Towards Automatic Lexicon Expansion

Jason Taylor

Word Count=6850

1st Year Report for Doctor of Philosophy

University of Edinburgh

2018

Abstract

Text-to-Speech Synthesis (TTS) and Automatic Speech Recognition (ASR) typically rely on a lexicon to ensure correct word pronunciations. In deployed systems, the lexicon must be updated regularly for neologisms, especially new foreign names and abbreviations. While necessary, this process is manual and expensive. My PhD aims to develop reliable automatic methods to expand the `Combilex` Speech Technology Lexicon. After verifying the potential of using abstract metaphones directly in TTS voice-building, I propose expanding `Combilex`'s accent-independent, Base-form (BF) lexicon. To this end, I discuss preliminary results of Grapheme-to-Metaphone (G2M) tests and other potential solutions for expansion.

* * *

I am the recipient of an ESRC-funded studentship via the Scottish Graduate School for Social Sciences. Under a mutual agreement with the regional body I completed 60 credits of research courses in the School of Social and Political Science during semesters 1 and 2 of the academic year 2017-2018. Alongside my studies that started in mid-October 2017, I was employed as a TA for the Speech Synthesis, Speech Processing and Accelerated Natural Language Processing courses as well as a marker for the latter two.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Jason Taylor*)

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction to Speech Technology Lexica

1.1 The Pronunciation Lexicon's Value in ASR and TTS

A pronunciation lexicon is a computer data object containing the spelling and pronunciation of words (1, p.207:8). Text-to-Speech Synthesis (TTS) and Automatic Speech Recognition (ASR) typically rely on a lexicon to ensure correct word pronunciations. Creating and expanding high quality lexica is however an expensive process. This is because human linguists are required to devise new entries manually. The goal of my PhD is to develop reliable, automatic methods for expanding the pre-existing `Combilex` Speech Technology Lexicon (2).

The lexicon remains necessary both commercially and in research for ASR and TTS alike. Recently developed paradigms aimed at removing its requirement directly convert between speech and text by means of neural sequence-to-sequence models. In ASR, such end-to-end models (3; 4) have achieved performance comparable to that of traditional, multi-component, WFST-based systems. However, (5) indicates a minimum of 12,000 hours are required to deliver such state-of-the-art performance. This is approximately 10 times the size of large, free-to-use datasets like the Multi-Genre Broadcast CHallenge (MGB) data (6) or LibriSpeech (7) dataset. Owing to the cost of acquiring and processing such quantities of data, high quality end-to-end ASR models are beyond reach for many researchers and developers.

Gathering sufficient speech data to build an end-to-end TTS model is also problematic. The original Tacotron system (8) is trained on 24.6 hours of a single female speaker for example. Moreover, TTS is a de-compression process: its input (text char-

acters) carries much less information than is expected of its output (a waveform expressing nuances of human speech). For instance, the text string *convict* should be synthesised with distinct pronunciations depending on whether it is a verb or a noun. When a system is deployed, the ability to control the pronunciation of such homographs is essential. End-to-end TTS (9; 10; 11) struggles to disambiguate homographs, resorting to a lexicon to phonetize graphemic input. On a wider note, automating and simplifying the TTS pipeline makes controlling pronunciations complicated. For instance, to ensure the correct pronunciation of digits like *1999*, 'verbalisation' is carried out whereby digits are spelt out how they should sound, e.g. *nineteen ninety-nine*. This requires creating a complex number grammar. These problems with end-to-end lessen their appeal, meaning the lexicon is still used in DNN and Unit-Selection/Hybrid TTS systems as they remain preferable for deployment.

1.2 Some English Pronunciation Lexica

Figure 1.1 shows a sample entry of the word *absorbing* in two English pronunciation lexica: the Carnegie Mellon University Pronouncing Dictionary (12), or CMUDict, and `Combilex`. The CMUDict is the most commonly used open-source lexicon: it's pronunciations are in a General American accent (GAM) and were collected via crowdsourcing.

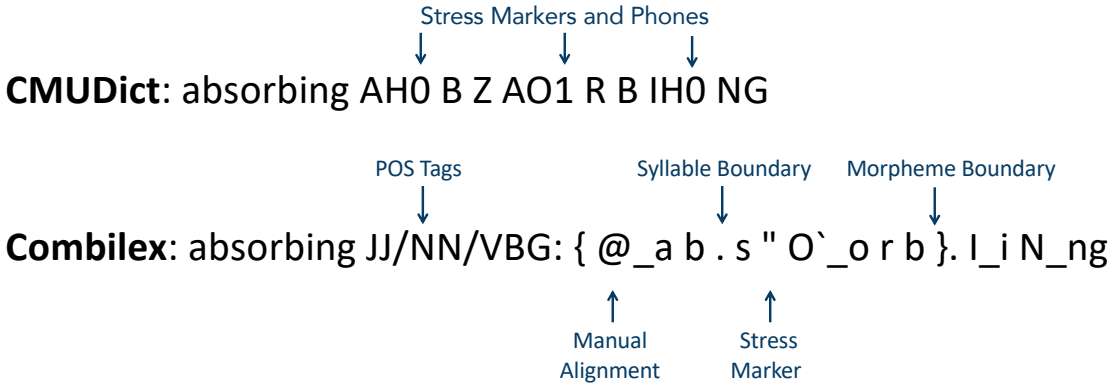

Figure 1.1: Example Entry from the CMUDict and `Combilex`

`Combilex` is a commercial lexicon developed at CSTR and comes in a variety of accents including Receieved Pronunciation (RP), GAM and Edinburgh Scottish. These surface-form accents are generated from an abstract, Base-form (BF) lexicon which is explained in more depth in section 2.1. Approximately 15% of Combilex's entries

(~20,000) were manually crafted by a single linguist, with the remainder selected via morphological composition as described in (2). Consequently, the CMUDict contains more proper names than Combilex.

As shown, lexica can optionally store linguistic precepts beyond the phonetic level, such as lexical stress markers. `Combilex` also contains syllabification, Part-of-Speech (POS) tags and morpheme boundaries. These are useful for front-end TTS tasks such as POS tagging and homograph disambiguation (13). An advanced feature of `Combilex` is that given a single entry, multiple morphological derivations can be generated via finite-state transducers (FSTs). These FSTs offer words that when selected by hand, add complete linguistic entries to the lexicon.

1.3 PhD Aim

`Combilex` currently contains 141,341 word pronunciations. Although marginally higher than the open-source CMUDict with 135,091 entries, it is much smaller than Google's internal lexicon which holds 435,000 pronunciations according to (14).

Expanding `Combilex` is desirable to compete with industrial sized lexica. However, its compilation thus-far has already been costly. Usually, linguists manually check all new words during a lexicon's expansion. Commercially this must occur regularly to keep up with the ever-changing language use of customers (especially, unseen foreign names and abbreviations). Therefore, my PhD aims to develop automatic methods to expand Combilex reducing the time linguists spend verifying new entries.

The first step in lexicon expansion is deciding how pronunciations should be represented. There is never a single pronunciation of a word. It depends on a speaker's demographic profile (geographical origins and social factors, see (15)), the speaker's environment (including interlocutor identity: for instance machine or human, see (16)), the word's phonetic context (e.g. *the* sounds different when pronounced before a consonant or vowel), and nuanced, hard to quantify articulatory phenomena (like speech speed and the *warmth* of the vocal tract). The first question I set out to answer therefore was how should I choose the representation for new `Combilex` pronunciations?

Chapter 2

The Accent-Independent Lexicon

2.1 Keyword Metaphones

How should a single pronunciation of a word be decided? The notion of accent-independence employed in lexica like `Combilex` (2) and Unisyn (17) involves storing abstract pronunciations based on evidence of phonological variation in English. According to (18), phones found in certain keywords have separate realizations in different accents of English. The list of keywords used in `Combilex` may be found in (19).

Rather than using a set of phones to reflect a speaker's pronunciation, `Combilex` stores entries with a super-set of BF metaphones. For example, Figure 2.1 shows the `Combilex` metaphone string for *whiter* in grey. The 'W' metaphone was crafted as it has been noted that in different parts of the English-speaking world the first phone uttered in *whiter* varies. For instance in Scottish English, it is the 'W' phone, or the voiceless labio-velar approximant [ʍ] (IPA). In RP, it is the 'w' phone, or the voiced labio-velar approximant [w].

The advantage of storing entries in an accent-independent manner is new surface-accent lexica can be generated with only the FST rules to convert the metaphones to phones of the surface-accents in question. However, gathering the specific phonology of a variety of speech still requires expensive manual intervention, albeit less than would involve development of entirely new lexica from scratch for every new dialect in a language. Therefore it will be more efficient to add new words to the BF lexicon rather than to each surface-form lexicon individually.

This led me to ask how valuable is the BF lexicon compared to it's surface-form descendants? In deployed TTS systems that reflect an individual's speech for instance,

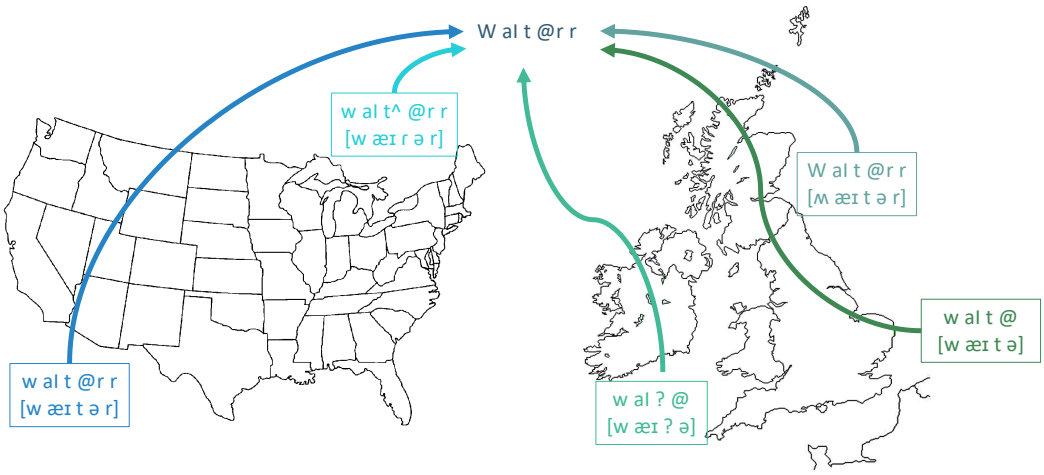

Figure 2.1: General Pronunciations of 'whiter' abstracted from accents to craft meta-phone sequence in grey (first line of variants in `Combilex` phones, second line in IPA).

specific tailoring of word pronunciations and phonesets is good practice. However, this step is time-consuming and often lexica are used off-the-shelf, even if there is a mismatch in the accent with the voice talent's speech. In a scenario where only a GAM lexicon is available with RP speech, foreseeable pronunciation issues in the TTS output arise. For instance, GAM English has a single pronunciation for *whiter* and *wider* (with a voiced alveolar tap), while Received Pronunciation (RP) has two (voiceless and voiced alveolar stop respectively). If a TTS system were fed only a single phonestring for these two words, how could it maintain the pronunciation distinction as would be proper in RP speech?

Instead, could better quality TTS be obtained by the accent-independent, BF lexicon directly? This maintains distinctions lost during its transformation to surface-forms. Furthermore, accent- and speaker- specific pronunciations of the voice talent would map onto fluid meta-phones rather than strict, phonetically-defined phones during voice building. Therefore, the accent-independent lexicon could fit to pronunciation variation better than a lexicon whose accent is mismatched with that of the voice talent. As described below, I built such systems and put them to listening tests to answer these questions.

2.2 TTS Experiments

I aimed to uncover how a mismatch between the accent of the lexicon and the speaker data affected DNN-based TTS quality. Due to the complex ways in which a lexicon

exerts an effect, I conducted experiments from two methodological angles. First, to broadly gather how a mismatch performs in comparison with a matching accent in the lexicon and voice data, I held forced-choice A/B preference tests on randomly selected stimuli. Second, similar to the argument of (20), I chose to focus comparison on specific differences between the output of two systems. In this case, every surface-accent lexicon contains two words with the same pronunciation in that accent but with two separate pronunciations in another accent. For instance, the *whiter* and *wider* pair explained in the previous section. I checked whether listeners could distinguish between these words by conducting an intelligibility test of the word pairs with a mismatched accent in the lexicon and voice data. I conducted the same experiments using the `Combilex` accent-independent keyword lexicon next.

In the setup it is usual to include phonetic input features to the DNN acoustic and duration models such as place/manner of articulation, voicing, tongue height/frontness and lip-rounding (21). To avoid pre-determining how metaphones could be realised phonetically, I could not define articulatory features for them. This could have posed a problem in comparing the effect of `Combilex`'s BF lexicon to surface-forms such as RP and GAM, as the number of input features had to be different for the BF and surface-form lexica. To directly compare the BF lexicon with an RP lexicon, I had to ensure phonetic features did not have an effect on TTS output quality. Hence I conducted a preference test on voices built with RP speech data and the RP lexicon with and without phonetic questions prior to testing the BF lexicon.

2.2.1 Voices

To match and mismatch the accent of the lexicon with that of the voice talents I used RP and GAM voice data. The voices were built from 2600 utterances of a male RP speaker recorded for the Hurricane challenge (22) and from 1131 utterances of the Arctic RMS US male speaker recordings. All training and test utterances had a 16kHz sampling rate.

I built voices utilising DNN acoustic and duration models with Merlin (23). The DNN models were 6-layer feedforward networks with 1024 hidden units per layer using the tanh non-linear activation. Stochastic gradient descent was used for optimisation with a learning rate of 0.002 for 25 training epochs. The WORLD vocoder (24) was used for acoustic feature extraction and synthesis of test waveforms. With phonetic features, the input vector size was 772, which reduced to 632 when removed

for the experiments with the BF lexicon. These sizes include all the frame features specified in (25).

I modified Festival to phonetise text in training and test utterances using `Combilex's` BF, RP and GAM lexica. The training utterances were aligned to their speech recordings using the default HTK 5-emitting state aligner bundled with Merlin.

2.2.2 A/B Listening Tests

I performed 4 forced-choice A/B listener tests in total. Each test had 20 randomly selected sentences synthesised by the 2 voices (40 stimuli each test). All utterances contained between 4 and 15 words (all in lexicon), and the order of systems for each question were randomised. Listening tests took place in soundproofed booths with headphones. To ensure a sufficient number of listeners (26), I recruited 41 participants paid £8 each for approximately 45 minutes. Each of the tests collected 20 preferences from all participants, giving a total number of 820 per test. The binomial significance test as implemented in Scipy was performed on the aggregate of all answers on each section to establish the probability participants selected one voice over another due to chance rather than a controlled change. The sequence of tests aimed to find out the following respectively:

1. Do listeners notice a difference between voices built on RP speech data using GAM versus RP surface-form lexica?

2. Do listeners notice a difference between voices built on GAM speech data using RP versus GAM surface-form lexica?

3. Do listeners notice a difference between voices built with and without phonetic input features?

4. Do listeners notice a difference between voices built on RP speech data using the RP versus BF lexica?

2.2.3 Intelligibility Test

I conducted an intelligibility test to demonstrate how a mismatch in accent between the surface-accent lexicon and voice data can lead to a loss of word distinction retained in the BF lexicon. I asked participants to type the final word of utterances played to them. Each utterance was synthesised and tested with each word in a predetermined set of

word pairs. For the RP speech data mismatched with the GAM lexicon, I used pairs of words where the voiced/voiceless distinction is not found on alveolar stops in the GAM lexicon and where a voiced alveolar tap is used instead. The utterances played to the listeners were: *this shirt is wider*/*whiter* and *she was a graceful rider*/*writer*.

To test the GAM speech data mismatched with the RP lexicon, I synthesised utterances with words showing a separate difference between the accents. There are pairs of words in GAM distinguished by whether the pronunciation is rhotic or not (27), where only a single non-rhotic pronunciation is used in RP. I therefore asked participants to type the last word in: *show me the sauce*/*source* and *Henry has red paws*/*pores* to test whether the distinction is maintained.

For comparison, listeners were also played the utterances synthesised from systems with the BF lexicon and where the accent of the lexicon matched the speech data in question. Since the distinctions are maintained in the correct surface accents and the BF lexicon, it was expected the intelligibility of the words with these lexica to be much higher than with a mismatched lexicon.

2.3 Results

2.3.1 A/B Listening Tests

Figure 2.2 shows results for RP and GAM speech data trained with RP and GAM lexica. For the RP voice (the top bar), the preference for the RP lexicon was strong with a p-value < 0.0001. Overall, speech with the RP lexicon sounded crisper and more natural, probably because the alignments between the recorded RP speech and the phonetised text were more consistent with the RP lexicon than with the GAM lexicon.

Despite the RP lexicon preference, there are instances where DNNs learn characteristics of the RP speech not specified in the lexicon. Case in point: accent-specific sounds are synthesised for which no explicit phone is present in the input string. For example, in the GAM lexicon there is no phone corresponding to the voiced palatal approximant ([j] in IPA), which in the RP training data frequently occurs after an alveolar stop and the GOOSE vowel in the word *residual*. Yet the synthetic RP speech contained the sound in this position even when the GAM lexicon was used. Likewise if there is an extra phone in the mismatched lexicon not uttered by a speaker in the speech data, the acoustic model may learn that the acoustics are no different. For ex-

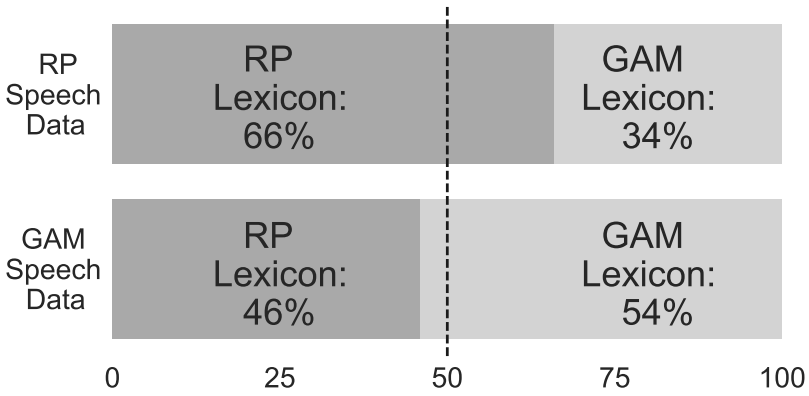

Figure 2.2: Preference percentages of stimuli for voices trained with correct and incorrect surface-form lexica

ample in the GAM lexicon *court* is "k Or r t", but since in these phonological contexts during training the RP speaker did not pronounce a rhotic sound, test utterances do not contain the rhotics either. Consequently I found during some informal listening that *court* sounds homophonous with *caught*, as it should in RP speech despite two different phonestrings in the GAM lexicon.

The above cases demonstrate ways in which DNNs adapt to RP speech when fed phone information from a GAM lexicon. However, if phonological contexts missing in the training data are synthesised at test time, the network outputs less convincing acoustic parameters. For example the metaphone sequence "m A r" in the word *Mario* was not seen in the training set, and during more informal listening I discovered the vowel was reduced to a schwa-like sound when synthesised.

My overall impression is that the GAM lexicon with the RP speech sounds acceptable, but is clearly inferior to using the RP lexicon. The reader may listen to samples of each voice online[1].

When a voice was built to match the GAM lexicon to the GAM speech data (lower bar of Figure 2.2 it was significantly preferred to using the RP lexicon with a p-value of 0.023. The synthesis of two words show separate reasons why listeners did not prefer the RP lexicon. The first is because the RP pronunciation can sound unnatural with GAM speech data: *water* contained the voiceless plosive instead of the voiced alveolar tap found in GAM speech. In the other word, *competition*, the schwa found between the labial and alveolar plosives was pronounced with a rhotic. This is because

[1]Visit homepages.inf.ed.ac.uk/s1649890/baseform/ for samples

GAM speech is rhotic, but the RP lexicon does not possess placeholders for rhotics after schwa. This explains why the schwa was aligned to a schwa followed by a rhotic in the GAM training speech, which in turn accounts for the rhotic sound synthesised with the RP lexicon in *competition* (see listening samples webpage). Overall, however, the utterances with both lexica were intelligible, although as with the RP speech data, the matching surface-accent lexicon (GAM) was preferred with the GAM speech data.

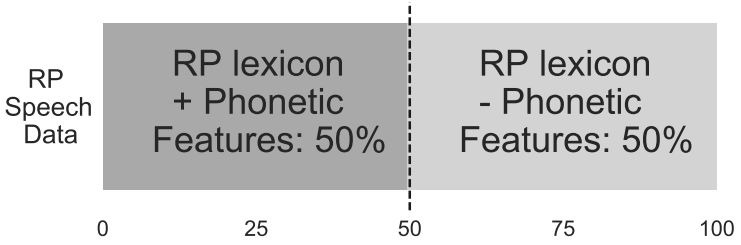

Figure 2.3: Preference percentages of stimuli synthesised with voices built with and without phonetic features

Figure 2.3 shows an equal preference between voices trained with and without the phonetic features, meaning they are redundant when building a voice with our RP speech data. I wanted to confirm this was the case to justify using an accent-independent lexicon with metaphone transcriptions with no strict phonetic realisation. In addition, it justifies removing the phonetic features from the voice built with the RP lexicon for the final AB test with the BF lexicon. It also allows for comparison between the results with the RP speech data with the GAM lexicon in Fig 2.2. This means the accent-independent BF lexicon can be compared with the matching RP lexicon and mismatching GAM lexicon.

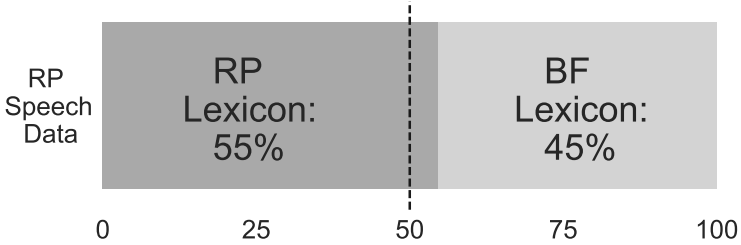

Figure 2.4: Preference percentages of stimuli for RP voice with RP and BF lexica

Figure 2.3 shows the preference between the RP voice built with the RP and BF lexica. The RP lexicon was slightly preferred with a significant effect with a p-value of 0.009. This effect size is much smaller than the comparison with the GAM lexicon

with the RP speech data. Indeed, the synthesised utterances with the RP and BF lexica sounded very similar. The breakdown of preferences for each question is shown in Figure 2.5, which shows a varied pattern of preference for each lexicon (sometimes BF is preferred and sometimes RP).

Analysing in closer detail utterances where the RP lexicon was preferred, a pattern is found with the BF lexicon: all the metaphone inputs contained rhotic placeholders. For instance, in utterance 1 the word *their* has the metaphones "D @r r" (where @ is schwa) in the BF lexicon and the phones "D @" in the RP lexicon. Alignment accuracy is affected by the presence of this 'empty' metaphone as it occupies minimally 25 milliseconds of duration in the training waveforms. This is because the aligner must emit at least a single 5 millisecond frame from each of its 5 states. In addition the rhotic is allotted a minimum duration of 25 milliseconds by the DNN duration model at test time. The acoustic effect of longer durations is a prolongation of the preceding vowel: schwa. The utterance with the RP and BF lexica would sound almost identical were it not for the longer vowel durations which were only identified in extremely quiet lab conditions. The same is true of words in utterances 3, 6, 13 and 17 with the rhotic metaphones present in the BF lexicon's *clear-cut*, *resources*, *injured* and *radar*.

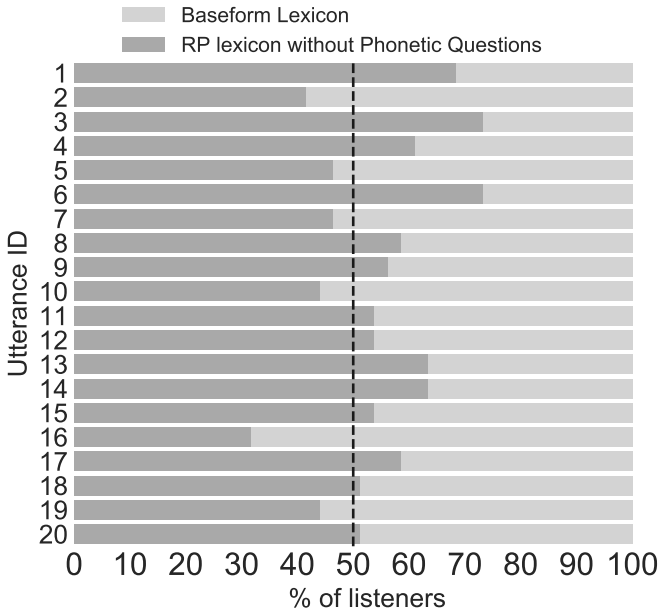

Figure 2.5: Preference Results each utterance in BF test

These are small and subtle differences I think will be easily resolved by modifying the alignment and duration models' minimum phone duration thresholds. I am currently testing whether this fixes the issue. Nevertheless, overall the BF lexicon still

appears more promising for automatic tailoring to a voice talent's speech than mismatched surface-accent lexica.

2.3.2 Intelligibility Test

Figures 2.6 and 2.7 show the results of the intelligibility test for the RP and GAM voices respectively. Each bar represents a percentage of 82 stimuli - 2 responses per 41 participants - in which the correct word was typed. Two responses are grouped into one bar, because listeners heard both words in the pair individually. Each pair of bars is grouped by the lexicon used to build that voice.

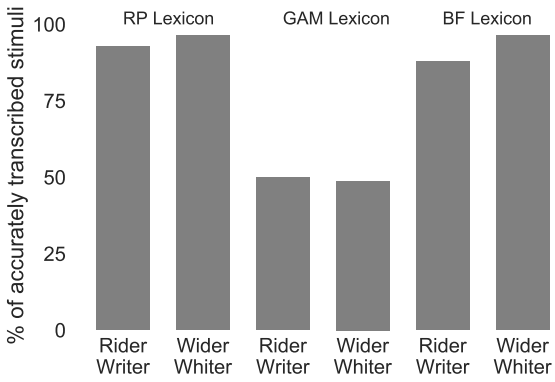

Figure 2.6: Disambiguation rate of RP voice with RP, GAM and BF lexica

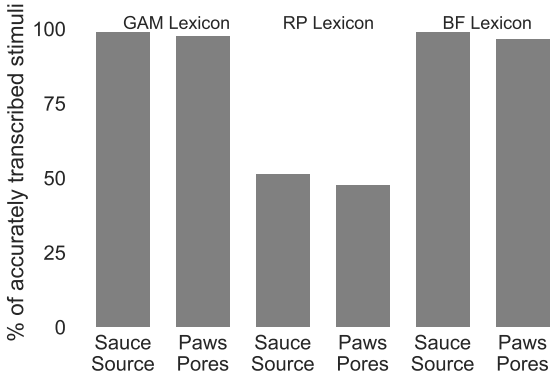

Figure 2.7: Disambiguation rate of GAM voice with GAM, RP and BF lexica

The drops in the mismatched surface-form groupings show where only one word out of the two was entered by listeners. This is because the other word in the pair sounded identical, thus listeners could not distinguish *whiter* from *wider* with the RP speech data or *source* from *sauce* in the GAM speech data. They could however with

the BF lexicon. An ability to maintain distinctions in pronunciations is important to reflect the speech of the accent in question, and once again demonstrates the advantage of using the BF lexicon over a mismatched surface-accent lexicon.

2.4 Conclusions

These results show that the BF lexicon works almost as well as a lexicon matched to the accent of the voice talent. With some further work, the BF lexicon has the potential to work equally as well in DNN-based TTS voices as a matched surface-accent lexicon. The fact that the BF lexicon can be used re-inforces the motivation for expanding the BF lexicon, not individual surface-accent lexica, with new entries. These findings have been submitted to ICASSP 2019.

Chapter 3

Expansion of the Accent-Independent Lexicon

3.1 Grapheme-to-Metaphone (G2M) modelling

Several approaches exist for creating new accent-independent word entries. The standard method for dealing with words Out-of-Vocabulary words (OOVs) in ASR and TTS is Grapheme-to-Phoneme modelling (G2P). This process involves using lexicon entries to train a statistical model that predicts a pronunciation given the text of a word. Since the accent-independent lexicon operates with metaphones, G2P becomes G2M: Grapheme-to-metaphone modelling.

This summer, Maiia Bikmetekova (MB) conducted preliminary tests of G2M with the accent-indepedent lexicon for her MSc dissertation. Her results indicated G2M performed comparably in terms of Word and Phone Error Rates (WER and PER) to surface-form G2P. I have built the RP, GAM and BF lexica-based G2P and G2M (henceforth G2P/M) models with a Bi-LSTM architecture as implemented in OpenNMT (28). For MB, this setup rendered lower error rates than Sequitur (a standard, open source G2P toolkit) and the Festival TTS system's G2P Classification and Regression Tree (CART).

3.1.1 Initial Results

Figure 3.1 confirms MB's findings that G2M performs comparably to G2P. The hyperparameters for the Bi-LSTM were: 6 bi-directional layers with 500 units each, a learning rate of 0.0001, dropout of 0.1, Luong's global attention (29), the ADAM optimiser

and mini-batches of 64. The Bi-LSTM converged after ~100,000 training steps.

The total number of incorrect phonestrings divided by the size of test set gives the Word Error Rate (WER). This indicates how many words in total were incorrectly predicted. The Phone Error Rate (PER) is calculated by summing the total levenshtein distance for every prediction in the test set by the sum of the lengths of all gold (meta-)phonestrings together. This quantifies how many (meta-)phones out of 100 were wrong. The error rates presented are low relative to G2P research from the literature. For instance, a similar system trained on the CMUDict obtained 25.8% WER in (30). This is because `Combilex` has a very high consistency across its entries thanks to 85% of its entries being derived from the same 20,000 root words and containing less names than in the commonly used CMUDict. Names tend to bear unusual pronunciations leading to worse performance in terms of error-rates.

Table 3.1: *RP, GAM and BF Lexica Grapheme-to-Phoneme and Metaphone Performance. P and W columns represent PER and WER expressed in % respectively*

	Lexicon					
	RP		**GAM**		**BF**	
	P (%)	**W (%)**	**P (%)**	**W (%)**	**P (%)**	**W (%)**
Bi-LSTM	1.1	4.6	1.2	5.0	1.1	4.9

Moreover, G2P and G2M error rates are highly sensitive to the kind of word entries included during preprocessing. In many G2P research papers, preprocessing lexica for G2P involves removing entries such as: homographs, foreign words and apostrophes. In the table above, words of fewer than 4 letters, foreign words and multiple (meta-)phonestrings for homographs were excluded. The numbers in Figure 3.2 show the effect of incrementally including these kinds of words in the RP G2P model. By '+separate root morphemes' I mean if a word such as *run* were in the training set, then words like *runner* and *running* were also in the training set. Words in the test set therefore only contain words derived from unseen root morphemes. In a practical setting, G2P will be used on words with seen and unseen morphemes in training. Since this last setup pretends they are all unseen, it portrays the worst-case performance (WER 34.3%). Conversely the top case in Figure 3.2 presents the best case scenario for predictable words (WER = 4.6%).

The table includes varying sizes of the training, val and test splits with the addition

Table 3.2: *G2P performance of Bi-LSTM with ablated RP lexicon entries*

Word Types	**Dataset sizes (relative to previous step in words)**	**PER (%)**	**WER (%)**
- 4 letter-words - homographs - foreign words - separate root morphemes	Train: 90,770 Val: 11,346 Test: 11,346	1.1	4.6
+ 4 letter-words - homographs - foreign words - separate root morphemes	Train: 94,903 (+4,133) Val: 11,862 (+516) Test: 11,863 (+517)	1.1	4.9
+ 4 letter-words + homographs - foreign words - separate root morphemes	Train: 104,713 (+9,810) Val: 13,080 (+1,218) Test: 13,085 (+1,222)	2.3	11.9
+ 4 letter-words + homographs + foreign words - separate root morphemes	Train: 106,566 (+1,853) Val: 13,327 (+247) Test: 13,394 (+309)	2.8	13.3
+ 4 letter-words + homographs + foreign words + separate root morphemes	Train: 104,201 (-2,365) Val: 15,993 (+2,666) Test: 21,986 (+8,592)	7.1	34.3

of each preprocessing step. Generally 80/10/10 was the split in percentage terms. Interestingly, the RP lexicon size increases by 10% from 118,628 to 130,878 when multiple pronunciations of a homograph are included. Importantly, What is labeled '-homographs' includes a single pronunciation of the word (whichever appears in the lexicon first).

I split the root morphemes across the training, val and test sets heuristically to get close to a 80/10/10 split but the actual percentage divide was 74/10/15. This is unequal because of the way I implemented the splitting. Correcting this is in on my to-do list.

Figure 3.2 shows the true variability of the WER in G2P dependent on the kinds of words that are included from the lexicon. The addition of words fewer than 4 letters, homographs, foreign words and separate root morphemes increase error rates. Consequently, G2P/M is not as reliable a method for automatic expansion of unseen morphemes as initially supposed.

Furthermore as Figure 1.1 shows, metaphones are not the only linguistic informtion stored in `Combilex`. However, target G2P data can be modified so that the Bi-LSTM outputs much of this additional information. For instance, (14) shows stress markers and syllabification can be jointly predicted in this way.

Including homographs and foreign words in the baseline, RP G2P performed with PER of 2.8% and WER of 13.3% (the fourth line of Figure 3.2). When lexical stress markers were predicted that increased PER to 3.2% and WER to 14.8%. With syllables they also increased to3.3% and 15.1% respectively. A further feature of Combilex is that precise graphemic alignments to each phoneme are provided. When I attempt to jointly predict these for the test set, the PER lowers to 2.3% while the WER further increases to 15.9%. I am now in the process of running these same tests with the BF lexicon, although I expect identical effects. Importantly, predicting the additional information did not bring about large increases in error rates. This shows promise for using G2P/M to predict this additional linguistic information. To fully understand the effects of predicting this extra information, I shall predict the scores on the extra information individually e.g. by separating stress markers from the (met)phones before evaluation.

As previously mentioned, `Combilex` contains few foreign names but G2P/M have higher errors for these as they have less predictable pronunciations. I would like to uncover how the current Bi-LSTM works for foreign names in particular. Assuming it would perform poorer than above, I would like to test whether the process could be improved to recognise when a name is foreign by training separate language-origin specific G2M models. The adequate model could be selected by computing a semantic representation of the OOV in a transcription, like its sentence embedding. This could be compared to other words of specific language-origins to employ separate language-origin G2M models. I would expect language-origin specific G2M to achieve higher performance due to expected consistency with the training data.

3.2 Towards Replacing Human Verification

A high-quality lexicon demands that new entries be completed without errors. This is vital for commercial TTS because listening test results are highly dependent on the correct pronunciation of words. It is still the case that human linguists verify pronunciation entries, and this makes high quality lexica expensive to maintain.

Using audio could increase the reliability of G2P/M methods to generate pronunciations for new words. Recent papers employing audio to verify G2P hypotheses include (31; 32; 33; 34), where the source of audio could be user-corrected OKGoogle queries, crowdsourced audio of names and Euronews videos. Comparison between these papers are difficult due to the use of separate datasets. Nonetheless, G2P-based methods that align audio to hypotheses appear more reliable for learning pronunciations than OOV detection and recovery methods (35; 36; 37). These aim to detect OOVs from an ASR hypothesis with a classifier and retrieve the pronunciation of them with the system's acoustic model. They tend to be unreliable as there are no obvious features that will identify an OOV in the output of an ASR system, as all candidate words are already in vocabulary. I therefore plan to implement experiments testing the robustness of G2M candidate selection via audio.

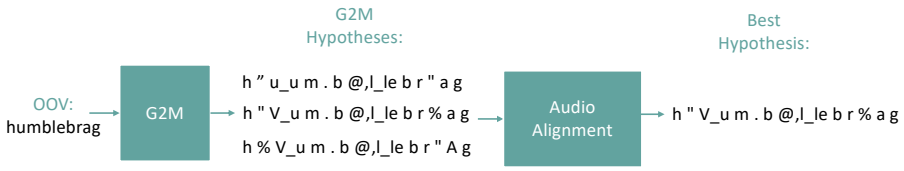

Figure 3.1: Proposed pipeline for G2M audio verification

There are a wealth of research questions regarding audio verification of G2M candidates. Firstly, does it perform better for certain kinds of words (e.g. morphologically dervied words from seen roots) and worse for others, like foreign names? Secondly, what kind of audio is required to learn pronunciations successfully - can speech in non-ideal conditions be used, e.g. with background noise and overlapping speakers? In addition, what are the benefits and challenges of using crowdsourced audio for pronunciations as opposed to regular corpora?

Analysis has shown a G2M model will not always predict the correct metaphone strings amongst its candidates. How could I verify whether this were the case? Morphological (de)composition is one method for generating reliable new entries. This takes a pre-existing word and creates new candidate words from it's root by adding

and removing bound morphemes. The resultant entries are generated from principled linguistic rules and are complete. Therefore they are reliable. If the metaphone strings of a morphologically (de)composed candidate match the G2M candidate, could this match the confidence of human-like verification? I plan to compare this method with using crowdsourced pronunciations as a substitute for human verification in year 3.

Chapter 4

Conclusion

In this report I introduced issues surrounding automatic lexicon expansion. Having demonstrated the accent-indepdendent, BF lexicon's potential by direct use in DNN-based TTS systems, I have set to work on expanding it. I established that G2M performs comparably to G2P and that extra linguistic information can be jointly predicted such as stress markers, syllabification and graphemic alignments with relatively small increases in WER. I presented methods and research questions with the aim of removing human verification of new lexicon entries. These include audio verification of G2M predictions, morphological (de)composition and crowdsourcing pronunciation.

Bibliography

[1] P. Taylor, *Text-to-Speech Synthesis*. Cambridge: Cambridge University Press, 2009.

[2] K. Richmond, R. Clark, and S. Fitt, "On Generating Combilex Pronunciations via Morphological Analysis," in *Interspeech*, 2010. [Online]. Available: https://pdfs.semanticscholar.org/2167/ee32dbdb1d5607e4d75828d76bf417dd455f.pdf

[3] R. Prabhavalkar, K. Rao, T. N. Sainath, B. Li, L. Johnson, and N. Jaitly, "A Comparison of Sequence-to-Sequence Models for Speech Recognition," 2017. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2017-233

[4] H. Soltau, H. Liao, and H. Sak, "Neural Speech Recognizer: Acoustic-to-Word LSTM Model for Large Vocabulary Speech Recognition." [Online]. Available: https://arxiv.org/pdf/1610.09975.pdf

[5] D. Amodei et al, "Deep Speech 2 : End-to-End Speech Recognition in English and Mandarin," 2016. [Online]. Available: http://proceedings.mlr.press/v48/amodei16.pdf

[6] P. Bell, M. Gales, T. Hain, J. Kilgour, P. Lanchantin, X. Liu, A. McParland, S. Renals, O. Saz, M. Wester, and P. Woodland, "The MGB challenge: Evaluating multi-genre broadcast media recognition," in *Proc. ASRU*, 2015. [Online]. Available: https://ieeexplore.ieee.org/document/7404863

[7] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an ASR Corpus Based on Public Domain Audio Books." [Online]. Available: https://www.danielpovey.com/files/2015_icassp_librispeech.pdf

[8] Y. Wang, R. J. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. A.

Saurous, “Tacotron: Towards End-To-End Speech Synthesis,” 2017. [Online]. Available: https://google.github.io/tacotron

[9] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. J. Skerry-Ryan, R. A. Saurous, Y. Agiomyrgiannakis, and Y. Wu, “Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4779–4783. [Online]. Available: https://google.github.io/tacotron/publications/tacotron2.

[10] W. Ping, K. Peng, A. Gibiansky, S. Sercan, S. Arık, A. Kannan, S. Narang, B. Research, J. Raiman, Openai, and J. Miller, “Deep Voice 3: Scaling Text-to-Speech with Convolutional Sequence Learning,” in *ICLR*, 2018. [Online]. Available: https://arxiv.org/pdf/1710.07654.pdf

[11] Y. Taigman, L. Wolf, A. Polyak, and E. Nachmani, “VoiceLoop: Voice Fitting and Synthesis via a Phonological Loop,” in *ICLR*, 2018. [Online]. Available: https://arxiv.org/abs/1707.06588

[12] CMU, “The Carnegie Mellon Pronouncing Dictionary,” 2018. [Online]. Available: https://github.com/cmusphinx/cmudict

[13] K. Gorman, G. Mazovetskiy, and V. Nikolaev, “Improving homograph disambiguation with supervised machine learning,” in *LREC*, 2018. [Online]. Available: http://wellformedness.com/papers/gorman-etal-2018.pdf

[14] D. Van Esch, M. Chua, and K. Rao, “Predicting pronunciations with syllabification and stress with recurrent neural networks,” in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2016. [Online]. Available: https://www.isca-speech.org/archive/Interspeech_2016/pdfs/1419.PDF

[15] W. Labov, *Sociolinguistic patterns / William Labov*. University of Pennsylvania Press Philadelphia, 1973.

[16] G. Howard, *Communication Accommodation Theory: Negotiating Personal Relationships and Social Identities across Contexts*. Cambridge University Press, 2016.

[17] CSTR, "Unisyn Lexicon Release, version 1.3," 2018. [Online]. Available: http://www.cstr.ed.ac.uk/projects/unisyn/

[18] J. C. Wells, *Accents of English*, ser. Accents of English. Cambridge University Press, 1982. [Online]. Available: https://books.google.co.uk/books?id=yIunVTcLg8MC

[19] S. Fitt and S. Isard, "Synthesis of regional English using a keyword lexicon," in *EUROSPEECH 99*, 1999. [Online]. Available: http://www.cstr.ed.ac.uk/projects/unisyn.html

[20] J. Chevelu, D. Lolive, S. Le Maguer, and D. Guennec, "How to Compare TTS Systems: A New Subjective Evaluation Methodology Focused on Differences," in *Interspeech*, 2015. [Online]. Available: https://hal.inria.fr/hal-01199082

[21] H. Ze, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, may 2013, pp. 7962–7966. [Online]. Available: http://ieeexplore.ieee.org/document/6639215/

[22] M. Cooke, C. Mayo, and C. Valentini-Botinhao, "Intelligibility-enhancing speech modifications: the Hurricane Challenge," in *Interspeech*, 2013. [Online]. Available: http://www.cstr.ed.ac.uk/downloads/publications/2013/Cooke_IS13.pdf

[23] Z. Wu, O. Watts, and S. King, "Merlin: An Open Source Neural Network Speech Synthesis System," in *9th ISCA Speech Synthesis Workshop*, 2016. [Online]. Available: http://dx.doi.org/10.21437/SSW.2016-33

[24] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: A Vocoder-Based High-Quality Speech Synthesis System for Real-Time Applications," *IEICE Transactions on Information and Systems*, vol. E99.D, no. 7, pp. 1877–1884, 2016.

[25] O. Watts, G. E. Henter, T. Merritt, Z. Wu, and S. King, "From HMMs to DNNs: Where do the improvements come from?" in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016. [Online]. Available: http://www.cstr.ed.ac.uk/downloads/publications/2016/watts2016hmms.pdf

[26] M. Wester, C. Valentini-Botinhao, and G. E. Henter, "Are We Using Enough Listeners? No! — An Empirically-Supported Critique of Interspeech 2014 TTS Evaluations," in *Interspeech*, 2015, pp. 3476–3480. [Online]. Available: http://www.cstr.ed.ac.uk/downloads/publications/2015/wester:listeners:IS2015.pdf

[27] S. Fitt, "The Treatment of Vowels Preceding 'r' in a Keyword Lexicon of English," in *Proceedings of ICPhS 99*, 1999. [Online]. Available: https://www.era.lib.ed.ac.uk/handle/1842/1190

[28] G. Klein, Y. Kim, Y. Deng, J. Senellart, and A. M. Rush, "OpenNMT: Open-Source Toolkit for Neural Machine Translation," 2017. [Online]. Available: https://arxiv.org/pdf/1701.02810.pdf

[29] M.-T. Luong, H. Pham, and C. D. Manning, "Effective Approaches to Attention-based Neural Machine Translation," Tech. Rep., 2015. [Online]. Available: http://aclweb.org/anthology/D15-1166

[30] K. Rao, F. Peng, H. Sak, and F. Beaufays, "Grapheme-to-phoneme conversion using Long Short-Term Memory recurrent neural networks," 2015. [Online]. Available: https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/43264.pdf

[31] A. Bruguier, F. Peng, and F. Beaufays, "Learning Personalized Pronunciations for Contact Name Recognition." [Online]. Available: https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/45415.pdf

[32] A. Bruguier, D. Gnanapragasam, L. Johnson, K. Rao, and F. Beaufays, "Pronunciation learning with RNN-transducers," 2017. [Online]. Available: https://www.bruguier.com/pub/general_prons.pdf

[33] A. T. Rutherford and F. Peng, "Pronunciation Learning for Named-Entities through Crowd-Sourcing." [Online]. Available: https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/43086.pdf

[34] I. Sheikh, D. Fohr, I. Illina, and G. Linares, "Modelling Semantic Context of OOV Words in Large Vocabulary Continuous Speech Recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2017. [Online]. Available: https://hal.inria.fr/hal-01461617/document

[35] I. Szöke, M. Fapšo, L. Burget, and J. Cernocký, "Hybrid word-subword decoding for spoken term detection," Tech. Rep. [Online]. Available: http://noel.feld.cvut.cz/gacr0811/publ/SZO08b.pdf

[36] L. Qin, A. W. Black, F. Metze, and M. Dredze, "Learning Out-of-Vocabulary Words in Automatic Speech Recognition." [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.417.972

[37] E. Egorova and L. Burget, "Out-of-Vocabulary Word Recovery Using FST-based Subword Unit Clustering in a Hybrid ASR System." [Online]. Available: http://www.fit.vutbr.cz/research/groups/speech/publi/2018/egorova_icassp2018_0005919.pdf