Using Real Words for Recording Diphones

Susan Fitt

Centre for Speech Technology Research
University of Edinburgh
s.fitt@ed.ac.uk

Abstract

This paper focuses on the creation of word-lists for making diphone recordings for speech synthesis. Such lists often consist of nonsense words, which has the advantage that the phonetic environment can be constrained, and it is easy to produce lists containing all possible combinations. However, this approach has the disadvantage that non-experts may find it difficult to read the nonsense-word transcriptions. For this reason, we investigate here the issues associated with the use of real words in creating diphone recordings.

1. Introduction

We have produced an accent-independent lexicon [1] based on Wells's keywords of English [2]. This is aimed at increasing the variety of synthesis and recognition possible, and thus increasing the acceptability of speech technology. If we are to synthesise different accents, it becomes increasingly likely that we shall need to record the voices of non-specialists in order to collect the diphones needed. Making it easier for non-specialists to record word-lists will also enable users to record their own voices for synthesis.

For obvious reasons it is more difficult for non-specialists to read nonsense-words than to read real words. Due to the non-phonetic nature of English spelling, nonsense-words for recording English diphones generally have to be presented as some kind of phonetic transcription in order to make sure that the right sounds are elicited. For example, there is no one spelling that corresponds to schwa, and there is no way to distinguish orthographically between the voiced "th" in 'there' and the voiceless "th" in 'thing'. So, anyone recording a word-list must first learn a set of phonetic symbols. Using real words for recording avoids this problem. However, we exchange this for different problems, which we investigate here.

Where necessary for clarity, keysymbol transcriptions are shown in vertical brackets, e.g. |th|. Keywords are given in small capitals, e.g. NURSE; with a bold highlight where necessary, e.g. COMM**A**. Example words from the lexicon are given in single quotes, e.g. 'use', while double quotes are used for other quoted information, e.g. "th"

2. Lexicon

2.1. Structure of base lexicon

The lexicon contains several features which aid the selection of appropriate words. Each entry consists of six fields separated by colons:

use:1:VB/VBP: { y * uu z } :{use}:194941

These are discussed briefly below.

2.1.1. Orthography

This field consists of the orthography in lower case.

2.1.2. Ordering/semantics

This is only included for homographs, words whose pronunciation varies by part-of-speech, or words with free variants. For instance, as well as 'use' (verb), our lexicon contains a second entry, 'use' (noun):

use:2:NN: { y * uu s } :{use}:194941

We have ordered these words in what we consider to be the most likely ranking; so, for instance, 'mow' ("to cut", belonging to keyword KNOW) is ranked number 1, while 'mow' ("to grimace", belonging to keyword MOUTH) is ranked number 2. For many words, though, the preference is not so obvious and the ranking is rather arbitrary.

For homographs such as 'mow' above, we include a brief semantic description in this field, while free variants such as 'either', whose first syllable may for UK speakers rhyme with either 'bye' or 'bee' are simply numbered.

2.1.3. Part of Speech

We use the Penn Treebank categories [3], derived partly automatically using the Penn Treebank tagger, partly semi-automatically by cross-checking for instance singular and plural nouns, and partly by hand.

2.1.4. Pronunciation

This is a complex field which we cannot describe in full here. However, there are several features which are important to note. The transcriptions in the base-lexicon are accent-independent, in that they form the basis of all the accents of English covered by the lexicon; at present this includes UK, US, Australian and New Zealand accents. The symbols used are keysymbols, a kind of meta-phoneme which describes the primary distinctions in English accents (see for example [1]). An example of a keysymbol pair is |@@r| and |er|:

NURSE { n * @@r r s }
PERT { p * er r t }

Although in many accents of English these words contain the same vowel, in Scottish English they are distinct, with 'pert' using a front vowel similar to the vowel in 'pet'. We differentiate between the two in the lexicon, so that we can use the distinction in synthesising Scottish English and simply ignore it in other accents.

The transcriptions include stress, syllabic and morphological markers. It should also be noted that the symbol set varies slightly from that presented in earlier papers reporting this work.

2.1.5. Enriched orthography

This field consists of the orthography broken down into the same morphemes as are annotated in the phonetic transcriptions. A certain amount of orthographic adjustment has been done here to aid morpheme matching, for instance 'using' has the pronunciation and enriched orthography:

{ y * uu z }.> i ng > :{use}>ing>

This field was derived by an algorithm matching the transcription with the orthography. Once an alignment was achieved, the root was automatically checked against other roots and morphophonemic rules, so that we were able to produce {use}>ing> rather than {us}>ing>.

2.1.6. Word frequency

An important element, particularly in the context of selecting words for diphone production, is word frequency, and this has been allocated a field in the lexicon. Word frequencies were derived from the collation of a number of on-line sources, including on-line newspaper articles and books. Extra weighting was given to spoken sources. The frequencies are based solely on appearance of orthographic strings, so homographs or other multiple entries are given the same frequency.

2.2. Post-lexical rules

The accent-independent base lexicon contains all the distinctive lexical information necessary for producing various accents of English. However, some important information is not explicitly transcribed in the lexicon; this consists primarily of allophones and variable rules.

So, in order to obtain accurate pronunciations, we must apply post-lexical rules to the base transcriptions. An example of an obligatory allophone is t/d-tapping in American English. For 'waiting' this would give us

{ w * ei t }.> i ng > → { w * ei t^ }.> i ng >

Variable rules include the glottal stops which many accents of English use to replace |t|; many also use |i n| or a syllabic |n| to replace |i ng| in words such as 'waiting'. So, for 'waiting' in UK, we have the options:

{ w * ei t }.> i ng >	no rules
{ w * ei t }.> i n > or { w ei t }.> n >	"-ing" rule
{ w * ei ? }.> i n >	glottal rule

etc. Of course, for US accents, we could combine the "-ing" rule with the tap rule.

Two points should be noted here. Firstly, the rules apply either to single words or to complete strings. Some rules apply across word-boundaries, for instance in non-rhotic accents |r| is deleted phrase-finally in 'far', but retained before a vowel in 'far away'. So, to form running text, we need to first concatenate the base forms and then apply the rules. Secondly, output retains some redundancies. The output is transcribed in what we term "basic keysymbols", such as the NURSE - PERT distinction, which is redundant in many accents, and "output keysymbols" (allophones etc.), which are accent-specific.

3. Selection of words

As we have seen, although the master lexicon contains the *necessary* information for describing the different accents of English, the lexicon alone does not contain *sufficient* information. For synthesising different accents we must convert it into accent-specific lexica. We cannot use a single word-list for recording all accents, since the accents use different allophones and variable rules.

3.1. Target diphones.

If we take an RP output of the lexicon, there are 69 distinct keysymbols, not including stress, syllable boundaries or morphological markers. Figure 1 (unreduced symbol set) shows the distribution of these symbols. The most frequent, |i| as in KIT, has about 72,000 occurrences; it is followed by |s|, |t|, |n|, and |@| (COMM**A**). The least frequent are |oou| (ADI**O**S, 183), |i@| (ID**EA**, 99), |x|, (LO**CH**, 51) |ll| (**LL**ANDUDNO, 7) and |oir|, with just one occurrence in COIR. Adding in word ends, this gives us a theoretical target set of 71 x 71, i.e. 5041 diphones. However, there are complications (see for example [4]); some of these are discussed below.

3.1.1. Syllable positions

There are a number of restrictions on syllable positions in English, which reduce the diphone target set. However, few of these are absolute, and some of them vary by accent. |ng|, for example, is usually only found after short vowels, but in our lexicon we have 'munchen' ("Munich") { m * uu ng . k @ n }, 'oink' { * oi ng k } and 'boing' { b * oi ng }. Also, |ng| can generally occur syllable-finally, as in 'sing', or in final clusters such as 'sink'. But, in Birmingham and Liverpool, some speakers realise final |ng| as |ng g|, with words such as 'sing' pronounced as { s * i ng g }, and 'singer' { s * i ng g }.> @r > rhyming with 'finger' { f * i ng . g @r }. If we make a general specification that we require |ng| + vowel diphones, we will not find them in an output lexicon of these accents.

However, it is not enough to scour the output lexicon for adjacent pairs. There are numerous pairs, especially vowel-vowel, which do not occur in the lexicon but are needed across word-boundaries, for example |ei . ei| in 'day eight'. Some allophones only appear at word joins. So, we need to consider word-pairs. Note that we must join the word pairs before applying the post-lexical rules to derive the allophones. Since the conditioning environment for these allophones may extend over several segments, we need to be careful to select adequate word pairs.

Also, future additions to the lexicon, especially foreign words or names, may add new pairs to the lexicon. This is more problematic and suggests that we may need to record nonsense words for the missing diphones, or risk being unable to synthesise new words. This is a problem for our aim of facilitating recordings by untrained speakers.

3.1.2. Co-articulation

Co-articulation and other phonetic phenomena favour the inclusion of separate diphones for certain clusters or syllable positions. For example, the |r ii| pair in 'be**rea**ving' is not the same as the |r ii| pair in 'ret**rie**ving'. Syllable position may also have an effect on some segments. Recordings with just one |i t| diphone, taken from the word 're**it**erate', i.e. across a

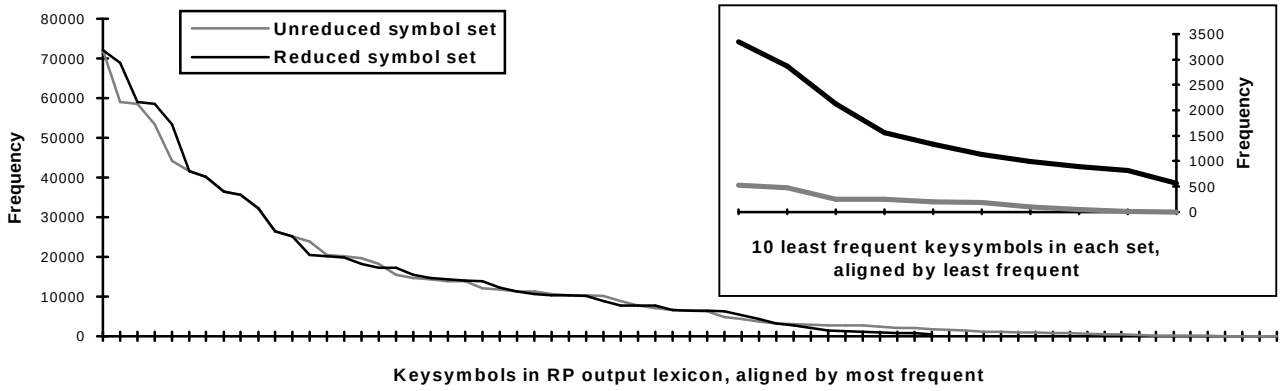

Figure 1: Frequency of keysymbols in RP output lexicon – full set and reduced set.
Inset: Figure 1a: Frequency of 10 least frequent symbols in each version of lexicon.

syllable boundary, do not produce a natural synthesis of |t| in codas, such as 'sit'. So, we included initial clusters with stops or |s| in our target diphone list, and both syllable boundary conditions for stops.

3.2. Phonetic environment

The first consideration in selecting words was phonetic environment. We attempted to take all symbol pairs from the middle syllable of a trisyllabic word (unless the targets included word boundaries), stressed (except for schwa or syllabic consonants), and not in consonant clusters unless these are specifically required. /iy/ is a potential exception to these criteria; it most often occurs word-finally, in words such as HAPPY, and in this environment it is more stable than in trisyllables such as 'an**y**way'.

Out of about 118,000 words there are around 35,200 trisyllables; of these, about 12,000 have the primary stress on the middle syllable, such as 'absurdest', 5,300 have schwa in the middle syllable, such as 'acronym', and about 2100 have a syllabic consonant in this position, such as 'battlefield'. On the face of it this would seem a reasonable number of words from which to select suitable candidates for recording.

However, many of these words are related and so contain similar strings of symbols, for instance 'civilise', 'civilize', 'civilised' and so on. Taking our 19,400 candidate trisyllables, e.g. 'civilise', we get:

{ s * i . v l }.> ae z >

Removing morphological markers and taking the central syllable plus the adjacent symbol on each side, we get:

i . v l . ae

There are around 9,300 of these; still apparently a good quantity for selection.

Of course, the problem is that these words do not provide an even distribution of symbols; there are some diphones which do not appear at all in this set, for example |th * @@r| as in 'thirst'. There are some which only appear in less than ideal contexts, for example |l * @@r|, whose only instance is in a cluster in the word the rather unlikely word 'deblurring'. Others account for a large quantity of the set, for example |s * e| as in 'ascending' with 269 examples.

3.3. Word-level criteria

It quickly became obvious that, even where the diphones were found in the right context, the phonetic criteria alone did not produce a good selection of words. An obvious case is homographs, or words with varying pronunciations such as 'economics', which we avoided altogether. Sometimes, though, recording a word-list will reveal a homograph not yet listed in the dictionary; furthermore, a homograph will occasionally be the only instance of a particular diphone.

Other types of word may also cause difficulty for speakers. For example, if we run through the trisyllable set in alphabetical order, the first word to produce the sequence |b @| is 'ababa' { * a . b @ . b @ }. It fulfills the phonetic criteria, but fails on the grounds of recognisability. In a lower-case lexicon, with no context, it is likely that some speakers will not recognise it and will mispronounce it. Much better is the next example, which is 'abacus' { * a . b @ . k @ s }. One preference, then, is to avoid proper names. Another point to note is that selecting alphabetically is not ideal as it will result in a bias towards the beginning of the alphabet; it is preferable to have a spread of words so that if a diphone is not produced well in the recordings, it is more likely that it can be found in another word.

Word frequency is an important factor in asking speakers to record words. For example, words in the lexicon with zero frequency include 'amphitheatral', 'solemniser' and 'volatilizable', which a speaker might stumble over or not know how to pronounce. On the other hand, the most frequent words are 'the', 'i', 'and', 'to' and 'you', which are also poor selections. Many of this will be disallowed in any case as they have two entries, for stressed and unstressed pronunciations, and they will also be dispreferred as they are monosyllables. The most frequent trisyllables as specified earlier are better candidates: the list starts with 'usually', 'another', and 'probably'. 'Probably' would then provide the |b @| diphone rather than "abacus", though 'probably' is not ideal as it is subject to reduction of the middle syllable.

For every extra consideration, there are a few diphones which no longer appear in the acceptable word set. At this point, we need to form priorities rather than absolutes, for

example a proper name may be acceptable if no other instance of the diphone is found. On the other hand, if the only target diphone which fits the phonetic criteria is found in a homograph, it is better to take the diphone from a non-trisyllabic word, or across a word-boundary.

3.4. Word pairs

Word-pairs were formed by making five combinations for each diphone, using the most frequent words containing the right symbols, in adjective-noun, noun-noun or verb-noun pairs where available. This basic formula resulted in some unlikely combinations ('low-fat chair'), some dubious ones ('definite ooziness') and some better ones ('parrot eater'). For each combination a selection was made by hand.

4. Summary and Evaluation

Words were selected for diphones as follows: the words were ordered by frequency, then we took the first example of each diphone in a desirable environment. If there was no such example, we resorted to less desirable environments, for example taking |th @@r| from 'thirst'. At this stage we did not record diphones not found in the output lexicon.

Synthesis from diphones recorded in this manner was found to be of good quality (with the caveat that we have not made a direct comparison with diphones taken from nonsense words). Examples can be found on the project web pages (http://www.cstr.ed.ac.uk/projects/unisyn.html). The following points were identified as areas for improvement.

4.1. Context

After initial recordings it was evident that we needed to attach more weight to the phonetic environment around the diphone, even if this would mean we then had to select diphones from word-pairs rather than single lexical entries. For example, nasal co-articulation extends for a considerable distance across neighbouring segments, so words containing nasals are best avoided.

4.2. Missing diphones

We also need a global algorithm for identifying which of the potential diphones not found in the lexicon might be found in future words. We already have feature classifications and consonant cluster rules for the keysymbol set, and this should provide a basis for generating valid missing diphones, which would have to be recorded in nonsense words.

4.3. Word pairs

An automated way of making acceptable word-pairs is needed, perhaps by selecting them from running text. We are also currently lacking an automatic way of identifying input for generating cross-word allophone pairs; it is not practical to generate all word pairs (118,000 x 118,00 words!), apply the rules and then select allophones; we need to pre-select words which contain the right environment.

4.4. Checking variation

For the current work we checked variable rules with the speakers. For untrained speakers working without phoneticians we would need a way of doing this automatically. One possibility is to use rhymes, so a speaker from Birmingham might be asked, "Do you rhyme 'singer' with 'finger'. For some variables such as glottal stops this approach could be difficult to apply; we could investigate whether speech recognition might provide a solution.

4.5. Mispronunciations

A certain amount of mispronunciation is inevitable; for instance, as noted above, some speakers may pronounce 'probably' with two syllables rather than three. It is very difficult to avoid this in all cases unless the recording is monitored by a phonetician; one way to guard against such mispronunciations is to try to include more than one example of each diphone in the word list. Another way of improving the accuracy rate is to generate several candidate words for each diphone for the word-list, and hand-edit to weed out poor selections. This, though, does require the input of a phonetician. Another possibility would be to write reduction and elision rules, and disallow any words which fell within their scope. Some commonly mispronounced words, such as 'skeleton' (pronounced as if it were "skellington"), or 'nuclear' (pronounced as "nucular") could perhaps be marked in the lexicon as words to avoid.

4.6. Mapping redundant symbols

We wished to avoid collapsing non-distinctive keysymbols in the output lexica, such as NURSE-PERT in RP, as this requires an extra stage of processing and in a few cases disallows valid choices by the speaker, such as use of |x| as in LOCH. However, we must conclude that, particularly with some symbols having such low frequency, removing redundancy is preferable. For RP this reduces the symbol set from 69 to 49. Including word ends but not including clusters or different syllable positions, we have a maximum set of 2601 pairs, reduced from the previous 5041 – a considerable gain. Furthermore, as we can see in Figure 1a, the most gain is in the least frequent symbols; the symbol with lowest frequency is now |zh| with 563, followed by |dh| with 821 and |ur| with 888. This greatly increases the chances of finding suitable words containing the necessary diphones.

This work is funded by the UK Engineering and Physical Sciences Research Council through grant EPSRC GR/L53250.

5. References

[1] Fitt, Susan, and Isard, Stephen (1999). Synthesis of regional English using a keyword lexicon. *Proceedings: Eurospeech 99*, Vol. 2, pp. 823-6.

[2] Wells, John C. (1982). *Accents of English*. Cambridge: Cambridge University Press.

[3] Marcus, Mitchell, P., Santorini, Beatrice, and Marcinkiewicz, Mary Ann, (1993). Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*, Vol. 19, No. 2, pp. 313-30.

[4] Black, Alan W., and Lenzo, Kevin A. Diphone databases. http://www.festvox.org/festvox/

[5] Lenzo, Kevin A., and Black, Alan W (2000). Diphone collection and synthesis. *Proceedings: ICSLP 2000.*