The ISLE corpus of non-native spoken English[1]

Wolfgang Menzel[1], Eric Atwell[3], Patrizia Bonaventura[1], Daniel Herron[1], Peter Howarth[3], Rachel Morton[2], Clive Souter[3]

[1]Universität Hamburg, Fachbereich Informatik, Vogt-Kölln-Strasse 30, 22527 Hamburg, Germany
[2]Entropic Cambridge Research Labs, Compass House, 80-82 Newmarket Road, Cambridge, CB1 4LD, Great Britain
[3]University of Leeds, Woodhouse Lane, Leeds LS2 9JT, Great Britain

{menzel|herron|pbonaven}@informatik.uni-hamburg.de, rim@entropic.co.uk, {eric|cs}@scs.leeds.ac.uk, p.a.howarth@leeds.ac.uk

Abstract

For the purpose of developing pronunciation training tools for second language learning a corpus of non-native speech data has been collected, which consists of almost 18 hours of annotated speech signals spoken by Italian and German learners of English. The corpus is based on 250 utterances selected from typical second language learning exercises. It has been annotated at the word and the phone level, to highlight pronunciation errors such as phone realisation problems and misplaced word stress assignments. The data has been used to develop and evaluate several diagnostic components, which can be used to produce corrective feedback of unprecedented detail to a language learner.

Introduction

Project ISLE (Interactive Spoken Language Education) has the goal of integrating state-of-the-art Hidden Markov Model [HMM] speech recognition technologies into a computer-based package for intermediate level learners of English. The use of speech recognition [SR] will allow students to use spoken language as the most natural form of communication. More importantly it allows for on-line diagnosis and correction of both the communicative and grammatical adequacy of the spoken utterance and possible pronunciation errors made when speaking. The first goal is achieved in a customary way by prompting students with a small set of options they can select from. A low perplexity speech recognizer checks whether the student's selection was an appropriate one. Speaker adaptation techniques are used to compensate for accented, non-native speech.

Afterwards, custom designed components of the ISLE system are invoked to locate and describe phone- and stress-level pronunciation errors in the utterance (Herron et al. 1999). Thus, the system is in a position to produce detailed feedback to the student and to offer tailored practice for the errors encountered. The range of oral activities the student is engaged in includes reading exercises, producing minimal pairs, selecting a item from a list of options, and combining items from different selections. Although the technologies used and developed are theoretically valid for any language pairs, ISLE is focusing on Italian and German learners of English.

Purpose/Goals

To support the development of pronunciation training tools a corpus of non-native speech was required for three reasons:

1. to train the parameters and rules used in the recognition and diagnosis systems;
2. to test the performance of the system on a known data set; and
3. to evaluate the contribution of speaker adaptation for improving the reliability of the native British English recognizer.

While the last function requires only a word-level transcription the first two also demand that the corpus be annotated at the phone- and stress-level for pronunciation errors. Additionally, it was desirable to test the system on a variety of exercises of various complexities (or perplexities, more specifically), since the actual system was planned to have both simple and complex exercises.

The language material for testing the speaker adaptation was chosen from a non-fictional, autobiographical text describing the ascent of Mount Everest (Hunt, 1996). It was selected so that speakers/readers would not have to deal with reported speech or foreign words, which may cause them to alter their pronunciation. Approximately 1300 words of the text (82 sentences) were chosen, to be read by each speaker.

To test the recognition and error diagnosis capabilities a different kind of data was collected with the intention of capturing typical pronunciation errors made by non-native

[1] This research has been supported by the European Commission under the 4th framework of the Telematics Application Programme (Language Engineering Project LE4-8353).

speakers of English in controlled language learning situations. Therefore, the constraints on this kind of data come firstly from the exercise types which have been identified as being important by an initial user survey, and secondly from the tasks complexity for which a sufficiently high recognition accuracy can be expected.

The linguistic complexity for the speech recognizer is restricted by assuming that a mini-grammar can be written by the courseware author for the intended domain. Initially a perplexity of 6-10 was considered acceptable for exercises where alternative words and expressions were to be chosen from a given list. This part of the material consists of approximately 1100 words contained in 164 phrases.

Data

The linguistic material was divided into seven blocks (cf. Table 1). One third of the data used a large vocabulary and was not annotated (except that word-level errors were noted); this portion of the data was used solely for testing of adaptation of the SR system (Block A, B and C). The remaining phrases were focused on problem phones (as identified by language teachers), weak forms, words with potentially tricky stress patterns, and difficult consonant clusters. It was also divided by the type of exercise for which each phrase might be an answer: simple exercises, such as minimal pairs or multiple choice (Block D); fully-constrained reading exercises (Block E); or slightly-complex description exercises (Block F and G).

All these utterances were designed to cover problem phonemes and compounds. These may be L1 specific, as the following examples show:

German learners:

vowels — the difference between /eh/ and /ae/
the difference between /ao/ and /ow/

consonants — the substitution of /v/ by /w/
the substitution of /th/ and /dh/ by /s/ and /z/

Italian learners:

vowels — the pronunciation of /ih/ as /iy/
the pronunciation of /ae/ as /eh/

consonants — the omission of /hh/
the substitution of /th/ and /dh/ by /t/ and /d/

Speakers

Speech data was collected from 23 Italian and 23 German intermediate-level speakers of English. Volunteer speakers were sought from among the employees and students of four different project sites in Italy, Germany and the UK. The aim was to balance these for native language (German/Italian), whilst also collecting data from a small number of non-native speakers from other countries (Spanish, French, Chinese), and from native

Block	# Sents.	Linguistic Issue	Exercise Type	Examples
A B C	27 33 22	Wide vocabulary coverage (410)	Adaptation/ Reading	“In 1952 a Swiss expedition was sent and two of the men reached a point only three hundred metres from the top before they had to turn back.”
D	81	Problem phones Weak Forms	Minimal Pair Item selection/ combination	“I said bad not bed “She's wearing a brown wooly hat and a red scarf.”
E	63	Stress Weak Forms Problem Phones Consonant clusters	Reading	“The convict expressed anger at the sentence.” “The jury took two days to convict him.”
F	10	Weak Forms Problem Phones	Description/ Item selection/ combination	“I would like chicken with fried potatoes, broccoli, peas and a glass of water.”
G	11	Weak Forms Problem Phones	Item selection/ Combination	“This year I'd like to visit Rome for a few days.”

Table 1. Linguistic material

British English speakers. The latter two groups were included to allow comparison with other types of non-native learners, in order to monitor possible changes in system performance according to L1 model. We also initially intended to balance collection for sex, age and proficiency. Given the small number of speakers this has only been achieved to a limited degree. Table 2 gives an overview of the speaker sample. Proficiency ratings are based on a self-judgement of speakers.

L1	Sex		Proficiency				
	M	F	1	2	3	4	Total
German	13	10	-	-	8	15	23
Italian	19	4	27	11	4	1	23
Total	32	14	27	11	12	16	46

Table 2: Speaker sample

Recording Conditions

The phrases were recorded in non-noisy environments using high-quality headset microphones. In order to minimise the effect of growing familiarity with the recording tool, or of boredom with the exercise affecting the quality of recording for any subpart of the data, these were presented in a semi-randomised order. Difficult/long blocks (A, B, C and E) were interspersed with easier ones (D, F and G). The Everest text (Blocks A, B and C) was distributed so that no two Everest blocks appeared in sequence.

Speakers required between 20 minutes and one hour to record the entire set of phrases; they were able and encouraged to re-record those in which they realised they had made a large error (e.g., misreading one or more words). The data was recorded directly into WAV format, using a sampling rate of 16kHz at a resolution of 16 bits.

Before beginning recording, speakers were presented with an electronic form to collect demographic data: name, age, sex, country of origin, native language, and own English proficiency judgement. The date and location of recording were also collected. The sentences to be read one by one were presented to the speaker on the screen until they have finished. Alternatively, if a speaker required a break, the session could be suspended and the rest of the sentences recorded at a later date.

Annotation

The recorded data have been transcribed and annotated in a sequence of partly automated steps. After checking the quality of the recordings, three levels of reference transcription have been added to each waveform file. This was achieved using a British English recogniser (Young et al. 1999) performing forced alignment at the word level, based on the text in the sentence prompts or its cleaned-up version. The Hidden Markov Model of the recogniser provides a best-fit alignment from words in the prompt to the waveform.

Canonical pronunciations (phone annotations) then have been added and aligned by lexical look-up from the recogniser's pronunciation dictionary. Although International Phonetic Alphabet (IPA) labelling might have been desirable from a linguistic perspective, the chosen phone set was Entropic's UK English phone set (Power et al. 1996, see also the appendix). Note however that a mapping exists from IPA to the UK phone set if needed.

Primary stress has been marked for polysyllabic words, again by lexical look-up. To achieve this, the stress pattern was mapped onto the phone sequence (taking into account the word's part of speech, for stress-pair words, like *conduct*). The vowels or diphthongs of monosyllabic words are also marked as receiving primary stress.

After the transcription process, all the data collected for blocks D to G have been manually annotated with phone and stress-level errors. Six trained linguists (teachers and students from the Language Unit and Linguistics Department at Leeds University) served as annotators. They were asked to correct any differences between the automatically annotated phone sequence and the actual utterance, by marking insertions, deletions, and substitutions at the phone level. Stress errors were also flagged, although only the primary stress in each word was noted.

Two teachers of English as a foreign language also made broad judgements of the proficiency level of each speaker using a scale from 1 (beginner) to 5 (fluent). These might subsequently be used for monitoring the error detection process and comparing quantity of errors with teachers' perceptions of speaker proficiency.

Annotators were familiarised with the UK phone set, and allowed to do some practice annotation on the first set of pseudo-speaker data. The most significant problem here was in asking linguists trained in phonetics to ignore their desires to achieve a very narrow transcription using the IPA symbol set and associated diacritics, and instead do a broader transcription onto the coarser phone set of the recogniser. Obviously, initial work was slower, but after practice, annotators were able to complete work on each speaker in 5-6 hours, making the total time for all annotation approximately 300 hours. In cases of clear non-native phone interference, (for example the Italian trilled [r]), annotators were allowed to select a phone from alternative non-English phone sets, but they were encouraged, where possible, to select the 'closest' match from the UK English phone set. In order to improve the annotation quality, this selection was additionally cross-checked by a trained phonetician, and native speaker of the speaker's mother tongue.

The annotators were encouraged to edit annotations attached to the waveform, if the speaker has said

something other than the canonical pronunciation provided in the reference transcription. Such changes consist of deletions, insertions and substitutions of phones or a stress shift.
Phone deletions are shown by replacing a unit's label with a zero. Insertions are indicated by appending a hyphen and the extra unit label to the leftmost neighbouring unit, or adding the extra unit label and a hyphen to the rightmost neighbour. Substitution is shown by simply replacing the label with the observed one.

Error Type	Reference Label	Edited Label	Example
Deletion	h	0	Dropped 'h' in *how*
Insertion	f	f-ax	Word-final schwa insertion on *beef*
Substitution	uh	Uw	*Book* rhyming with *boot*

Table 3: Examples for phone error annotations

The annotation tool does allow the boundaries between units to be moved, but the annotators were instructed not to change the boundaries, even if they were wrong, since this would disturb the integrity of the alignment between the three annotation levels.

Annotators could add comments between angle brackets < > for non-speech events or comments on the pronunciation as listed in the instructions, based on the conventions used in the SpeechDat Corpus markup.

Stressed vowels/diphthongs in polysyllabic words were labelled with a P (primary) in the reference transcrip-tion. These labels are deleted and moved to another vowel if the stress pattern differs from the norm. A full stop is used to mark an unstressed phone (either vowel or consonant). De-stressing a syllable can cause a vowel reduction, so amendments to the stress labelling are often associated with phone re-labelling. Secondary stress is not marked, so a word should be labelled with exactly one primary stress. Although monosyllabic words were labelled with primary stress in the reference transcription, annotators were instructed not to change their canonical stress pattern, even if they have become de-stressed in the context of the utterance. For example, in Figure 1, it can be seen that the word *and* retains its primary stress marker, even though it is likely to have been de-stressed.

Of course, for phone errors which are not related to stress errors only the phone level is amended. An example of this is shown in Figure 1, where a schwa has been added (by an Italian speaker) to the end of the word *beef*. In the reference transcription, the best fit achieved by the recogniser has (quite reasonably) subsumed the schwa into the initial vowel *ae* of the following word *and*. Here, the annotator would add the schwa label *ax* to the preceding or following phone.

Annotation consistency

In order to provide a measure of agreement between and among the annotators, the consistency of each judge relative to the others (the inter-judge scores) as well as the consistency of one judge to him or herself (intra-judge scores) was calculated based on a subset of the data which was annotated twice: once in the 'normal' annotation, and again by each of the six annotators. For this purpose five pseudo-speaker blocks of data were created in addition to the blocks of individual speaker data, by selecting some utterances covering all speakers. All annotators marked up pseudo-speaker 1 first, then annotated some of the individual speakers, with pseudo-speaker blocks 2-5 interspersed in the remaining work, but with the order rotated. Pseudo-data block 1 was used as a training session. It contained additional native speaker material that was not in the other pseudo-speaker blocks, in order to give the annotators a range in their mind to help them gauge how to annotate native to beginner accents.

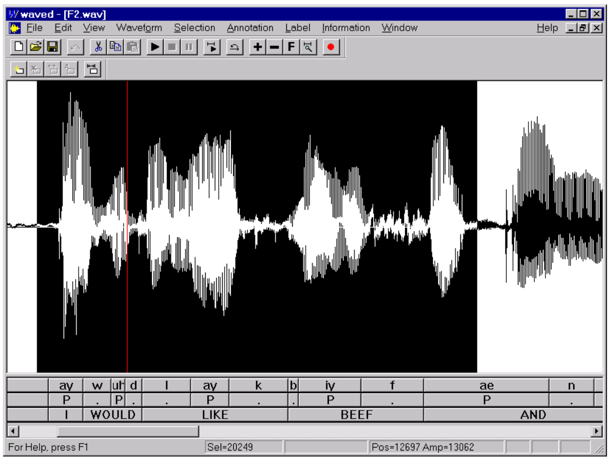

Figure 1: The annotation tool

Pairwise correlation of error identification between annotators can be classed for full hits (error is found in the same place, and diagnosed as the same error) or near hits (error is found in the same place, but diagnosed as a different error). Furthermore, the number of false alarms produced by a judge has been calculated. In Figures 2 and 3 below, near hits are drawn on top of the full ones. Annotators are identified by their initials.

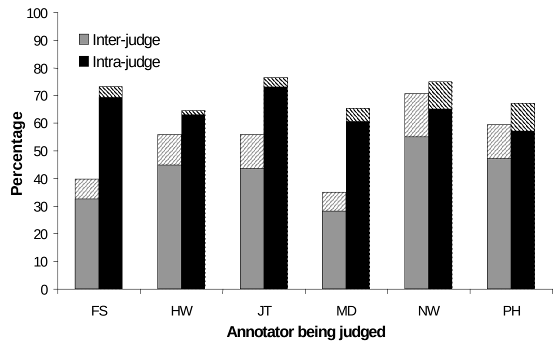

Figure 2: Hit rate across annotators

The first question to be asked of these data regarded the consistency of each judge relative to the others. In order to calculate the required hit and false alarm rate, each sentence from each block was examined pair-wise, comparing every annotator to all the others. For each phone in each word, it was assumed that the first annotator was the "golden" one, and had made the correct decision (leaving correct phones unaltered, and corrected errors when they occurred). Thus for each phone full hits, near hits, misses, correct rejections, and false alarms can be counted across all blocks, depending on whether another annotator agreed with the (correct-by-definition) decision of the first one. The mean number of near/full hits and misses is then calculated using as a denominator the number of instances in which (relative to the first annotator) there was an error. Similarly, the denominator for the number of false alarms and correct rejections is the number of phones for which he indicated there was no error.

Similar statistics can be computed for each judge relative to him/herself, because some subset of each block of pseudo speaker data was, in the normal annotation, assigned to that annotator. By comparing his/her decisions on the same sentences, one can clearly calculate the number of hits. Computing the number of false alarms is more problematic, however, since the already strange notion of a 'golden' annotator is strained. In order to avoid the difficult decision as to whether to 'believe' a

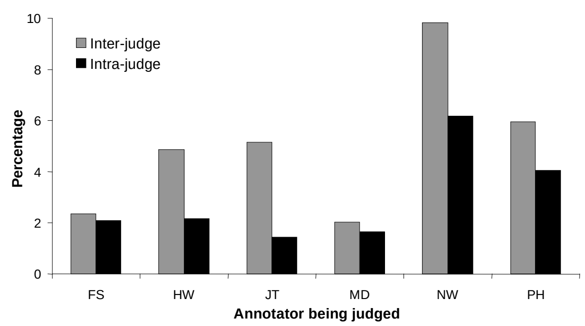

Figure 3: False alarm rate across annotators

selected annotator the first or the second time he annotates a given sentence, we shall consider false alarms and misses together, and report the mean as the effective false alarm rate. (This problem does not occur between judges, since we compute and report the false alarm rate in both directions, leading to asymmetry in the false alarm results).

Both, inter-judge and intra-judge agreement is rather low, with intra-judge scores being consistently better. Best hit-rate between judges shows only around 55% correspondence on deciding where and what an error is. Even localisation of the error alone shows at best a 70% agreement between annotators. This comes to no surprise and confirms the results on the consistency of phone-level annotations obtained elsewhere (Eisen et al. 1992). Consequently, the target one might reasonably set for diagnosis programs should be limited to only those errors which annotators agree on.

Data Characteristics

Although the corpus was intended to be balanced for proficiency, there is a clear difference between the language groups. The Italian speakers had an average of 0.54 phone errors per word with a standard deviation of 0.75, while the Germans had 0.16 phone errors per word with a standard deviation of 0.42. This difference is probably due in part to the greater phonological similarities between German and English than between Italian and English. Also, the actual types of errors also differed highly between the groups, with different phones affected, and different types of errors present on this phones (e.g., a higher incidence of phone insertions produced by the Italian natives).

Examples of pronunciation errors at each level, subdivided between German and Italian native speakers are given in Tables 4, 5, and 6 below. Annotators reported some difficulty in deciding which errors to mark at word level and which to mark as phone level – for example in the case of a spurious *s* being appended onto a noun or verb, it is difficult to decide whether the speaker is performing a systematic pronunciation error, or intending to pronounce a different word from the one in the prompt.

German		Italian	
prompt	was read as	Prompt	Was read as
not be	be not	Photographic	Photography
the	a	Than/then	That
month	week	Deserted	Desert
of	about	Like to	to like
-	more	-	The
in	-	To	-

Table 4: Examples of word level errors

In general, word level errors tend to be not systematic or easily predictable, whereas stress level errors have been observed largely as predicted. Phone level errors exhibit both, predictable (owing to L1 interference and attested in the EFL literature), as well as idiosyncratic behaviour.

German	Italian
′report	′photographic
′television	′convict / con′vict
‘contrast / contr’ast	′components

Table 5: Examples of stress level errors

German			Italian		
from	to	Example	from	to	example
oh	ow	Pr**o**duce	Eh	ey	s**ai**d
ax	ao	C**u**pboard	Eh	ae	b**e**d
uw	ao	Pn**eu**matic	Ae	ey	pl**a**nning
aw	ow	**Ou**tside	Ih	iy	t**i**cket
aa	ae	St**a**ff	Ay	iy	b**io**log**i**c**a**l
ih	iy	D**e**ssert	Oh	ow	
-	p	**P**neumatic	Ih	iy	
s	z	**S**aid	Ax	ae	
v	w	**V**isa	-	ax	sheep_
w	v	**W**eekend	-	hh	**h**onest
dh	d	**Th**e	Th	t	**th**in
-	w	Bisc**u**it	S	z	**s**leep
-	b	Thum**b**	Jh	g	**g**in**g**er
g	-	Finger	T	-	bai**t**
t	-	Desser**t**			

Table 6: Examples of phone level errors

Statistics extracted from the error-annotated corpus allow to identify the most common sources of English pronunciation errors for native speakers of Italian and German. Table 7 ranks phones according to their error sensitivity, i.e. the frequency with which a particular phone was affected by a pronunciation error. Table 8 shows the most productive phones, i.e. those contributing most to the overall number of errors. The data confirm the initial assumption that difficult phones in the language to learn are substituted with most similar phones of the native language of the speaker, or are deleted.

Schwa (ax) insertion accounts for approximately 6.7% of the errors Italian speakers ever made.

Short function words contribute considerably to the overall share of errors. Table 9 displays the words most frequently affected by errors, again in terms of their error productivity P (percentage of errors produce by this word) and their error sensitivity S (percentage of wrong occurrences of this word).

for Italian speakers			for German speakers		
uh	47.2%	often uw	z	20.4%	often s
ah	38.6%	often ax	ah	17.8%	often ax
er	37.8%	often eh-r	ax	15.9%	often uh
dh	37.2%	often d	v	13.9%	often f
ng	36.2%	often ng-g	zh	11.8%	often sh
ax	36.1%	often oh	th	9.5%	often s
ih	35.2%	often iy	dh	9.1%	often s

Table 7: Most error-prone phones

for Italian speakers			For German speakers		
ax	12.1%	often oh	Ax	21.7%	often uh
ih	11.9%	often iy	Ah	10.3%	often ax
ah	6.4%	often ax	Z	9.3%	often sh
dh	5.9%	often d	Ih	5.8%	often ax
eh	5.5%	often ey	T	5.6%	often -
d	5.0%	often d-ax	Dh	5.1%	often s
er	5.0%	often er-r	V	4.8%	often f

Table 8: Phones that account for most of the errors

of Italian speakers			Of German Speakers		
	P	S		P	S
a	8%	42%	To	9%	44%
the	6%	60%	The	8%	31%
to	4%	58%	A	6%	14%
said	4%	49%	Of	3%	27%
I	2%	18%	And	2%	31%
and	2%	55%	With	1%	41%

Table 9: Words that account for most of the errors

Conclusions

The ISLE corpus has turned out to be a limited but highly versatile spoken language resource for the development of pronunciation training tools for foreign language learning. It has been successfully used to

1. evaluate the accuracy on low perplexity speech recognition tasks, as they can typically be found in exercises for second language learners of English,
2. train and evaluate procedures for error localisation ,
3. evaluate rule-based techniques for diagnosing phone-level errors, and
4. evaluate procedures for word stress detection.

In constrast to other collections of learner speech (e.g. de Cock et al. 1997), the ISLE corpus comes with a highly detailed annotation, which includes word- and phone-level transcriptions, both canonical and actual.
Considering the extraordinary high costs involved in collecting and annotating such data (Ehzani and Knodt

(1998) estimated a price of up to one dollar per phone), the ISLE corpus makes a good compromise between investment and return.

Considerably more effort needs to be spent if a corpus is needed which is better balanced for age, sex and proficiency of speakers. Nevertheless, the available data gave a good impression about the extent to which annotators can be consistent with each other and with themselves in marking up speech with phone level errors using a broad transcription and a phone set tailored to the phonological inventory of the target languge. Furthermore it allows us to classify learners' pronunciation errors, relating them to mother tongue interference, difficulties in English phonology or to idiosyncratic learner behaviour. Errors have also been quantified, providing an indication of which are the most frequent, and therefore most deserving of the attention of pronunciation tutors, real or virtual.

The corpus will be available for non-commercial purposes through the European Language Resources Distribution Agency (ELDA).

References

de Cock, S., Granger, S. , Petch-Tyson, S. 1997. The Louvain International Database of Spoken English Interlanguage (LINDSEI) Project, rapport interne, Centre for English Corpus Linguistics, Université catholique de Louvain,

Ehzani, F. Knodt, E. 1998. Speech Technology in Computer-Aided Language Learning: Strengths and Limitations of a New CALL Paradigm. in: Language Learning & Technology, vol. 2, no. 1, p. 45-60.

Eisen, B., Tillmann, H. G., Draxler, Ch. 1992. Consistency of Judgements in Manual Labelling of Phonetic Segments: The Distinction between Clear and Unclear Cases, Proc. Int. Conf. On Spoken Language Processing ICSLP'92, p. 871-874.

Herron, D., Menzel, W., Atwell, E., Bisiani, R., Daneluzzi, F., Morton, R., Schmidt, J. A. (1999) Automatic Localization and Diagnosis of Pronunciation Errors for Second Language Learners of English. Proc. 6th European Conference on Speech Communication and Technology, Eurospeech '99, Budapest, vol. 2, p.855-858.

Hunt, J. 1996. *The Ascent of Everest*. Stuttgart: Ernst Klett Verlag, English Readers Series

Young, S., Kershaw, D. Odell, J., Ollason, D., Valtchev, V., Woodland, P. 1999. The HTK Book 2.2, Entropic, Cambridge.

Power, K., Morton, R., Matheson, C., Ollason, D. 1996. The graphvite Book 1.1, Entropic, Cambridge.

Appendix

Entropic GrapHvite UK Phone Set

Symbol	Example	Symbol	Example
Vowels		**Plosives**	
Aa	b**al**m	B	**b**et
Aa	b**a**rn	D	**d**ebt
Ae	b**a**t	G	**g**et
Ah	b**a**t	K	**c**at
Ao	b**ou**ght	P	**p**et
Aw	b**ou**t	T	**t**at
Ax	**a**bout	**Fricatives**	
Ay	b**i**te	Dh	**th**at
Eh	b**e**t	Th	**th**in
Er	b**ir**d	F	**f**an
Ey	b**ai**t	V	**v**an
Ih	b**i**t	S	**s**ue
Iy	b**ee**t	Sh	**sh**oe
Oh	b**o**x	Z	**z**oo
Ow	b**oa**t	Zh	mea**s**ure
Oy	b**oy**	**Affricates**	
Uh	b**oo**k	Ch	**ch**eap
Uw	b**oo**t	Jh	**j**eep
Semi-Vowels		**Nasals**	
L	**l**ed	M	**m**et
R	**r**ed	N	**n**et
W	**w**ed	Ng	thi**ng**
Y	**y**et	**Silence**	
Hh	**h**at	Sil	silence
		Sp	short pause