Generating multiple-accent pronunciations for TTS using joint sequence model interpolation

BalaKrishna Kolluru, Vincent Wan, Javier Latorre, Kayoko Yanagisawa, Mark J. F. Gales

Toshiba Research Europe Ltd, Cambridge, UK

{balakrishna.kolluru,vincent.wan,javier.latorre,kayoko.yanagisawa,mjfg}@crl.toshiba.co.uk

10.21437/Interspeech.2014-318

Abstract

Standard grapheme-to-phoneme (G2P) systems are trained using a homogeneous lexicon, for example one associated with a particular accent. In practice, a synthesis system may be required to handle multiple accents. Furthermore, a speaker rarely has a pure accent; accents vary continuously within and between regions of a country. Generating phonetic sequences for each accent is possible, but combining them to yield a single synthesis pronunciation is highly challenging. To address this problem, this paper considers a space of accents. The bases for these spaces are defined by statistical G2P models in the form of graphone models. A linear combination of these models define the accent space. By selecting a point in this continuous space, it is possible to specify the accent for an individual speaker. The performance of this approach is evaluated using an accent space defined by American, Scottish and British English. By moving around the accent space, it is shown that it is possible to synthesize speech from all these accents as well as a range of intermediate points.

Index Terms: phonetic sequence generation, accent space, interpolation

1. Introduction

Accent is the peculiar manner of pronunciation of a particular individual, location, or community. An accent may identify where speakers reside(d) (regional accent), their socio-economic status or ethnicity (social accent), or whether they have a different first language (foreign accent). As such, it is a crucial component of a speaker's voice identity. Not surprisingly thus, the ability to modify the accent is one of the most demanded features of customizable text-to-speech (TTS) systems, especially by users of augmented and alternative communicative (AAC) devices that need to use the TTS voice as their own to communicate with other people [1]. Accents may differ in various ways: phonemically (set of phonemes used), phonetically (acoustic realisation of the same phoneme), phonotactically (sequences of phones which are permitted), prosodically, or in terms of voice quality [2]. From a phonological point of view, accent differences may be reflected in their letter-to-sound mapping. These differences can consist of substitution of some phones, for example 'bath' /b ɑː θ/ in British English (Received Pronunciation, RP) versus /b æ θ/ in General American English (GA); or in insertions/deletions of some phones, for example in 'herb' /h ɜː b/ (RP) vs /ɝː b/ (GA).

In order to achieve high quality synthesis from TTS systems, homogeneous data has typically been required. This imposes constraints on factors such as speaker variety, speaking expression, speaker accent and recording condition. Advances in acoustic modelling have allowed the constraints to be relaxed allowing TTS models to be trained on diverse data such as datasets containing multiple speakers, expressions or recording conditions [3] and the synthesis quality of one speaker may be improved by training on many other speakers. By analogy training TTS models on more than one accent may also help to improve overall quality.

A change in the accent changes both the prosody and the phonetics. Acoustic variations, either in the spectrum or the prosody, if consistent, can be modelled by means of a set of transforms that change the statistical distribution of the acoustic features. This allows phone substitution in which one phone changes to another [4]. Insertions and deletions of phones, however, imply adding/removing states to the HMM sequence. These cannot be modelled by manipulating the statistical distributions of the phones and thus requires a change in the letter-to-sound module of the system.

The letter-to-sound conversion is one of the first steps of a TTS system. It converts a sequence of input words into a sequence of phones. In most systems, this conversion is done by a combination of a look-up table (a lexicon) and grapheme-to-phoneme (G2P) module that mainly deals with the out-of-vocabulary words. Within this framework, it is possible to switch between accents by switching between their associated lexica and G2P models. However, accents are not discrete entities. Although for convenience we may refer to the British accent or the American accent, for example, accents vary almost continuously across regions or social strata [5], with every individual probably configuring their own accent specific as a mixture of different components. The same speaker's accent may also vary over time [6]. Quantising that continuum with a set of lexica is difficult and costly. An alternative is to "parametrise" the accent space with some accent-axis that, when interpolated, can form all the other variations. However, if the accent-space were to be based on the phonetic sequence, it is not clear how such an interpolation could be carried out, because phone sequences are sequences of symbols.

Another way is to consider that accent space not as the space of phone sequence but as the space of the probabilities over the phone sequence. In that respect, Li *et. al.* [7] proposed a G2P model adaptation approach that combines acoustic models with graphonemes using the maximum likelihood criterion. Waxmonsky and Reddy [8] approach the phoneme generation for proper names by interpolating language-dependent and language-independent G2P models. Li *et. al.* [9] extend n-gram graphone model pronunciation generation to use mixture models. They report that interpolation is useful when pronunciation data is for a specific variant (or set of variants such as a dialect) of a language. Their approach uses G2P models to capture the variation in pronunciation between different aspects of the language.

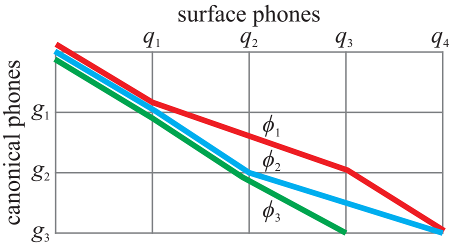

Figure 1: Lattice with the co-segmentation of grapheme and phoneme

In this paper, another approach that forms a continuous space of accents is described. The approach still uses a lexicon plus a G2P system as the first step but the phone sequence at this stage is a generic or canonical version (c.f. the base form of Combilex [10]). The canonical sequence is then converted to an accented sequence by means of a set of joint sequence models (JSMs) [11] which are interpolated at the probability distribution level. This second step is a phone-to-phone (P2P) mapping. Defining the JSMs as the bases of a space, the interpolation weights may be interpreted as points within an accent space. By selecting a point within this continuous space it is possible to specify accents for individual speakers.

This paper shows a proof of concept that choosing different points in accent space does indeed vary the accented phone sequence generated. The paper is organised as follows: Section 2 contains the description of the model used for the P2P accent interpolation; Section 3 describes the different experiments carried out to evaluate the performance; Section 4 discusses subjective evaluations; and Section 5 concludes.

2. Accent modelling for TTS

The goal of a TTS is to find the sequence of acoustic features $\boldsymbol{c}$ most probable for the sequence of input words $\boldsymbol{w}$ and a given model $\boldsymbol{\Theta}$. If the phone sequence $\boldsymbol{q}$ is considered a hidden variable, the synthesis process can be expressed as finding

$$\begin{aligned}\hat{\boldsymbol{c}} &= \arg\max_{\boldsymbol{c}} P(\boldsymbol{c}|\boldsymbol{w},\boldsymbol{\Theta}) && (1)\\ &= \arg\max_{\boldsymbol{c}} \sum_{\boldsymbol{q}} P(\boldsymbol{c}|\boldsymbol{q},\boldsymbol{w},\boldsymbol{\Theta})P(\boldsymbol{q}|\boldsymbol{w},\boldsymbol{\Theta}) && (2)\\ &\approx \arg\max_{\boldsymbol{c},\boldsymbol{q}} P(\boldsymbol{c}|\boldsymbol{q},\boldsymbol{\Theta}_{\text{ac}})P(\boldsymbol{q}|\boldsymbol{w},\boldsymbol{\Theta}_{\text{g2p}}) && (3)\end{aligned}$$

where $\boldsymbol{\Theta}_{\text{ac}}$ and $\boldsymbol{\Theta}_{\text{g2p}}$ refer to the acoustic model and G2P model respectively. If the G2P is deterministic then there is one single $\boldsymbol{q}$ for which $P(\boldsymbol{q}|\boldsymbol{w}) \neq 0$. This is the case in most TTS systems which obtain the phonetic sequence mostly from a lexicon. In those systems, the way to model different accents is by having multiple lexica. However, accents tend to change gradually from one area to another. Therefore, it can be difficult and costly to define a set of lexica to cover all possible accents over such a continuous space. A possible way to address this, would be to interpolate the phonetic sequence for each accent. The problem is that phonetic sequences are formed by discrete symbols, and it is not obvious how to interpolate between them. Another approach is to interpolate not the $\boldsymbol{q}$ directly but their probability distributions.

2.1. Joint sequence model

A joint sequence model (JSM) is a G2P model that defines $P(\boldsymbol{q}|\boldsymbol{w})$ as the marginalization over all the possible joint segmentations $\boldsymbol{\phi}$ of the phone sequence $\boldsymbol{q}$ and the grapheme sequence $\boldsymbol{g}$ associated with $\boldsymbol{w}$. A joint segmentation $\boldsymbol{\phi}$ specifies

- phone realisation: $\boldsymbol{q} = \{\phi_1^{(\boldsymbol{q})}, \ldots, \phi_{|\boldsymbol{\phi}|}^{(\boldsymbol{q})}\}$
- canonical phones: $\boldsymbol{g} = \{\phi_1^{(\boldsymbol{g})}, \ldots, \phi_{|\boldsymbol{\phi}|}^{(\boldsymbol{g})}\}$

where $|\ \boldsymbol{\phi}\ |$ is the number of segments in $\boldsymbol{\phi}$. Figure 1 gives a graphical description of this. The size of each of the units need not be the same provided the combination of grapheme sub-sequences and phone sub-sequences yield $\boldsymbol{g}$ and $\boldsymbol{q}$ respectively. The set of all possible combinations is denoted by $\boldsymbol{\Phi}(\boldsymbol{g},\boldsymbol{q})$. If the mapping between $\boldsymbol{w}$ and $\boldsymbol{g}$ is deterministic, then $P(\boldsymbol{q}|\boldsymbol{w},\boldsymbol{\Theta}_{\text{jsm}}) = P(\boldsymbol{q}|\boldsymbol{g},\boldsymbol{\Theta}_{\text{jsm}})$, where $\boldsymbol{\Theta}_{\text{jsm}}$ is the JSM model. Therefore, it can be written

$$\begin{aligned}P(\boldsymbol{q} \mid \boldsymbol{g},\boldsymbol{\Theta}_{\text{jsm}}) &= \frac{P(\boldsymbol{q},\boldsymbol{g} \mid \boldsymbol{\Theta}_{\text{jsm}})}{P(\boldsymbol{g} \mid \boldsymbol{\Theta}_{\text{jsm}})} && (4)\\ &\propto \sum_{\boldsymbol{\phi}\in\boldsymbol{\Phi}(\boldsymbol{g},\boldsymbol{q})} P(\boldsymbol{\phi} \mid \boldsymbol{\Theta}_{\text{jsm}}) \cdot P(\boldsymbol{q},\boldsymbol{g} \mid \boldsymbol{\phi},\boldsymbol{\Theta}_{\text{jsm}}) && (5)\end{aligned}$$

with

$$P(\boldsymbol{q},\boldsymbol{g}|\boldsymbol{\phi},\boldsymbol{\Theta}_{\text{jsm}}) = \prod_{j=1}^{|\boldsymbol{\phi}|} P(\phi_j^{(\boldsymbol{q})},\phi_j^{(\boldsymbol{g})} \mid \boldsymbol{\Theta}_{\text{jsm}}) \qquad (6)$$

2.2. Accent interpolation

In the case of multiple accents, each one can have its associated JSM model. It has been shown that JSM models can be interpolated to handle pronunciation variations in Arabic [9]. A similar approach can be used to handle pronunciation variations across accents as shown in Figure 2. In this case, the phonetic sequence also depends on the specific accent. Assuming that the accent can be defined by a vector of interpolation weights $\boldsymbol{\lambda}$, then for each accent

$$P(\boldsymbol{q}|\boldsymbol{\lambda},\boldsymbol{g},\boldsymbol{\Theta}_{\text{jsm}}) \propto \sum_{\boldsymbol{\phi}\in\boldsymbol{\Phi}(\boldsymbol{g},\boldsymbol{q})} P(\boldsymbol{\phi} \mid \boldsymbol{\lambda},\boldsymbol{\Theta}_{\text{jsm}})P(\boldsymbol{q},\boldsymbol{g} \mid \boldsymbol{\lambda},\boldsymbol{\phi},\boldsymbol{\Theta}_{\text{jsm}}) \qquad (7)$$

where

$$P(\boldsymbol{q},\boldsymbol{g}|\boldsymbol{\lambda},\boldsymbol{\phi},\boldsymbol{\Theta}_{\text{jsm}}) = \prod_{j=1}^{|\boldsymbol{\phi}|}\sum_{a} \lambda_a P(\phi_j^{(\boldsymbol{q})},\phi_j^{(\boldsymbol{g})} \mid \boldsymbol{\Theta}_{\text{jsm}}^{(a)}) \quad (8)$$

and a denotes the accent. For simplicity, $P(\boldsymbol{\phi} \mid \boldsymbol{\lambda},\boldsymbol{\Theta}_{\text{jsm}})$ is assumed to follow a uniform distribution w.r.t. $\boldsymbol{\lambda}$.

For languages with a clean grapheme-to-phoneme structure, $\boldsymbol{g}$ can be directly the grapheme sequence. However, for complex languages, such as English, this might be problematic, since G2P cannot produce as good results as a lexicon. Therefore, instead of $\boldsymbol{g}$ being the grapheme sequence, it can also be the canonical pronunciation obtained from a lexicon.

At synthesis time, solving (3) over all the possible $\boldsymbol{q}$ can be computationally very expensive. A simpler solution is to obtain from (7) the N best $\boldsymbol{q}$ and out of them select the one that maximises (3). Naturally, an even further simplification would be to use $N = 1$.

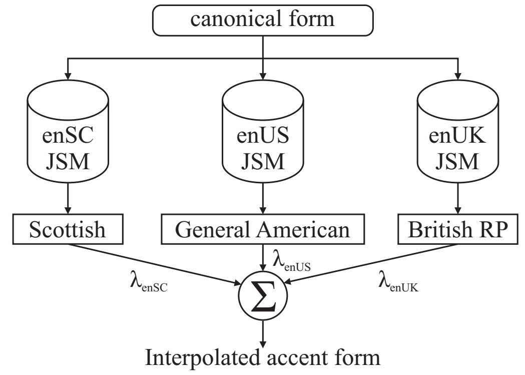

Figure 2: Interpolating different accents to generate a phonetic sequence

2.3. Estimation of the accent interpolation weights

If some phonetic sequence $\boldsymbol{q}$ for the intended accent is known, the optimum $\boldsymbol{\lambda}$ can be obtained as the one that maximizes (7). Assuming, for simplicity, that $P(\boldsymbol{\phi} \mid \boldsymbol{\lambda}, \boldsymbol{\Theta}_{\text{jsm}})$ to follows a uniform distribution w.r.t. $\boldsymbol{\lambda}$, this can be achieved by applying Jensen's inequality to (7) to get the auxiliary function

$$F(\boldsymbol{\lambda}) = \sum_{\boldsymbol{\phi}\in\boldsymbol{\Phi}(\boldsymbol{g},\boldsymbol{q})} \log P(\boldsymbol{q},\boldsymbol{g} \mid \boldsymbol{\lambda}, \boldsymbol{\phi}, \boldsymbol{\Theta}_{\text{jsm}}) \tag{9}$$

By forcing the constraints $\sum_a \lambda_a = 1$ and $\lambda_a \geq 0 \forall a$, $\boldsymbol{\lambda}$ can be considered as the mixture weights in a Gaussian mixture model [12]. Thus, (9) is maximized by

$$\hat{\lambda}_a = \frac{1}{N} \sum_{\boldsymbol{\phi}\in\boldsymbol{\Phi}(\boldsymbol{g},\boldsymbol{q})} \sum_{j=1}^{|\boldsymbol{\phi}|} \frac{\lambda_a P(\phi_j^{(\boldsymbol{q})}, \phi_j^{(\boldsymbol{g})} \mid \boldsymbol{\Theta}_{\text{jsm}}^{(a)})}{\sum_a \lambda_a P(\phi_j^{(\boldsymbol{q})}, \phi_j^{(\boldsymbol{g})} \mid \boldsymbol{\Theta}_{\text{jsm}}^{(a)})} \tag{10}$$

with

$$N = \sum_{\boldsymbol{\phi}\in\boldsymbol{\Phi}(\boldsymbol{g},\boldsymbol{q})} \sum_{j=1}^{|\boldsymbol{\phi}|} 1 \tag{11}$$

which can be solved iteratively. If instead of the phonetic sequence, some acoustic data $\boldsymbol{O}$ from the target speaker is given, the optimum $\boldsymbol{\lambda}$ can be obtained by considering $\boldsymbol{q}$ as a hidden variable and using an EM algorithm. Using a combined model $\boldsymbol{\Theta} = \{\boldsymbol{\Theta}_{\text{jsm}}, \boldsymbol{\Theta}_{\text{ac}}\}$ in which $\boldsymbol{\Theta}_{\text{ac}}$ represents a multi-accent acoustic model, the auxiliary function can be written as

$$Q(\boldsymbol{\lambda}, \hat{\boldsymbol{\lambda}}) = \sum_{\boldsymbol{q}} P(\boldsymbol{q}|\boldsymbol{O}, \boldsymbol{\lambda}, \boldsymbol{g}, \boldsymbol{\Theta}) \log(P(\boldsymbol{q}, \boldsymbol{O}|\hat{\boldsymbol{\lambda}}, \boldsymbol{g}, \boldsymbol{\Theta})) \tag{12}$$

Assuming independence between $\boldsymbol{\Theta}_{\text{ac}}$ and $\boldsymbol{\Theta}_{\text{jsm}}$ and using the Viterbi approximation

$$Q(\boldsymbol{\lambda}, \hat{\boldsymbol{\lambda}}) \approx C(\log P(\hat{\boldsymbol{q}}|\boldsymbol{g}, \boldsymbol{\lambda}, \boldsymbol{\Theta}_{\text{jsm}}) + \log P(\hat{\boldsymbol{q}}|\boldsymbol{O}, \boldsymbol{\Theta}_{\text{ac}})) \tag{13}$$

where

$$\begin{aligned} \hat{\boldsymbol{q}} &= \arg\max_{\boldsymbol{q}} P(\boldsymbol{q}|\boldsymbol{g}, \boldsymbol{\lambda}, \boldsymbol{\Theta}_{\text{jsm}}) P(\boldsymbol{q}|\boldsymbol{O}, \boldsymbol{\Theta}_{\text{ac}}) && (14) \\ C &= P(\hat{\boldsymbol{q}}|\boldsymbol{O}, \boldsymbol{\lambda}, \boldsymbol{g}, \boldsymbol{\Theta}_{\text{jsm}}, \boldsymbol{\Theta}_{\text{ac}}) && (15) \end{aligned}$$

During the E-step the optimum $\hat{\boldsymbol{q}}$ given current $\boldsymbol{\lambda}$ is obtained from (14). Then, the new $\hat{\boldsymbol{\lambda}}$ are obtained by inserting $\hat{\boldsymbol{q}}$ in (10). Interestingly, the role of the JSM model in (14) is basically the same as that of a language model in automatic speech recognition.

2.4. Model initialization

Ideally, the JSM models should be trained over the whole phonetic sequence associated with an utterance. However, for initialization they can be trained on a word-level rather than on an utterance-level. Initialization is achieved by training one JSM for each accent to map from the canonical pronunciation of each word to the corresponding surface pronunciation using the accent specific lexica.

3. Experiments

Joint sequence models have previously operated on grapheme-to-phoneme (G2P) mappings [11]. As discussed in Section 2, a phone-to-phone (P2P) mapping from a canonical pronunciation to accented (surface) pronunciation is modelled by the JSMs.

3.1. Data

Three English lexica, derived from Combilex [10], are used for these experiments: General American (enUS), Scottish (enSC) and British RP (enUK). Each lexicon has approximately 129K headword entries, with multiple pronunciations for some words. They were converted to the Toshiba Universal Phoneset (TUPS) to ensure that a single phone set was used throughout. TUPS is a unified phone set that can be used to define the inventory for multiple languages, whereby phones which are described by the same International Phonetic Alphabet (IPA) symbol share the same TUPS symbol. It is expected that any differences in the acoustic realisation of the phones represented by the same symbol would be captured by the acoustic model.

The ideal canonical pronunciations for these experiments would have been derived from the base form of Combilex. Unfortunately, these were not available at the time of writing. Therefore, the canonical form was based on the enUS lexicon. To create the canonical lexicon, multiple pronunciations were stripped away, leaving only the most common pronunciation for each word.

3.2. P2P performance evaluation

To study the impact of a phone-to-phone model, 10-fold cross-validation is performed on the three individual JSMs built for enUS, enUK and enSC. Each lexicon was randomly split into 10 equally sized subsets with the aim of ensuring that each individual subset was phonetically balanced. Within each fold, a JSM was trained from 9 of the subsets and tested on the remaining held out set.

Table 1 shows the average phone-error rate performance across all 10-folds for third and forth order multigrams on the three accents. From pilot experiments, the phone-error rate is lowest for four-gram models. In contrast, for a G2P task, [11] reported phone-error rate performance was lowest for around seven or eight grams. This is expected since the canonical pronunciation can be expected to be much closer to the surface pronunciation than a grapheme.

3.3. Interpolation of JSMs for accent

The effect of interpolating JSMs on the generated accented pronunciations may be observed by calculating the phone-error rates with respect to the two accents at different interpolation values.

Figure 3 shows the change in phone-accuracy with respect to enUS and enUK ground-truths when interpolating between the two accents. It can be seen that there is a smooth transi-

accent	3-gram	4-gram
enUK	0.484%	0.360%
enSC	0.427%	0.331%
enUS	0.223%	0.181%

Table 1: Average phone-error rate of 3- and 4-gram JSMs per accent.

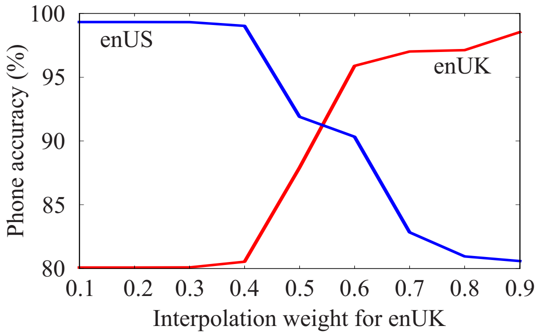

Figure 3: Phone accuracy with respect to the enUS and enUK ground-truth pronunciations as a function of interpolation weights between enUS and enUK accents. The weight for enSC is zero.

tion from one accent to the other as the interpolation weight changes. At one end, when the enUS model has the majority of the weight then more enUS phonetic sequences are predicted. As the interpolation weight moves towards enUK from enUS, a steadily increasing number of words are reassigned enUK pronunciations instead. Figure 4 illustrates two example words that change pronunciation at different points in the interpolation.

Similar results may be observed for enUS versus enSC and enUK versus enSC. By combining a large number of accents in this way it is hoped that an accent space can be formed.

4. Discussion

There are additional benefits of mapping between phone sequences compared to that of standard G2P. For example, it is possible to incorporate post-lexical effects such as liaison and other coarticulatory phenomena into the JSMs since the models can act on the complete set of phones of the entire utterance. This differs from the standard lexicon plus G2P approach which works on a word by word basis so post-lexical effects can only be added as a post-process.

The design of listening tests to evaluate mixed-accent TTS by must be considered carefully so that the subtle variations that occur between accents are correctly assessed. In unreferenced listening tests, subjects must be familiar with the accents being evaluated. Finding enough listeners to participate will become increasingly difficult as the number of accents increases. For example, the British RP accent is actually spoken by a very small proportion of people within the UK. Although most non-linguistically trained listeners will recognise the gross difference between the accents modelled in this paper, there may be more subtle variations that they may miss. Therefore, it is necessary to train the listeners to become properly familiar with the evaluated accents. In any case it is unlikely that listeners will know how a mixed-accent that consists of 30% Scottish and 70% American should sound. More generally, in unreferenced evaluations, listeners may have different mental references for each accent [13].

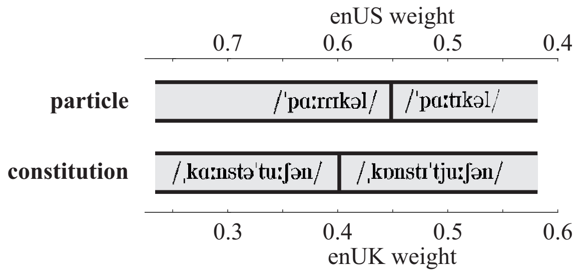

Figure 4: Words change pronunciations at different interpolation weights. The pronunciation is notated in IPA. The weight of enSC is zero.

Referenced tests are therefore preferred. One such test might ask subjects to say whether a sample of speech is closer to one accent or the other. However this requires a corpus in which each test sentence is spoken in each of the evaluated accents. The test is also more suited to evaluating broad accent classes than the subtle accents that can be obtained via an interpolation. Nevertheless it might be possible to analyse the interpolated accents in a similar way that [14] evaluated speaker-interpolation.

Alternatively, a similarity test may be performed if speech from a mixed-accent speaker is given as a reference: subjects can decide how close the speech generated using the proposed accent interpolation method sounds to the original mixed-accent speaker. For this purpose, future directions of this work include updating λ using the acoustics of the target speaker, and based on that, retraining the set of JSM models in a global fashion using speech data from real mixed-accented speakers.

5. Conclusion

This paper described a method of generating mixed accent lexicons via interpolation of JSMs for TTS applications that require high levels of personalisation. Instead of converting directly from grapheme to accented phoneme sequences, an additional step is introduced: the graphemes are first converted to a set of canonical phoneme sequences by standard approaches; they are in turn converted to accented phoneme sequences by JSMs.

The interpolation of the JSM based phone-to-phone models, each trained on different accents, successfully produces mixed-accents. The results in section 3.3 show that different words "flip" pronunciation at different interpolation weights. This interpolation weight, λ, marks a speaker in the accent space. Just as an accent may vary gradually across a geographical region, the interpolation weight causes a similar impact. The experiments in this work are very much a proof of concept and interpolate on a broad scale. However, the actual values of interpolation weights should be obtained from equation (3) taking into account the actual acoustic observations for those utterances in that accent.

Future work will include perceptual evaluations and adapting the JSMs to real speakers with mixed accents.

6. References

[1] J. Yamagishi, C. Veaux, S. King, and S. Renals, “Speech synthesis technologies for individuals with vocal disabilities: Voice banking and reconstruction,” *Acoustical Science and Technology*, vol. 33, no. 1, pp. 1–5, 2012.

[2] J. Wells, *Accents of English*. Cambridge University Press, 1982.

[3] V. Wan, J. Latorre, K. Yanagisawa, N. Braunschweiler, L. Chen, M. J. F. Gales, and M. Akamine, “Building HMM-TTS voices on diverse data,” *IEEE journal on Selected Topics in Signal Processing*, April 2014.

[4] M. Nicolao, J. Latorre, and R. Moore, “C2h: A computational model of h&h-based phonetic contrast in synthetic speech,” in *Proc. Interspeech*, 2012.

[5] W. Heeringa and J. Nerbonne, “Dialect areas and dialect continua,” *Language Variation and Change*, vol. 13, pp. 375–400, 10 2001.

[6] J. Harrington, S. A. Palethorpe, and C. Watson, “Monophthongal vowel changes in received pronunciation: an acoustic analysis of the Queen’s Christmas broadcasts,” *Journal of the International Phonetic Association*, vol. 30, pp. 63–78, 2000.

[7] X. Li, A. Gunawardana, and A. Acero, “Adapting grapheme-to-phoneme conversion for name recognition,” in *Proc. ASRU*, 2007.

[8] S. Waxmonsky and S. Reddy, “G2P conversion of proper names using word origin information,” in *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, ser. NAACL HLT ’12, 2012, pp. 367–371.

[9] T. Li, P. C. Woodland, F. Diehl, and M. J. F. Gales, “Graphone model interpolation and arabic pronunciation generation,” in *Proc. Interspeech*, 2011.

[10] K. Richmond, R. Clark, and S. Fitt, “On generating combilex pronunciations via morphological analysis,” in *Proc. Interspeech*, 2010.

[11] M. Bisani and H. Ney, “Joint-sequence models for grapheme-to-phoneme conversion,” *Speech Communication*, vol. 50, no. 5, pp. 434–451, 2008.

[12] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.

[13] J. Latorre, K. Yanagisawa, V. Wan, B. Kolluru, and M. J. F. Gales, “Speech intonation for TTS: Study on evaluation methodology,” in *Interspeech*, 2014.

[14] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, “Speaker interpolation in HMM-based speech synthesis system,” in *Proc. Eurospeech*, 1997, pp. 2523–2526.