Automatic error detection in alignments for speech synthesis

E Barnard and M Davel

Human Language Technologies Research Group
Meraka Institute, Pretoria
ebarnard,mdavel@meraka.org.za

Abstract

The phonetic segmentation of recorded speech is a crucial factor in the quality of concatenative systems for speech synthesis. We describe a a likelihood-based error detection process that can be used to flag possible errors in such a segmentation, with a view towards manual correction. It is shown that this process can be used to assist in the creation of high-accuracy segmentations. In particular, for an isiZulu corpus used in the creation of a unit-selection synthesizer, almost half of the errors that existed in a manual segmentation were detected by this process, while flagging fewer than a quarter of all segments. Different phoneme classes are handled with differing amounts of success, with vowels being the most troublesome.

1. Introduction and background

Speech synthesis is a technology of great potential significance in the developing world, since it provides a means to convert electronically stored information to a spoken format [1]. Hence, people with limited literacy, or disabilities that make it impractical to read, can gain access to documents that would otherwise not be available to them. Also, those without internet access can avail themselves of information in databases using a normal telephone service, if speech synthesis is utilized along with appropriate telephone services. Synthesis can therefore play a significant role in narrowing or bridging the digital divide.

Two broad categories of speech synthesis can be distinguished. Early synthesizers generally employed *model-based approaches,* where the speech-production process is described with a parametric model, and the parameters of the model are varied in time to produce synthesized speech. Perhaps the most successful model of this nature is based on the work of Klatt [2], in which the parameters are related to formant frequencies and amplitudes (and corresponding excitation sources). Work on these approaches continues to attract attention, and models employing parameters related to the positions of the articulators have been used with some success in recent years.

A significant amount of attention in both the academic and commercial arenas has, however, shifted to *concatenation-based approaches* [3]. Here, one no longer attempts to derive an explicit model of speech production – instead, speech segments are excised from a corpus of recorded speech, and spliced together to produce synthesized utterances. Concatenation is generally believed to produce utterances that are both more natural and more understandable than model-based synthesis, at the cost of increased effort in recording, preparing, storing, and searching the corpus of recorded speech.

Of these factors, the most expensive in practical cases is the preparation of recordings. In particular, the quality of synthesis depends sensitively on the accurate location of segment boundaries within the recorded speech, since these boundaries are used as markers for the excision of the segments that are concatenated. The production of these *time alignments* of phonetic sequences against spoken utterances is both time consuming and error prone. A trained transcriber has to listen to each segment, and manually (a) label the segment at a pre-determined level of phonetic refinement as well as (b) place boundary markers between the segment and its neighbours, according to specified conventions. Both of these tasks often involve debatable judgment calls, and gross errors as well as subtle misjudgements (from the perspective of synthesis quality) are common, especially when the transcribers are not highly experienced at this task.

Automation of the alignment process has therefore been used to a greater or lesser extent by many developers [4]. If a speaker-independent recogniser for the target language exists, it can produce a forced alignment against the orthography of the recorded speech (which is generally available, since the recordings are typically made from an orthographic script). Alternatively, a speaker-dependent recogniser can be trained on the recordings that form the synthesis corpus, if it is of sufficient extent. Both these approaches can produce initial alignments that greatly accelerate the manual transcription process, but neither is generally sufficiently accurate to eliminate the need for manual alignment altogether.

In this paper, we describe an approach that can further reduce the alignment time, by identifying likely errors in manual or automatic alignments. As described in Section 2, this approach uses whole-segment spectral information to evaluate segments more accurately. In Section 3 we demonstrate that a simple implementation of this approach can detect a significant fraction of alignment errors in an isiZulu corpus, which was used to build a basic isiZulu synthesizer. Section 4 concludes

with an overview of our method and results, and contains suggestions for extensions and refinement of this work.

2. Approach: difference of segment spectral means

In order to validate alignments, we start from the assumption that the great majority of segments have been aligned properly. Hence, for any measurement used to describe an acoustic segment, the average value of the measurement for a particular phone, calculated over all aligned utterances, should closely approximate the true value of that measurement if no alignment errors had been made. Consequently, any phonetic segment whose measurement value deviates substantially from that of the mean is possibly in error (either because one or both boundaries have been placed inappropriately, or because the segment is mislabelled, or possibly because the speaker produced a highly atypical phonetic variant).

This observation is equivalent to a simple algorithm for the detection of alignment errors, but practical implementation of the algorithms requires a number of choices, including:

- What measurements are used to describe phonetic segments? Ideally, the measurements should capture the nature of each segment, and both static and dynamic information is therefore relevant. However, the mean value of each measurement must be determined from a limited set of segments; we have therefore employed the mean spectrum (computed within 64 equal-sized spectral ranges on the Bark scale, spanning the range 0-8 kHz).
- How should differences between segments be measured? A number of metrics may be considered, but the paucity of data is again an important consideration. We have experimented with two metrics; both model the data using full covariance Gaussian distributions, but one method employs a pooled covariance matrix across all segments, and the other computes a covariance matrix for each phone individually.
- How should the threshold for candidate errors be set? False positive errors (i.e. flagging a candidate which was actually correct) are less troublesome than false negatives (errors which are accepted as correct segments). Thus, the chosen thresholds should err on the side of caution. However, as we shall see below, the variance of deviations within the different phone classes is substantial, so achieving a uniform level of conservatism across phones is a significant challenge. Our current approach therefore relies on a manual process for choosing the threshold for each class.
- How much context dependence must be modelled in the choice of classes? Co-articulation will certainly cause the realization of phones to vary depending on their surrounding phonetic context. Thus, if one has enough data to calculate accurate context-dependent models, such models are preferable. As we will see below, however, this was not the case for our isiZulu corpus.

3. Evaluation

We have evaluated our approach by testing its performance on a corpus used to build an isiZulu speech synthesizer (see [5] for details). Our isiZulu phone set consists of 47 phones. To obtain sufficient (though not complete) coverage of the diphones formed from this set, we selected 153 sentences from a public-domain text corpus that we had collected – the selection process attempts to cover the required diphones with the minimal number of sentences [6]. A male first-language isiZulu speaker from the Kwazulu-Natal region recorded these sentences. Our initial alignments were produced in a two-step process. Firstly, alignments that were obtained by mapping all isiZulu phones to the nearest English phones (using subjective, linguistically motivated criteria), and then using dynamic time warping to align the isiZulu utterances against an English voice which is distributed with the open-source Festival toolkit. This results in fairly crude alignments, since the isiZulu phones often are fairly different from the most similar English cognates.

These alignments were corrected by two isiZulu speakers with limited linguistic training (one was an undergraduate student in linguistics, the other had no formal linguistic training) and no prior experience with transcription. The transcribers aligned separate portions of the corpus, and did not crosscheck one another (though they were encouraged to discuss transcription conventions). Hence, these transcriptions were substantial improvements on those obtained automatically, but fell significantly short from those that would be obtained in a professional voice-development environment. We consider these alignments typical of what can be expected for first-time development of synthesizers in the developing world.

Below, we report on the effectiveness of our automatic process in correcting these manual alignments. The “correct” alignments, against which the improvements were measured, were those that we eventually used for synthesis. These were obtained by crosschecking of the manual alignments by the authors, and also by correction of additional errors that were detected during use of the corpus for synthesis. (The non-systematic nature of this process and the subjective evaluation inherent in alignment imply the likely presence of undetected alignment errors in our corpus, but random sampling indicates that these are sufficiently

rare not to impact our overall results.) A phone segment was scored as "incorrect" if it had the wrong label, or if either boundary differed from the correct location by more than 500 msec. The corpus consists of a total of 8 388 segments, and of these 152 were determined to be erroneous. Because several of our phonetic categories contained fewer than 20 segments, it was not feasible to use context-dependent models – in fact, we were also forced to use a pooled covariance matrix because of the limited number of segments in these categories. (It would be possible to use separate diagonal covariance matrices for most, but not all, classes. However, that would introduce additional complexity, since separate normalization schemes would be required depending on whether the pooled pr individual matrices are used. We have not pursued this approach.)

The overall success of our error detection process is summarized by the facts that (a) the error-detection process flagged 24.5% of all segments as possibly erroneous, and (b) this set contained 43.4% of all errors present in the corpus. Fig. 1 shows the fraction of errors detected and segments flagged as the detection threshold is varied, for four sounds: *"pau"*, *"i"*, *"g"* and *"r"*. (*"pau"* represents pauses that occur within or adjacent to an utterance.) It can be seen that the process was reasonably successful for these cases: in all cases, at least 60% of the errors could be detected by examining no more than 30 % of the segments. This is typical of the behaviour of phonemes that contained segmentation errors – the overall statistics are, however, impacted negatively by several phonemes for which no errors exist, for which candidates are nevertheless flagged.

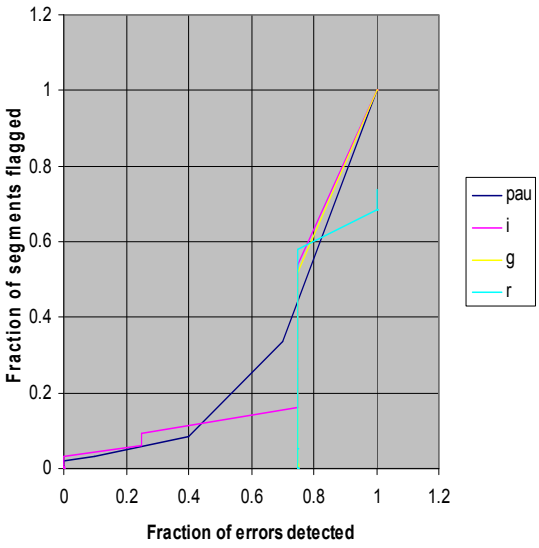
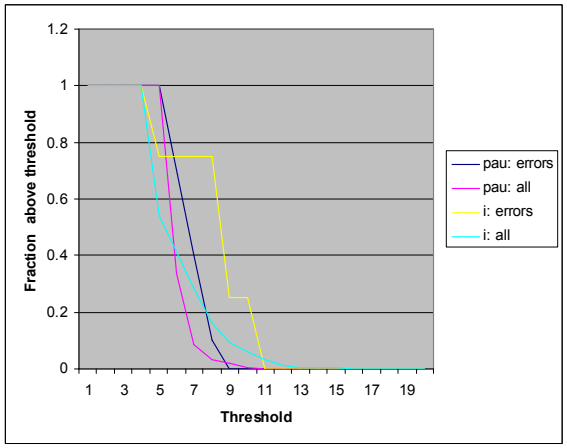
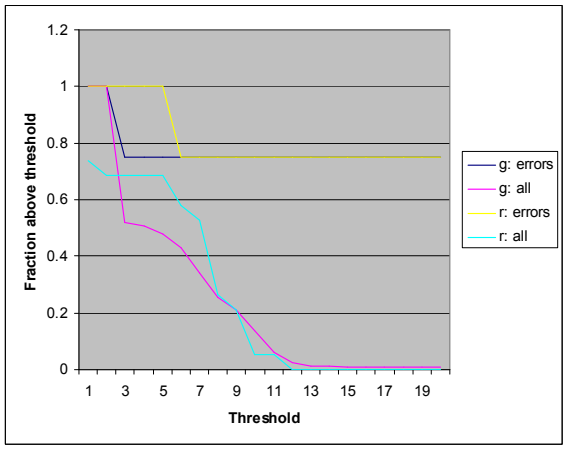


Figure 1: Detection-error trade-off curves for four different phonemes

For this process, manual thresholds were employed as discussed in Section 2. Fig. 2 shows why this was necessary: whereas any of a range of threshold values would be appropriate for the phonemes in Fig. 2(b), the phonemes in Fig. 2(a) require fairly specific (and different) threshold choices to successfully distinguish between correct and erroneous segments. Thus, manual selection of thresholds is currently required.



(a)



(b)

Figure 2: Fraction of all segments and erroneous segments flagged by the automatic process. In (a), the detection threshold must be chosen carefully, whereas any large value will be acceptable in (b)

In order to evaluate the performance of the detection process for different speech sounds, we divided the isiZulu phone set into seven broad categories, as shown in Table 1 (which also lists the number of samples of each sound contained in the corpus, as well as the number of errors made in such segments). The fraction of sounds in each category which were flagged by the

process, as well as the fraction of erroneous segments flagged, are listed in Table 2.

Class	*Example*	*Number of samples*	*Number of errors*
Clicks	**q**ala	71	1
Vowels	s**i**za	3105	52
Nasals	ha**m**ba	900	19
Fricatives/ affricates	**f**unda	914	20
Plosives	bo**ph**a	829	16
Silence / pause	<<pause>>	1917	27
Liquids/ glides	**l**anda	650	7

Table 1: Categories of isiZulu phones used for error analysis

Class	*Fraction of samples flagged*	*Fraction of errors flagged*
Clicks	0.25	1
Vowels	0.41	0.52
Nasals	0.20	0.30
Fricatives	0.09	0.60
Plosives	0.20	0.56
Silence / pause	0.14	0.30
Liquids/ glides	0.08	0.57

Table 2: Error detection statistics for different phone categories

These results indicate that the fricatives and liquids / glides are handled well by the automatic process, whereas the vowels and nasals are problematic. This contrast is probably a consequence of the features that were employed. For the more extended sounds such as vowels, the average spectrum is less reliable as an indicator of segmentation errors, both because of the relatively small effect of boundary errors on the average spectrum of a temporally extended phoneme and because of the significant intra-class variability that exists between different occurrences of such sounds.

4. Conclusions

We have shown that a simple segment-based approach can be used to detect a significant fraction of the errors that occur in both automatic and manual alignments of recordings used to construct a speech synthesizer. In practice, this approach is probably most useful when applied after one round of manual alignment has been performed, since the fraction of erroneous segments that remain undetected after automatic alignment is not satisfactory for system development. However, when larger corpora of recordings are used, this conclusion may no longer hold.

This work can be extended in a number of ways. It would be interesting to design more refined features – including spectral dynamics – to describe the individual segments, and also to develop more sophisticated error metrics – hopefully eliminating the need for manual threshold selection. The application of this approach to larger corpora of recordings, for which other automatic alignment strategies would be sensible, is also of interest.

5. Acknowledgements

Aby Louw was the main developer of the isiZulu synthesis system, and assisted in several aspects of this research.

6. References

[1] Dutoit, T.. *An introduction to text-to-speech synthesis.* Berlin:Springer. 1999

[2] Klatt, D. 1987. "Review of text-to-speech conversion for English", *Journal of the Acoustical Society of America*, 82:737–793.

[3] Hunt, A.J and Black, A.W. "Unit selection in a concatenative speech synthesis system using a large speech database", *Proceedings of ICASSP-96*, Vol. 1, pp. 373 – 376. Atlanta, Georgia: 373–376, 1996

[4] Clark, R.A J., Richmond, K. and King, S.. "Festival 2: build your own general purpose unit selection speech synthesiser". *Proceedings of the 5th ISCA Speech Synthesis Workshop*, Pittsburgh, PA:173-178. 2004

[5] Louw, J.A, Davel, M. And Barnard, E. "A general-purpose IsiZulu Speech Synthesiser", *South African Journal of African Languages*, To be published, 2006

[6] Taludkar, P "Optimal Text Selection Module Version 0.2", *LLSTI Project Report* 2004 Available at http://www.llsti.org/pubs/text_selection.pdf [Accessed on 16 October 2006]