric would also be useful for awarding marks to students who put transcription symbols that are close to the target but not wholly correct. It would be appropriate for this to be based more on auditory distance than articulatory distance, perhaps taking into account studies that have used confusability matrices.

Recording the test – Arguments have been put forward in favour of making audio, or audio and video, recordings of phonetics orals and dictations. There are two main reasons given. Firstly, the external examiner can moderate the marks better if they can see what went on for themselves. Secondly, in the event of a student appealing against their mark, the recordings can be used to check if the mark was fair. These arguments have to be weighed against the argument that students may find the test much more stressful in the presence of video-cameras and microphones (and a technician may

have to be present to operate the equipment) which is likely to impair their performance. There is also the issue of trusting the professionalism of the examiners, not to mention the cost of recording and filming these sessions which take place over three or four days in each academic year.

References

- Abercrombie, D. (1958) The recording of dialect material. *Orbis* 111, 232-5.
- Cucchiari, C. (1996) Assessing transcription agreement: methodological aspects. *Clinical Linguistics & Phonetics* 10, 131-155.
- Heselwood, B. & Howard, S. (2008) Clinical phonetic transcription. In M.J.Ball, M.Perkins, N.Müller & S.Howard (eds.) *The Handbook of clinical linguistics*. Oxford: Blackwell.
- Jakobson, R., Fant, G. & Halle, M. (1952) *Preliminaries to speech analysis*. Cambridge, Mass.: MIT Press.
- Kelly, J. & Local, J. (1989) *Doing phonology*. Manchester: Manchester University Press.
- Ladefoged, P. (1990) Some reflections on the IPA. *Journal of Phonetics* 18, 335-346.
- Laver, J. (1994) *Principles of phonetics*. Cambridge: Cambridge University Press.

(3856 words, excluding Appendices).

APPENDIX A. Consonants and vowels on the practical phonetics syllabus.

Consonants:

p^h, p, b t^h, t, d t^h, t, d c^h, c, ʃ k^h, k, g q^h, q, ɟ, ʔ

m, m̥ m̄, n, n̥, n̄, ɱ ɲ ŋ N

ɸ, β f, v θ, ð s, z ʃ, ʒ ʂ, ʐ, ʑ, ʎ x, ɣ ɣ, ʁ ħ h

ʌ ɬ, ɮ

w, v, j, ɹ, ɻ, ɰ, ɱ, l, ɭ, ʟ

ɾ, ɽ, ʙ, ɹ, ʀ

p' t' c' k' q'

b d f g ɟ

⊙ || !

Vowels – Primary cardinal vowels 1-8, Secondary cardinal vowels 9, 10, 11, 16

Voice qualities – modal, creaky, breathy, nasalized

APPENDIX B. Procedure for the Orals test.

PROCEDURE FOR PHONETICS ORALS

You will be given a card at random with 6 IPA consonant symbols and 4 IPA vowel symbols on it. Some symbols may have a diacritic attached. They will all be sounds we have done in class and which are in the Sound Files accompanying the Workbook.

There will be an IPA Chart available for you to consult if you are unsure of a symbol.

You are to produce each consonant in an [a__a] or [ɑ__ɑ] frame as we have done in class, and each vowel as a sustained monophthong.

You can have several attempts but stop when you feel you have produced it as well as you can.

It is important that one of the two examiners does not know what the symbols on your card are, so please do not say anything which would let him/her know. Refer to the sounds as 'first consonant', 'third vowel' etc.

It is worth attempting a sound even if you think you can't do it – each sound is marked out of 4 so you might get some credit for it.

Relax beforehand (!) – your vocal tract will perform better if it is not tense. It will also perform better if you don't slouch in the chair or lean on the table.

Remember – we've been through it as well!

APPENDIX C. Examiners score-cards.

1st Examiner:

Student's name:

Card No.:

CONSONANTS	VOWELS
1	1
	2
2	
	3
3	
	4
4	
5	
6	

2nd Examiner:

Student's name:

Card No.:

CONSONANTS	VOWELS
1	1
	2
2	
	3
3	
	4
4	
5	
6	