Experiments in rapid development of accurate phonetic alignments for TTS in Afrikaans

Daniel R. van Niekerk
Multilingual Speech Technologies,
North-West University, Vanderbijlpark.
Human Language Technologies,
Meraka Institute, CSIR, Pretoria.
Email: dvniekerk@csir.co.za

***Abstract*—The quality of corpus-based text-to-speech (TTS) systems depends on the accuracy of phonetic annotation (alignments) which directly influences the process of acoustic modelling. In this paper we discuss the rapid development of accurate alignments for new languages in an under-resourced context based on an investigation of automatically obtained alignments of an Afrikaans speech corpus. We show that certain classes of inaccuracies can be effectively detected by acoustic analyses. Furthermore we discuss systematically addressing identified inaccuracies and evaluate the impact of such a process on synthesis quality of a statistical parametric synthesiser in this context.**

I. Introduction and motivation

Recently the development of tools enabling the efficient machine learning of pronunciation [1] and acoustic models [2] have allowed the rapid development of robust and intelligible "baseline" systems (implemented with few language-specific modules) in under-resourced environments [3]. However, successful deployment is hampered by the lack of naturalness achievable by such basic systems based on speech corpora of neutral prosody and careful pronunciation. This is especially the case in contexts where users are not familiar with TTS technology and its idiosyncrasies [4].

The rapid development of systems based on small speech corpora (comparable in size to [5]) of more naturally read speech raises additional considerations not only related to increased prosodic variation, but also increased phonetic variation (i.e. the possible realisations of specific phonemes). Examples of such variation are natural phonological processes such as assimilation and deletion (which is often language- and speaker-specific). In the ideal case well estimated context-specific acoustic models based on phonemic transcriptions are essentially appropriate phonetic models, however when the corpus size is limited and more contexts are shared between models, more phonetic variation may degrade synthesised speech quality in certain contexts.

Furthermore, additional concerns (typical in under-resourced contexts) that have the potential to affect acoustic modelling success include:

- Source text often lacks sufficient coverage from different domains and/or exhibit quality problems such as code switching and large amounts of unpronounceable (by standard spelling rules) tokens. This compromises phonetic coverage and potentially also fluency during recording.
- Language-specific knowledge and/or implementations are usually not immediately available and thus systems often rely on generic text analysis and pronunciation prediction modules (see Section II for more details).
- Practitioners and voice talents involved in recordings are often not familiar with the technological constraints or are inexperienced, making the reduction of phonetic variation at this stage impractical.

In this paper we investigate the above concerns starting with an analysis of automatically obtained phonetic alignments of a typical corpus (Section II). In Section III we explore the possibility of automatically detecting potential problems (discrepancies) in alignments based on acoustic analysis. Sections IV and V discuss a systematic way of addressing discrepancies and the details of the intervention on our corpus. This is followed by an analytical and perceptual evaluation on the resulting acoustic models, a discussion on the contribution of this work and proposal of future work (Sections VI and VII).

II. Speech corpus and error analysis

A single speaker Afrikaans speech corpus carefully recorded in a professional studio environment for the purpose of TTS development is presented here. It is based on text consisting of 1005 sentences selected for diphone unit coverage [6]. The corpus was automatically aligned using a Hidden Markov Model-based (HMM) forced-alignment process (as described in [7]) to the output of our system's natural language processing (NLP) front-end, resulting in the corpus statistics presented in Table I. The front-end currently implements basic tokenisation, phrase break insertion based on punctuation, pronunciation prediction via grapheme-to-phoneme (G2P) rewrite rules and rule-based syllabification. The G2P rules, trained using the Default&Refine algorithm [1] from a phonemic pronunciation dictionary, is the sole source of pronunciation prediction, resulting in exactly one possible pronunciation for each word. This represents a "baseline" TTS system and reflects a typical scenario when developing TTS for under-resourced languages. Components that are notably absent compared to more advanced systems are part-of-speech tagging, morphological analysis and sophisticated phrase break prediction.

Utterances	Phrases	Words	Syllables	Phones
1005	1134	9225	15153	40451

TABLE I
Original corpus statistics

The corpus was manually inspected using Praat [8] with word, syllable and phone alignments visible. During this process, the following set of discrepancies were identified and labelled according to the source of the discrepancy:

A: Pronunciation mismatches due to under-articulation in continuous speech (e.g. deletions, assimilation or reduction).
B: Gross alignment errors (i.e. alignment errors that are clearly visible during inspection).
C: Alignment discrepancies due to unexpected pauses.
D: Mispronunciations (usually due to speaker error or extreme dialectal variation from expected case).

E: Label or alignment discrepancies due to transcription mismatches (usually due to uncaught reading mistakes).
F: Pronunciation mismatches due to foreign words.
G: Pronunciation mismatches due to G2P inaccuracies.

Discrepancies falling into these classes were marked on the word level. With the exception of A and C, all classes could be consistently identified but due to the tediousness of the manual process it is likely that a certain number of existing cases were overlooked. In the case of A, the phenomena occur in varying degrees making it only possible to consistently identify extreme cases. Table II contains frequencies of affected speech units (cases are not necessarily mutually exclusive):

Class	Utterances	Words
A	> 105	> 122
B	39	60
C	272	352
D	12	12
E	59	85
F	52	68
G	95	104
Total	> 493	> 791

TABLE II
DISCREPANCY FREQUENCIES

Having no discrepancies in these classes would be ideal in the sense that it would mean that the corpus is both error free and that the TTS front-end is perfectly predicting all relevant aspects (pronunciation, phrasing, etc.) of the specific speaker's speech. This would allow the acoustic modelling to proceed optimally given features that can be predicted from the text and thus effectively used during synthesis.

Based on the frequency of occurrence of discrepancies (conservatively about 49% of utterances and 9% of words), it is fair to assume that the quality of acoustic models will be significantly affected. Moreover, our experience is that this is a typical example when working in an under-resourced environment (due to the difficulties listed in Section I).

The following section explores the possibility of automatically detecting the discrepancy classes presented in this section.

III. DETECTING DISCREPANCIES

In order to flag the most significant discrepancies we evaluated a few features directly obtainable from the process of alignment and acoustic modelling.

Alignment log-likelihoods: using the log-likelihoods obtained from the forced-alignment process performed in Section II. Mean word scores are calculated from scores output by HTK [9].

Mel-cepstral distance: comparing synthetic speech samples with original instances from the corpus (similar to the distortion measure used in [10]). We evaluated two approaches: firstly simply training HTS [2] models (see Section V) on the full corpus and synthesizing all the training utterances and secondly in a 10-fold cross validation fashion where 10% test-sets where held out for comparison. The comparison of individual speech samples were attempted using both a dynamic time warping (DTW) approach as well as using the durations of the original speech alignments to allow direct frame-by-frame comparisons (as done in [10]). For the feature extraction we followed the convention used in [11] (in the section on DTW for alignment), simply using the Euclidean distance between static and delta (without energy; 24 coefficients in total) Mel-Frequency Cepstral Coefficients (MFCCs extracted with Edinburgh Speech Tools [12] every 5ms, using 25ms Hamming windows). We also experimented with the Mahalanobis distance (with the covariance in each experiment calculated on the training set) without additional success (results are not reported here).

Absolute duration differences: calculating the duration difference between synthesised samples and aligned original samples from the corpus.

A. Results

By examining detection rates of words identified in Section II using these features, we identify which of the features are most effective at highlighting discrepancies from the different classes defined above, but also gain insight into how these discrepancy classes are potentially affecting the acoustic modelling process. Figures 1 to 7 show receiver operating characteristic (ROC) curves when using basic statistics (the mean distance or absolute duration difference) on the word level to flag potential discrepancies.

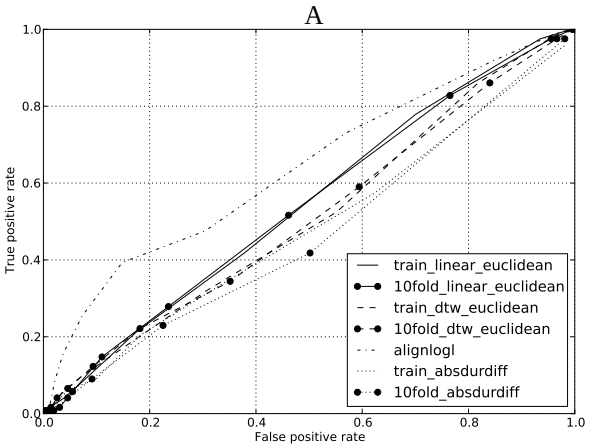

Fig. 1. ROC curves for class A for the features described in Section III

As mentioned in Section II, class A discrepancies were the most difficult to mark consistently, ranging from assimilation and deletions, to vowel reduction and voicing changes in some consonants. Considering the non-homogeneous nature (acoustically) of this class, the shape of the best ROC curve is understandable. It is also noted that none of the features based on comparison with TTS acoustic models are effective here. This is not unexpected, as the "full-context" TTS models can be expected to model these effects more effectively than simple triphone models used during alignment.

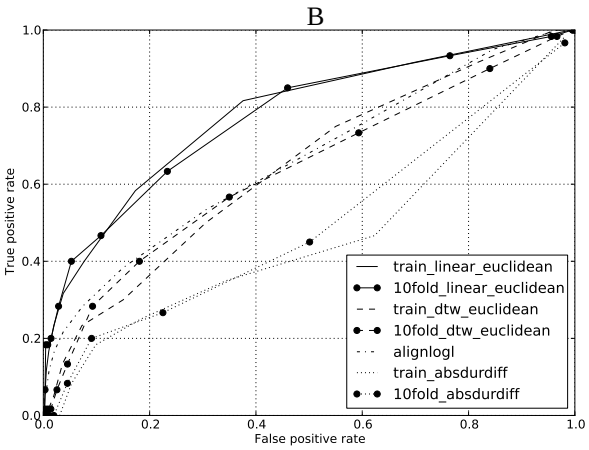

Fig. 2. ROC curves for class B for the features described in Section III

Class B discrepancies most often co-occurred with other problems (especially class E) to the extent that for the purposes of troubleshooting alignments, almost all of these cases might be reclassified into other classes based on underlying cause.

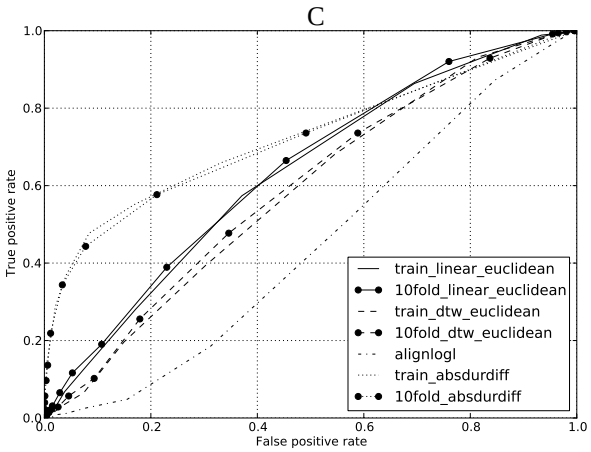

Fig. 3. ROC curves for class C for the features described in Section III

In class C, there is again more than one exact cause for pauses (or silences) that are not modelled by the TTS front-end. Two of the most prevalent being: undetected phrase breaks and glottal closures (which occurs before some vowels and often between two vowels that do not flow into each other). The best curve (based on the duration difference measure) presumably detects the longer pauses due to missed phrase breaks (which are not predictable from phonological context) well, but is less effective in the case of glottal closures which are shorter and phonologically more predictable.

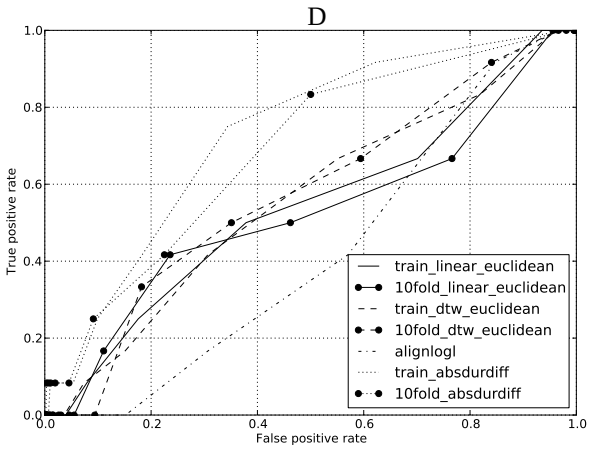

Fig. 4. ROC curves for class D for the features described in Section III

The remaining classes all present clear (easily identifiable) discrepancies between the audio and predicted labels (and associated alignments), with class E being the easiest to detect and G relatively difficult (presumably because of the fact that G2P errors occur in a more consistent/predictable way).

Comparison of the alignment log-likelihood with the cepstral distance features in general seems to suggest that the cepstral distance is more effective at detecting the identified discrepancies in this

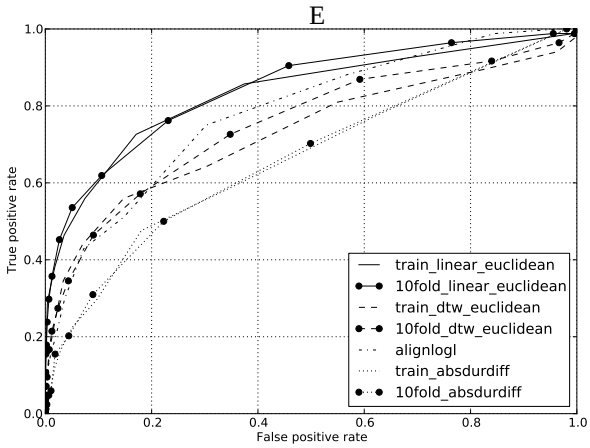

Fig. 5. ROC curves for class E for the features described in Section III

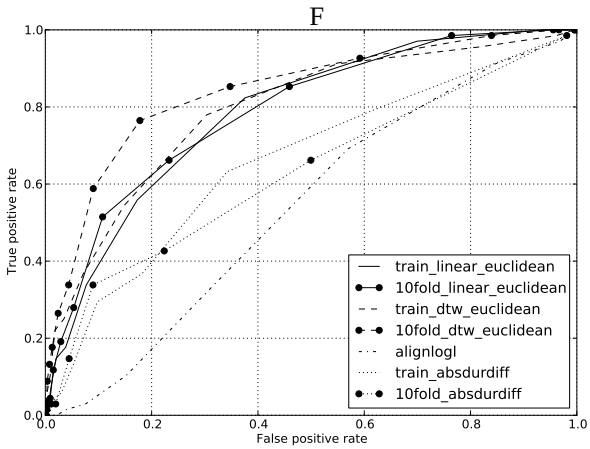

Fig. 6. ROC curves for class F for the features described in Section III

context. The exception is A discussed above. In the case of class C, the duration difference is clearly a sensible measure.

These results should allow us to design automated or machine-assisted corpus refinement processes or may be applied during the synthesis process (similar to the way in which alignment log-likelihoods are used in some unit-selection systems [13]), especially when corpora of similar nature are used.

In the following section we consider how such refinements can be efficiently effected.

IV. Addressing identified discrepancies

Technically, the simplest idea is to discard parts of the data exhibiting large discrepancies. We briefly experimented with this approach by systematically removing utterances based on the mel-cepstral distance measure without succeeding in improving the perceivable quality of HMM-based synthesis or measured cepstral difference (as in section VI-A).

An alternative is to resolve discrepancies before acoustic modelling. Manual correction of alignments might result in a more accurately annotated corpus, but would lead to a disconnect between the labels predicted by the TTS front-end and acoustic models which might lead to a degradation in synthesis quality, especially

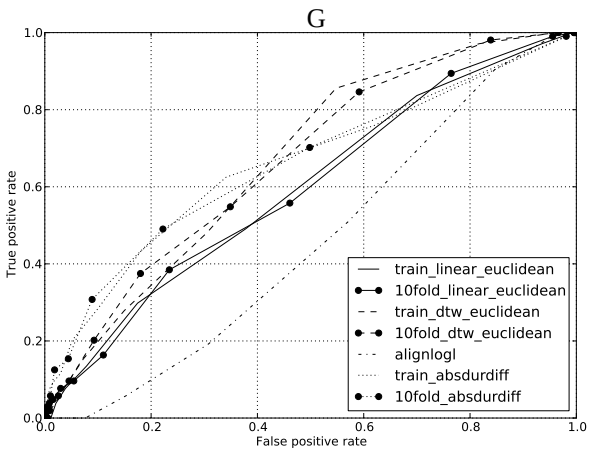

Fig. 7. ROC curves for class G for the features described in Section III

for contexts where consistent mismatches (such as class A) might have allowed acoustic models to compensate for discrepancies [14].

Given these considerations and the analysis presented in Sections II and III, we evaluate a framework for resolving discrepancies based on allowing multiple pronunciation variations during the alignment stage, similar to what is described in [13], but additionally allowing us to include pronunciation variation that only occurs in certain utterance contexts (e.g. deletions due to cross-word contexts). This framework (depicted in Figure 8) allows us to resolve difficult cases automatically (relying on alignment models to select an appropriate transcription) and capture relevant variation (e.g. speaker-specific pronunciations) for possible use at synthesis time.

By multiplexing predictions from complete TTS front-ends via a context-specific pronunciation dictionary (where different instances of a word occuring in different contexts are allowed to have distinct pronunciations), one is afforded flexibility to implement speaker-specific pronunciation variation using the most compact or appropriate mechanism available (e.g. speaker-specific lexicon, phonological rules, etc.) which is then presented to the alignment stage as a set of straightforward context-specific pronunciation variants (on the word level). Training of HMM models for alignment follows the standard recipe (see [9] Chapter 3), using the "standard front-end" for initial pronunciations and selecting more appropriate alternatives during the "re-alignment" stage prior to final embedded re-estimation and forced-alignment.

In the final sections of this paper we investigate the impact of addressing identified discrepancies, in the way described in this section, on HMM-based acoustic models. The next section starts by describing the details of intervention and acoustic modelling on our corpus.

V. Afrikaans voice

For alignment of our corpus we decided to address the problem classes within the framework defined in the previous section, using our original TTS front-end as the "standard-frontend" and one additional "speaker-specific front-end". The letters in parentheses in Figure 8 indicate where each of the discrepancy classes were addressed and a description follows.

Our speaker-specific front-end starts out by inheriting all processes and resources from the original front-end. We then specialise this implementation to address **class A and D discrepancies** by applying phonological rules occurring in continuous speech after initial pronunciation prediction (described in section V-A) and a pronunciation addendum which overrides the standard pronunciations respectively. **Class C and E discrepancies** are addressed in the orthographic

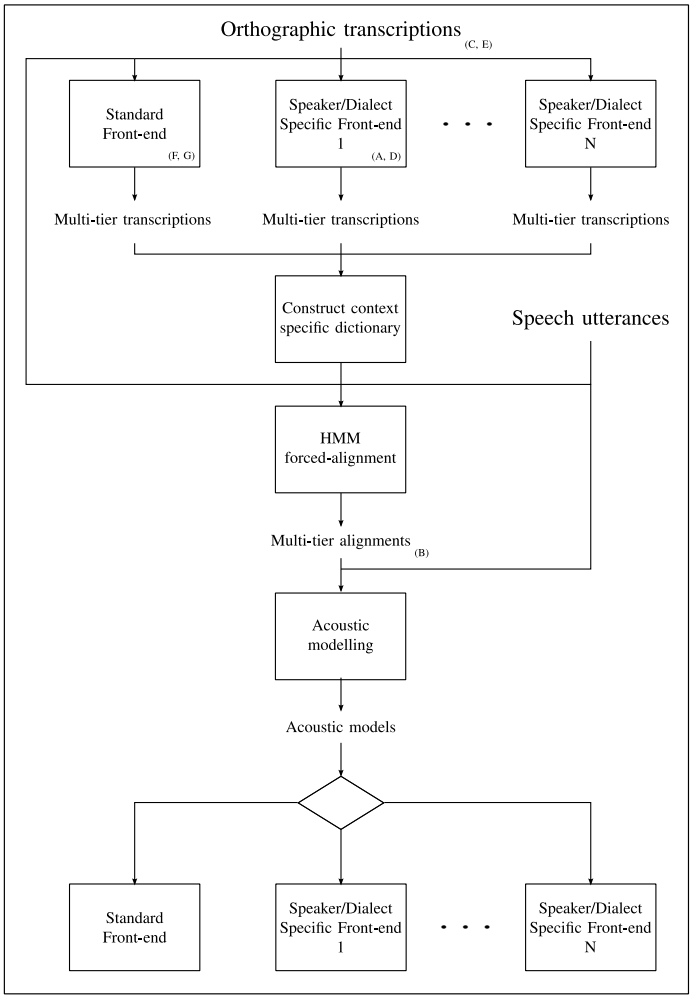

Fig. 8. Alignment and acoustic modelling process

transcriptions by inserting commas (our original front-end inserts phrase breaks based on punctuation) and correcting any transcription errors. The remaining **F and G discrepancies** are addressed in the original front-end in the pronunciation addendum and main lexicon respectively. We did not explicitly address issues marked as **class B**, believing that resolving other classes would generally lead to resolution of these problems (see discussion in Section III). However, if these were to be addressed, a logical way would be to edit alignments directly, assuming that any remaining cases of this nature are simple failures of the automated alignment process. Table III quantifies the extent of manual intervention as described in this paragraph.

Original front-end	
Additional pronunciation entries	98
Speaker-specific front-end	
Additional pronunciation entries	111
Phonological rules	3
Transcriptions	
Sentences edited	317

TABLE III
MANUAL INTERVENTION

Although we also implemented optional insertion of pauses (as

done in [13]) between words, this resulted in 1208 insertions (compared to the 352 manually identified - see Table II). Thus, in order to handle this information automatically we would need to further investigate the accurate classification of phrase breaks as done in [15].

A. Phonological rules

As stated, phonological rules generating alternative word pronunciations in certain contexts were implemented. No attempt was made to define an exhaustive list of known rules for Afrikaans, but rules were based on inspection of the lowest scoring words (based on the alignment log-likelihoods and cepstral distances described in Section III). This led to the implementation of 3 phonological rules addressing deletion of phones in different contexts of continuous speech:

1) Where the first sound of a word can be deleted (e.g. the word sequence "van die" is often realised [fɑni] instead of [fɑndi]).
2) Where the last sound of a word can be deleted (e.g. "met die" is often realised [mɛdi] instead of [mɛtdi]).
3) Where the [r] can be deleted at the end of the first syllable [fər] in a polysyllabic word (e.g. in the word "verklaar").

B. Acoustic modelling

In this section we describe the acoustic modelling process for the purpose of evaluation. Re-aligning the corpus with the above amendments (Section IV) resulted in the corpus statistics in Table IV. HMM-based acoustic models were trained using the standard demonstration script available as part of the HMM-based Speech Synthesis System (HTS) version 2.2 [2]. Model labels (features) similar to [16] were used (as far as these features were available - recall details of the TTS front-end implementation described in Section II). For the model tying decision tree, phone and word contexts as well as phonetic classes defined in the phone set (e.g. broad phonetic classes and features such as plosives, nasals, vowels, voicing, etc.) were used to define questions.

Utterances	Phrases	Words	Syllables	Phones
1005	1428	9216	15094	40233

TABLE IV
NEW CORPUS STATISTICS

For the experiments in Section VI-A we randomly selected a small set of utterances to serve as a test set. Alignments for the test set were manually corrected by checking all segment labels, phrase breaks and in some isolated cases moving segment boundaries. This process resulted in the corpus statistics as shown in Table V.

Set	Utterances	Phrases	Words	Syllables	Phones
Train	955	1356	8772	14351	38263
Test	50	69	444	741	1949

TABLE V
TRAINING AND TEST SET STATISTICS

VI. RESULTS

In the following sections we evaluate the effects of the work presented in Section V on resulting synthetic speech.

A. Mel-cepstral distance

Firstly, we measure the mel-cepstral distance between the synthesised test utterances (Table V) and original speech samples, using the phone label sequences as corrected. The original and updated alignment procedures are compared in this way by complete re-alignment and re-training of acoustic models for subsets of the training utterances ranging from 1% to 100%. The result is presented in Figure 9.

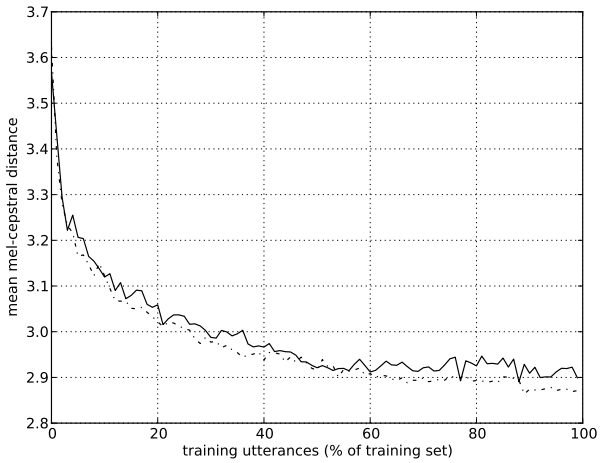

Fig. 9. Results: Mean mel-cepstral distance on the held-out test set. The solid line represents the mean distance when using the original alignments and the dashed line when using the updated alignments.

We observe consistently lower distances over the test set when using the updated alignments (based on the process in Figure 8), suggesting that the process is robust and results in a positive effect (as defined by our test set) on acoustic models.

B. Perceptual evaluation

Ultimately, we would like determine whether the improvement in spectral quality suggested by results in the previous section combined with the expected improvement in phrase prosody (due to the clarification of phrase breaks in the training corpus) translates into an overall improvement in perceived quality of the synthesised speech in our context.

While the alignment process (Figure 8) is able to select different pronunciations for different instances of the same word, our current implementation requires selecting a consistent strategy (i.e. we have to choose either the original or speaker-specific front-end) during synthesis time.

For this evaluation we choose to use the original front-end to determine the effect of enforcing this style (of pronunciation) from the given corpus. Our preliminary informal trials confirmed that the synthesis of contexts covered by the implemented phonological rules (e.g. "van die") using the original front-end results in noticeably clearer realisations of phones compared to acoustic models based on the original alignments.

To test overall preference we asked 11 respondents to indicate preference between 13 single-phrase sentence pairs synthesised with the original front-end, but using acoustic models based on the different alignment sets (Tables I and IV). For each sentence comparison we also invited respondents to list specific words in the utterances that influenced their decision as well as general comments.

The results of this evaluation (143 utterance comparisons) is presented in Table VI, showing the number of utterances preferred according to alignment set (or no preference). According to McNemar's test statistic using the chi-squared distribution with 1 degree of freedom, the 95% confidence level is given by:

$$\frac{(|b-c|-0.5)^2}{b+c} \geq 3.841 \quad (1)$$

A chi-squared value of 1.894 implies that the perceptual difference on our sample of short sentences is not significant.

Original	New	No preference	Total	χ^2
48	63	32	143	1.894

TABLE VI
PERCEPTUAL PREFERENCE

We discuss our results in the following section, followed by conclusions drawn from the work presented.

C. Discussion

Based on the results presented in VI-A and the inspection of synthesised examples of contexts covered by the implemented phonological rules, we argue that the work presented in Section V results in clearer phonetic models (i.e. models that are acoustically closer in nature to their phonetic labels). In Section I we argued why this might be beneficial in the case of acoustic modelling in sparse contexts. Furthermore, a more phonetically (as opposed to phonemically) labelled corpus could ease research and development of polyglot systems or systems based on resource sharing where phone sets from different languages need to be combined or compared.

In Section VI-B we asked respondents to indicate general preference, inviting them to comment on aspects of the speech that influenced their decisions. According to these comments decisions were most often made based on prosodic differences and respondents were less sensitive to under-articulation in samples based on the original alignments. Though we suspect that the clarification of phrase breaks in the updated alignments had a positive effect on prosody and phone durations in some samples, the lack of information (such as word emphasis) included in the front-end implementations could explain some of the variable results. Another point worth noting is that the modelling of fundamental frequency (f0) in HTS is tied to the phone identity and thus the process of training followed by application of acoustic models in a different context (e.g. by using the original front-end during synthesis) might affect f0 generation negatively.

A more extensive perceptual experiment including multi-phrase sentences (including evaluation of unit-selection voices) could result in significant further insight.

In the following section we present our conclusions.

VII. CONCLUSION

In this paper we considered the rapid development of accurate alignments for new languages in an under-resourced context based on an investigation of automatically obtained alignments of an Afrikaans speech corpus.

Here we highlight contributions and conclusions:

- An analysis of automatically obtained alignments for our corpus resulted in the definition of classes of discrepancies in alignments according to underlying cause.
- Methods of acoustic analysis were presented and evaluated for their utility in detection of identified discrepancies. We show that a number of problems can be effectively detected and that the mel-cepstral distance measure is more effective than the alignment log-likelihood scores for detecting a number of problem classes.
- We described an alignment framework and demonstrated how the identified problem classes can be addressed effectively. It was shown that following such a process on our corpus led to synthesised speech that is closer to natural speech (in terms of mel-cepstral distance on our manually aligned test set).
- A perceptual evaluation has highlighted further aspects that need to be considered in order to improve the quality of synthesised speech in this context.

To summarise, the work in this paper has taken a step in the direction of rapidly developing a more phonetic (as opposed to phonemic) annotation of a speech corpus. Future work could focus on further automation of this process, for example by applying the information obtained in Section III to propose phonological rules and alternative pronunciations (as implemented in Section V) without manual intervention, as well as incorporating phrase break detection as done in [15].

Towards the ultimate goal of improved synthesis quality, further work on automatically adapting the TTS front-end predictions based on acoustic analyses (similar to [17]) is needed. Improved prosodic modelling and synthesis also needs to be addressed in our Afrikaans system in particular.

VIII. ACKNOWLEDGEMENTS

Georg Schlünz assisted with the labelling of discrepancies in the corpus. The authors would also like to thank all participants involved in evaluating speech samples and reviewers for valuable comments and suggestions.

REFERENCES

[1] M. Davel and E. Barnard, "Pronunciation prediction with Default&Refine," *Computer Speech and Language*, vol. 22, pp. 374–393, 2008.

[2] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. W. Black, and K. Tokuda, "The HMM-based speech synthesis system (HTS) version 2.0," in *The 6th International Workshop on Speech Synthesis*, 2006.

[3] Meraka Institute, "Lwazi Project Final Report," CSIR, Pretoria, South Africa, http://www.meraka.org.za/lwazi, Tech. Rep.

[4] A. S. Grover and E. Barnard, "The Lwazi Community Communication Service: Design and Piloting of a Voice-based Information Service." in *WWW 2011*. IW3C2, 2011, pp. 433–442.

[5] J. Kominek and A. Black, "The CMU arctic speech databases," in *The 5th ISCA Speech Synthesis Workshop*, 2004, pp. 223–224.

[6] J. P. Van Santen and A. L. Buchsbaum, "Methods for optimal text selection," in *Proceedings of EUROSPEECH*, Rhodes, Greece, September 1997, pp. 553–556.

[7] D. R. Van Niekerk and E. Barnard, "Phonetic alignment for speech synthesis in under-resourced languages," in *Proceedings of INTERSPEECH*, Brighton, UK, September 2009, pp. 880–883.

[8] P. Boersma, *Praat, a system for doing phonetics by computer*. Amsterdam: Glott International, 2001.

[9] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Veltchev, and P. Woodland, *The HTK Book (for HTK Version 3.3)*. http://htk.eng.cam.ac.uk/: Cambridge University Engineering Department, 2005.

[10] J. Kominek, T. Schultz, and A. Black, "Synthesizer Voice Quality of New Languages Calibrated with Mean Mel Cepstral Distortion," in *Proceedings of the International Workshop on Spoken Language Technology for Under-Resourced Languages (SLTU)*, 2008.

[11] A. W. Black and K. Lenzo, *Building Synthetic Voices*, http://www.festvox.org/bsv, 2007.

[12] P. Taylor, R. Caley, A. Black, and S. King, *Edinburgh speech tools library*, 1999.

[13] R. A. J. Clark, K. Richmond, and S. King, "Multisyn: Open-domain unit selection for the festival speech synthesis system," *Speech Communication*, vol. 49, no. 4, pp. 317–330, 2007.

[14] M. Makashay, C. Wightman, A. Syrdal, and A. Conkie, "Perceptual evaluation of automatic segmentation in text-to-speech synthesis," in *Proc. ICSLP*, vol. 2, Beijing, China, Oct. 2000, pp. 431–434.

[15] K. Prahallad, E. Raghavendra, and A. Black, "Learning Speaker-Specific phrase breaks for Text-to-Speech systems," in *Proceedings of Speech Synthesis Workshop (SSW7)*, 2010.

[16] K. Tokuda, H. Zen, and A. W. Black, "An HMM-based speech synthesis system applied to English," in *IEEE Speech Synthesis Workshop*, 2002.

[17] C. Bennett and A. Black, "Prediction of pronunciation variations for speech synthesis: A data-driven approach," in *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, 2005.