TESTING THE CONSISTENCY ASSUMPTION: PRONUNCIATION VARIANT FORCED ALIGNMENT IN READ AND SPONTANEOUS SPEECH SYNTHESIS

Rasmus Dall[1], *Sandrine Brognaux*[2 3], *Korin Richmond*[1], *Cassia Valentini-Botinhao*[1], *Gustav Eje Henter*[1], *Julia Hirschberg*[4], *Junichi Yamagishi*[1], *Simon King*[1]

[1]The Centre for Speech Technology Research, University of Edinburgh, UK
[2]Cental, ICTEAM, Universite Catholique de Louvain, Belgium
[3]TCTS Lab, University of Mons, Belgium [4]Columbia University, USA

ABSTRACT

Forced alignment for speech synthesis traditionally aligns a phoneme sequence predetermined by the front-end text processing system. This sequence is not altered during alignment, i.e., it is forced, despite possibly being faulty. The consistency assumption is the assumption that these mistakes do not degrade models, as long as the mistakes are consistent across training and synthesis. We present evidence that in the alignment of both standard read prompts and spontaneous speech this phoneme sequence is often wrong, and that this is likely to have a negative impact on acoustic models. A lattice-based forced alignment system allowing for pronunciation variation is implemented, resulting in improved phoneme identity accuracy for both types of speech. A perceptual evaluation of HMM-based voices showed that spontaneous models trained on this improved alignment also improved standard synthesis, despite breaking the consistency assumption.

Index Terms: speech synthesis, TTS, forced alignment, HMM

1. INTRODUCTION

An essential preprocessing step of the speech data for text-to-speech synthesis (TTS) is alignment. That is the segmentation of raw speech waveforms into the phonemes of the utterance for the purpose of later use as units in unit selection synthesis, or to train models for statistical parametric speech synthesis (SPSS). Alignment is normally performed using an automatic method called forced alignment, as manually aligning speech is both expensive and error-prone [1].

In English speech synthesis, the standard forced alignment procedure lets the TTS front-end produce a phonemisation which the algorithm is then forced, hence the name, to find boundaries of in the acoustics. This phonemisation may be incorrect and the phonemes wrong, e.g. when reductions or deletions occur. As such the phonemes may not exist in the utterance, although due to the forced nature of the method these will still be "found". This is not a major issue in unit selection as the join cost will discourage any badly aligned units from being selected. In SPSS a join cost is not used, but it is usually assumed that the phonemisation is sufficiently close to correct that this is not an issue. As a consequence, it is assumed that any bad units are either averaged out as "noise", or that by being consistent across training and synthesis mismatches will not affect output speech. This is what we here call the consistency assumption between training and synthesis, namely that making the same mistakes consistently may "accidentally" have positive effects, such as appropriate phoneme reductions [2]. The assumption may be an extrapolation from automatic speech recognition, in which manual alignments do not improve word error rates over forced alignment [3], but to the best of our knowledge this has not been directly tested in synthesis. [4] provides evidence that the assumption holds but does not discuss the finding as their focus was on retaining the consistency. Furthermore, for SPSS based on spontaneous conversational speech data, it is worth noting that there are significant differences between the appearance of conversational phenomena in standard read prompts and spontaneously produced speech [5, 6, 7, 8, 9, 10]. Thus forced alignment could produce more, and more serious, errors than when aligning read speech, which may impact synthesis quality. Earlier work on spontaneous TTS admitted problems with speech alignment [11, Ch. 3]. This was solved through data selection, artificial stretching of the spontaneous speech and a proprietary alignment system which we do not have the details of. Unfortunately, no evaluation of this was performed.

In fact, a recent study in French [10] has demonstrated that correcting these differences can lead to improved synthesis quality. Using a corpus of sports commentaries [12] with hand-corrected alignment, an improvement in synthesis quality was achieved when using these manually corrected phonemisations for training and synthesis [10]. This shows that manually corrected transcriptions can benefit synthesis. It is, however, unclear whether this is due to the better phoneme accuracy in the alignment or due to a more natural pronunciation during synthesis. This paper's focus is on the first issue, whether better phoneme accuracy can improve standard synthesis despite being inconsistent across training and synthesis.

In Section 2 we present evidence that the standard phonemisation and forced alignment procedure produces many, and serious, mistakes, particularly with respect to spontaneous speech. In Section 3 the pronunciation variant alignment procedure is presented and objectively evaluated. Section 4 describes a perceptual evaluation of the resulting synthesis systems. The results are discussed in Section 5 before concluding in Section 6.

2. FORCED ALIGNMENT ACCURACY

In order to support the claim that spontaneous speech is less conformative to standard phonemisations than read speech, and to obtain a gold standard development set, the small corpus of read and spontaneous sentences from [13] was analysed. The corpus contains 50 sentences which were uttered in a normal conversation by a British English female voice talent. These sentences were orthographically transcribed and subsequently given to the voice talent to read aloud as standard prompts (the talent was unaware that she had uttered these sentences earlier) to obtain read-speech versions of them. The read and spontaneous versions of these 50 sentences thus contain exactly the same content and only differ in their acoustic realisation.

	Del	Add	Sub	Total	PER
Read					
Automatic	149	10	151	310	19.1%
Annotator 1	33	30	69	132	8.1%
Annotator 2	3	9	36	48	3.0%
Annotator 3	77	62	123	262	16.1%
Spontaneous					
Automatic	202	17	180	399	25.2%
Annotator 1	11	15	42	68	4.3%
Annotator 2	4	15	18	37	2.3%
Annotator 3	142	46	131	319	20.1%

Table 1. Overall differences between the agreed gold phonemisation, the standard automatic system and the annotators, with the gold phonemisation used as reference in each case. PER = Phone Error Rate.

2.1. Gold Standard Alignment

To create a gold standard phonemisation we first obtained an automatic alignment of the corpora and then manually corrected it. The automatic alignment was done using a large British English female Average Voice Model created using the Voice Cloning Toolkit [14] and adapting it to a corpora of 1176 read or 1146 spontaneous sentences from the voice talent speaker. The standard phonemisation was obtained using Festival and the RP British English version of the Combilex dictionary [15, 16]. For the spontaneous speech the transcription also included pausing, information which was provided for the model as it helped alleviate the cascading issue (see Section 2.3). Alignments of the 50 critical sentences of each type were then obtained using the respective models.

Next, the automatic alignments were independently manually corrected by two annotators. Where these hand-corrections disagreed the labellers met to discuss and agree upon a final phonemisation. In questions about whether to keep or to change the original Festival phonemisation, the Festival phonemisation was often preferred. Since human labellers correcting automatic transcriptions are biased toward the initial transcription [1], this agreed upon phonemisation is thus doubly biased *toward* the standard Festival phonemisation. To see the effect of this bias, a third transcriber corrected the output of the pronunciation variant system we propose (see Section 3). Finally, it is worth noting that the focus was on phoneme identity and not phoneme boundary; thus phoneme boundaries were only corrected if grossly incorrect, e.g., when a phoneme was deleted.

2.2. Transcriber Accuracy

To evaluate the phoneme accuracies of alignments we used the mean percentage deviation in Levenshtein distance (Phoneme Error Rate, or PER) with the manually corrected alignments as the gold standard (Table 1). While not the suggested method of [3, 1] it is standard [1, Ch. 1.32]. We used this as a measure of transcription accuracy since it is quick and easy to determine, and the agreed phonemisation of the two original annotators constituted our gold standard for development. See Section 4 for a perceptual evaluation of the resulting synthesis systems.

Table 1 shows each annotator, and the standard alignment system, compared to the phonemisation agreed by the two initial annotators. The automatic alignment is surprisingly bad in comparison, with over 19% of all phonemes wrong for the "simple" case of clear read speech prompts. For the spontaneous speech the PER is even higher, being above 25%. The standard Festival transcription adds very few phonemes. However, it deletes many phonemes and performs many substitutions, particular for the spontaneous speech. Of the additions, most are additional end of word stops, particularly "t" but often "d", and the main substitutions are "t"s for glottal stops and end of word "z" for "s". Together these account for 35% of all mistakes in the spontaneous and the pattern is similar for the read speech. What is notable here is that only the glottalisation could be considered speaker specific and non-standard in RP English (though common in many dialects). On the other hand, deletion of end of word stops and devoicing are common, in fact the main differences found here are similar to those in [17].

The third transcriber is almost as different from the phonemisation agreed upon by the two other transcribers as to the standard automatic method. This does not mean the third transcriber is closer to the standard Festival phonemisation. In fact they are further away, with a PER of 22.6% (read) and 32.0% (spontaneous) respectively (Table 3). The bias toward the transcribers' starting point thus appears very large. This is borne out when comparing to the lattice-based methods (see Section 3).

2.3. Cascading Deletion Errors

Due to the much greater amount of reductions and deletions in spontaneous speech, some utterances experience an issue of cascading alignment errors. This is not an issue of poor boundary alignment, but of poor automatic phonemisation. An example in our corpora is a realisation of the two words "basically because". In the read speech the automatic transcription is appropriate, but in the spontaneous the produced pronunciation is "basly 'cause". This produces not only the problem of non-existent phonemes being found, but also the much more serious issue of phonemes being "pushed" later in the utterance, putting every single phoneme further down the line out of alignment. This creates phonemes examples for model training which are grossly wrong. While the problem with because/'cause is arguably a lexicalised difference which could be resolved at transcription time, the issue of basically/basly is not, and such situations will cause problems in the trained models. Lexicalised differences were treated as orthographic transcription errors and corrected prior to alignment.

3. PRONUNCIATION VARIATION IN FORCED ALIGNMENT

We investigated a lattice-based forced alignment system to improve the phoneme identity accuracy on both read and spontaneous speech. Lattices have long been used to allow for pronunciation variation in automatic speech recognition and by researchers interested in speech segmentation, e.g., [18, 19, 20]. This is often avoided in synthesis due to the consistency assumption, given that the phonemisation found at training time cannot necessarily be produced during synthesis. This assumption, as discussed above, may not be correct.

The lattice-based system relies on two sources of information: hand-written variant options and pre-encoded pronunciation variants. The Combilex pronunciation dictionary contains pre-encoded pronunciation variants for a large portion of the dictionary, the average number of variants per word is 1.82. In the standard "full" pronunciation, normally pulled from the dictionary for both alignment and synthesis, these variants are not utilised. The system used here is based on the underlying context-dependent rewrite rules of Combilex as described in [21]. The system first finds all variants present in the dictionary and populates the lattice, realised as an FST, with them before expanding it using the additional expert-written variant options. These options are context-dependent rewrite rules written in terms of regular expressions. They generally take the form of a

phoneme with its left and right context and the resulting variant option, but can be any regular expression matching a valid string of symbols in the language of the Combilex dictionary. Thus the left and right context can be specific phonemes, features of the phoneme such as voicing, nasality and similar, but also word, syllable and phrase boundaries. For the set of manual rules, 14 rules were implemented based on the most common differences between the gold standard and automatic transcription from Section 2.1. Here are three examples from broad to quite specific:

- Any z can be devoiced.
- Any end-of-word t can be a glottal stop.
- A schwa after an f and before a stop can be deleted.

Note that the rules being based on the differences in the 50 sentences means they may be speaker dependent. We consider this acceptable as it is of interest whether or not a relatively small manual effort can improve the alignment procedure.

3.1. Method

Read and spontaneous corpora of respectively 1176 and 1146 sentences, plus the 50 matching sentences of the gold standard, were used as training data. Note that the 50 sentences both are in the training data and serve as test data. This is perfectly reasonable as we are attempting to create the best possible alignment of a known training set given no previous phoneme information. We are never trying to align unseen test data as the training data *is* the test data in our case. Festival was used as the front-end for producing the standard phonemisation of each utterance and a modified version of the multisyn tools for the alignment [22].

For alignment the multisyn tools rely on HTK [23]. A standard procedure was followed. As the lattices are initially created as FSTs these were converted to the HTK SLF format for alignment. After a first series of embedded training, the lattices were introduced simultaneously with the optional short pause models. The standard procedure was followed by a non-standard pause removal step, where short pauses under 40 ms were removed from the labels and models re-aligned and estimated. This was found to improve performance.

Four systems were built: the standard alignment method, lattice alignment using only Combilex pronunciation variants, lattice alignment using only manual rules and lattice alignment with both manual and Combilex variants. Each system was run in a variety of configurations, which for brevity is not detailed here, from which the best method was selected for use. In all cases this was with no standard multisyn phoneme substitutions, an additional pause removal step, five-state monophone models and flat-start alignment.

3.2. Results

Using the agreed gold standard transcription as the comparison point, Table 2 lists the performance of each system. The proposed full system, i.e., lattices with both rules and Combilex variants, improves the phoneme accuracy in both the read (2.8%) and spontaneous (5.4%) case compared to the standard method. However, the manual rules are better in both cases, improving the standard system by 3.9% and 6% respectively. Using lattices particularly reduces the number of phoneme additions, meaning they are more likely to delete phonemes compared to the traditional method. This is important as the cascading errors of Section 2.3 occur because of additional phonemes. For the spontaneous speech, substitution errors are also reduced whereas deletions increase slightly, though much less than the reduction in additions.

	Del	Add	Sub	Total	PER
Read					
Standard	10	149	151	310	19.1%
Lattice w. Combilex	6	139	184	329	20.2%
Lattice w. Rules	20	106	120	246	15.2%
Lattice w. Both	22	101	142	265	16.3%
Spontaneous					
Standard	17	202	180	399	25.2%
Lattice w. Combilex	9	178	199	386	24.4%
Lattice w. Rules	37	133	134	304	19.2%
Lattice w. Both	38	130	145	313	19.7%

Table 2. Overall differences between the reference agreed phonemisation and the different automatic systems.

	A1	A2	A3	Gold
Read				
Standard	17.3%	19.2%	22.6%	19.1%
Lattice w. Combilex	20.1%	20.2%	16.9%	20.2%
Lattice w. Rules	15.7%	15.2%	13.9%	15.2%
Lattice w. Both	16.8%	16.7%	9.1%	16.3%
Spontaneous				
Standard	23.0%	25.7%	32.0%	25.2%
Lattice w. Combilex	23.4%	25.1%	26.1%	24.4%
Lattice w. Rules	18.0%	19.9%	20.8%	19.2%
Lattice w. Both	18.6%	20.8%	16.5%	19.7%

Table 3. Overall PER differences between the automatic systems and the various annotators. A[1–3] are annotators; Gold is the agreed phonemistation between A1 and A2.

While it seems clear that the manual rules reduce the errors, it is not clear that the Combilex pronunciation variants do. This is likely due to the double bias toward the standard phonemisation. If using the third transcriber as the gold standard (Table 3), both the dictionary variants and the manual rules are beneficial, and in fact complementary. The manual rules still perform better on their own, but this may be due to these being tuned toward the test set. The third transcriber also favours the combined system, but massively disfavours the standard phonemisation, showing, again, the transcriber bias toward the initial transcription. Despite this, it is clear that the variant systems outperform the standard phonemisation in terms of phoneme accuracy.

4. SYNTHESIS EVALUATION

While we have shown improved PER, this need not translate into better synthesis quality. Furthermore, the discrepancy between annotator opinion and system PER casts doubt on the exact system performance. We therefore performed a synthesis-based evaluation, i.e., a task-oriented evaluation of the resulting alignments [24], which is also the ultimate target. Specifically, HTS 2.3beta [25] was used to train eight HMM voices: four of each speech type, each using either standard alignment or one of the three lattice systems. The same corpora as for the alignment were used, meaning 1176 read and 1146 spontaneous sentences recorded in a hemi-anechoic studio by a female British English voice talent for the purpose of speech synthesis. The 50 test sentences were excluded from voice training in all cases.

4.1. Listening Test Design

A listening test based on MUSHRA [26] was run. This is similar to a MOS test, but allows side-by-side comparison of stimuli (same sentence but different systems) on a sliding scale from 1–100. Stimuli were unlabelled and ordered randomly. Natural versions of both read

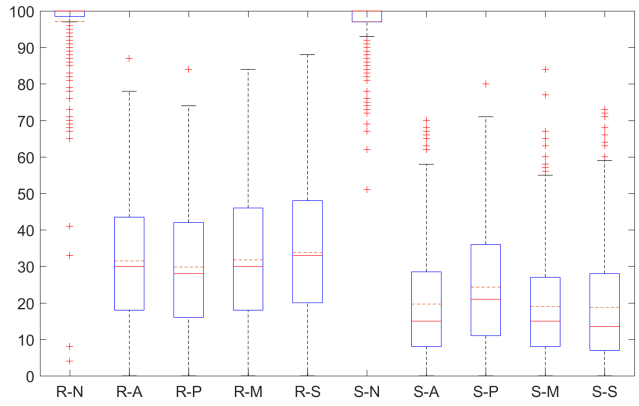

Fig. 1. Boxplot of perceptual test results. R = Read, S = Spontaneous, N = Natural; M = Lattice w. Manual rules, P = Lattice w. Combilex variants, A = Lattice w. Both, S = Standard method. Solid lines are medians, stapled means and boxes 25 and 75% quantiles.

System	R-N	R-A	R-P	R-M	R-S	S-N	S-A	S-P	S-M	S-S
S-S	*	*	*	*	*	*	1	*	1	-
S-M	*	*	*	*	*	*	1	*	-	
S-P	*	*	*	*	*	*	*	-		
S-A	*	*	*	*	*	*	-			
S-N	1	*	*	*	*	-				
R-S	*	<0.005	*	<0.05	-					
R-M	*	1	<0.05	-						
R-P	*	<0.05	-							
R-A	*	-								
R-N	-									

Table 4. Adjusted p-values after Holm-Bonferroni correction for the Wilcoxon signed-rank test. * = $p<0.001$. Labels as in Figure 1.

and spontaneous speech were included with the synthetic systems. Participants were asked to rate the stimuli according to how natural they sounded, with at least one stimulus at 100 and the others rated in relation to this. No designed reference sample was presented since each set included multiple natural productions. 32 paid participants were recruited and performed the experiment in a sound-insulated booth wearing Beyerdynamic DT770 PRO headphones. Each rated 15 sentences from either of two non-overlapping subsets randomly selected from the 50 test sentences, along with an initial practice sentence not included in the analysis. This amounts to 16 evaluations of each speech sample, for a total of 480 datapoints per system. All test materials are available at [27].

4.2. Results

The results of the test are graphed in Figure 1. Table 4 shows all system pairs compared using a Wilcoxon signed-rank test, after Holm-Bonferroni correction to avoid false positives. Natural speech is, unsurprisingly, rated significantly higher than synthetic speech. In contrast to [13] there was no significant difference between the natural speech types. However, this is probably due to the natural speech being so clearly more natural than the synthetic, causing differences between these two types to become much smaller. All the read speech-based voices were rated significantly higher than the voices built on spontaneous speech. For the read speech the standard alignment produces significantly higher rated speech than the other types, and the Combilex variants significantly lower, although the effect size is quite small. For spontaneous speech, on the other hand, the lattice system with only Combilex variants was rated significantly higher than all others, with no differences between the rest.

5. DISCUSSION

The consistency assumption between training and synthesis only partly holds. For read speech synthesis it seems to hold, as the standard method achieves higher ratings. Informal listening to the output of the proposed systems suggests that these systems produce hyper-articulated speech, which could reduce subjective naturalness. Arguably, however, we are getting what we ask for. At synthesis time we *ask for* the hyper-articulated version of the sentence, though we do not normally get it because of serendipitous reductions obtained due to the consistency assumption. Once we break that, a hyper-articulated version is produced. However, if we truly wish to control synthesis output, we should rather aim to have a better, more complete, acoustic model as provided by the proposed system. Methods for controllable, perhaps even gradeable, reduction of a sentence should then be developed, for instance utilising the reduced variants already encoded in dictionaries. While synthesis from the phonemisation found by the pronunciation variant alignments was not evaluated in a formal perceptual test, preliminary subjective evaluations are promising, indicating that pronunciation variant synthesis systems should be worth investigating. Synthesis from spontaneous speech, in contrast, appears to benefit from the use of pronunciation variants. It is encouraging that simply applying pre-encoded pronunciation variants helps us learn a better model, particularly on difficult, spontaneous speech data, where a fully pronounced alignment is highly inappropriate. While including manual rules always improved accuracy, as measured by the manual transcription, they did not increase perceived naturalness in synthesis. This may be due to them being overfitted to the test sentences, and thus not entirely suitable for the full training data. Synthesis models from read speech were rated more natural than spontaneous speech. This is not surprising since spontaneous data is much more varied and difficult to model, as exemplified by the much less accurate alignments. However, using pronunciation variant forced alignment pushed spontaneous speech towards closing the gap, and likely forms part of the alignment method used in [11]. It is also possible that the difference found between read and spontaneous is due to difference in phoneme accuracy and not specifically the type of speech, i.e. that read speech with a lower accuracy can benefit from such variant modelling.

6. CONCLUSION

We have reported on an investigation of the consistency assumption between training and synthesis in TTS. A pronunciation variant based forced-alignment method was implemented and its application to speech synthesis evaluated. It was found that standard synthesis with read speech did not benefit from these variants, though the underlying acoustic model arguably was more correct. For spontaneous speech, including pronunciation variation yielded an improvement, showing that the consistency assumption between training and synthesis only holds when minor errors are made. We suggest further improvements could be attained by incorporating automatic pronunciation reduction at synthesis time, this is considered future work.

7. ACKNOWLEDGEMENTS

R. Dall is supported by the JST CREST uDialogue Project. S. Brognaux is supported by FNRS. This work was partly funded by EPSRC grant no. EP/I031022/1 (Natural Speech Technology) and the Scottish Executive. As per EPSRC guidelines all experimental materials are available at [27].

8. REFERENCES

[1] C. Van Bael, *Validation, Automatic Generation and Use of Broad Phonetic Transcriptions*. PhD thesis, Radboud University Nijmegen, 2007.

[2] W. N. Campbell, *Computing Prosody*, ch. Synthesizing spontaneous speech, pp. 165–186. Springer, 1997.

[3] C. Van Bael, L. Boves, H. van den Heuvel, and H. Strik, "Automatic Phonetic Transcription of Large Speech Corpora," in *Proc. LREC*, (Genoa, Italy), pp. 4–11, 2006.

[4] Y. J. Kim, A. Syrdal, and M. Jilka, "Improving tts by higher agreement between predicted versus observed pronunciations," in *Proc. SSW5*, (Pittsburgh, USA), 2004.

[5] E. Shriberg, "Disfluencies in SWITCHBOARD," in *Proc. ICSLP*, (Philadelphia, PA, USA), pp. 11–14, 1996.

[6] H. Bortfeld, S. D. Leon, J. E. Bloom, M. F. Schober, and S. E. Brennan, "Disfluency rates in conversation: effects of age, relationship, topic, role, and gender.," *Language and speech*, vol. 44, pp. 123–47, June 2001.

[7] J. E. Fox Tree, "The Effects of False Starts and Repetitions on the Processing of Subsequent Words in Spontaneous Speech," *Journal of Memory and Language*, vol. 34, no. 6, pp. 709–738, 1995.

[8] S. Goddijn and D. Binnenpoorte, "Assessing Manually Corrected Broad Phonetic Transcriptions in the Spoken Dutch Corpus," in *Proc. ICPhS*, (Barcelona, Spain), pp. 1361–1364, 2003.

[9] S. Brognaux and T. Drugman, "Phonetic variations: Impact of the communicative situation," in *Proc. Speech Prosody*, (Dublin, Ireland), 2014.

[10] S. Brognaux, B. Picart, T. Drugman, and D. Louvain, "Speech synthesis in various communicative situations: Impact of pronunciation variations," in *Proc. Interspeech*, (Singapore, Singapore), 2014.

[11] S. Andersson, *Synthesis and Evaluation of Conversational Characteristics in Speech Synthesis*. PhD thesis, University of Edinburgh, 2013.

[12] S. Brognaux, B. Picart, and T. Drugman, "A New Prosody Annotation Protocol for Live Sports Commentaries," in *Proc. Interspeech*, (Lyon, France), 2013.

[13] R. Dall, J. Yamagishi, and S. King, "Rating Naturalness in Speech Synthesis: The Effect of Style and Expectation," in *Proc. Speech Prosody*, (Dublin, Ireland), 2014.

[14] 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.

[15] K. Richmond, R. a. J. Clark, and S. Fitt, "Robust LTS rules with the Combilex speech technology lexicon," in *Proc. Interspeech*, (Brighton, UK), pp. 1295–1298, 2009.

[16] K. Richmond, R. Clark, and S. Fitt, "On Generating Combilex Pronunciations via Morphological Analysis," in *Proc. Interspeech*, (Makuhari, Japan), pp. 1974–1977, 2010.

[17] J. Fackrell, W. Skut, and K. Hammervold, "Improving the accuracy of pronunciation prediction for unit selection tts," in *Proc. Interspeech*, 2003.

[18] D. Binnenpoorte, C. Cucchiarini, H. Strik, and L. Boves, "Improving automatic phonetic transcription of spontaneous speech through variant-based pronunciation variation modelling," in *Proc. LREC*, (Lisbon, Portugal), pp. 681–684, 2004.

[19] J. M. Kessens, M. Wester, and H. Strik, "Improving the performance of a Dutch CSR by modeling within-word and cross-word pronunciation variation," *Speech Communication*, vol. 29, no. 2, pp. 193–207, 1999.

[20] S. Paulo and L. C. Oliveira, "Generation of word alternative pronunciations using weighted finite state transducers," in *Proc. Interspeech*, (Lisbon, Portugal), 2005.

[21] K. Richmond, V. Strom, R. J. Clark, J. Yamagishi, and S. Fitt, "Festival multisyn voices for the 2007 blizzard challenge.," in *Proc. Blizzard Challenge Workshop*, (Bonn, Germany), 2007.

[22] R. A. J. Clark, K. Richmond, and S. King, "Festival 2 Build Your Own General Purpose Unit Selection Speech Synthesiser," in *Proc. SSW*, (Pittsburgh, USA), 2004.

[23] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book (for HTK Version 3.4)*. 2006.

[24] C. Van Bael, H. Van Den Heuvel, and H. Strik, "Validation of phonetic transcriptions in the context of automatic speech recognition," *Language Resources and Evaluation*, vol. 41, pp. 129–146, 2007.

[25] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. W. Black, and K. Tokuda, "The HMM-based Speech Synthesis System Version 2.0," in *Proc. SSW*, (Bonn, Germany), pp. 294–299, 2007.

[26] International Telecommunication Union Radiocommunication Assembly, Geneva, Switzerland, *Method for the subjective assessment of intermediate quality level of audio systems*, June 2014.

[27] R. Dall, "Experiment materials for "testing the consistency assumption: pronunciation variant forced alignment in read and spontaneous speech synthesis"," in *The Centre for Speech Technology Research, http://dx.doi.org/10.7488/ds/1314*, 2016.