INTERSPEECH 2009 BRIGHTON

Phonetic alignment for speech synthesis in under-resourced languages

D.R. van Niekerk, E. Barnard

Human Language Technologies Research Group, Meraka Institute, CSIR, Pretoria /
School of Electrical, Electronic and Computer Engineering,
North-West University, Potchefstroom, South Africa
dvniekerk@csir.co.za, ebarnard@csir.co.za

Abstract

The rapid development of concatenative speech synthesis systems in resource scarce languages requires an efficient and accurate solution with regard to automated phonetic alignment. However, in this context corpora are often minimally designed due to a lack of resources and expertise necessary for large scale development. Under these circumstances many techniques toward accurate segmentation are not feasible and it is unclear which approaches should be followed. In this paper we investigate this problem by evaluating alignment approaches and demonstrating how these approaches can be applied to limit manual interaction while achieving acceptable alignment accuracy with minimal ideal resources.

Index Terms: speech synthesis, phonetic speech segmentation, resource scarce language

10.21437/Interspeech.2009-266

1. Introduction

The widely spoken languages of the developed world have understandably received much attention from language technology specialists, including the development of large corpora of accurately annotated speech data. This enables the effective development of high quality, optimised corpus-based language systems such as Automatic Speech Recognition (ASR) and Text-to-Speech (TTS). When considering the automation of new corpus development for any of these languages, one has many resources to call upon in aid of such processes. However, when developing corpora and systems for new languages (especially those unrelated to well resourced languages, such as the languages of African origin), there are no analogous resources to build upon. Furthermore, skills shortages and limited economic viability of the lesser spoken languages hamper any prospects of developing large, high quality, manually constructed corpora. For these reasons, when considering the construction of systems such as TTS for these languages, speech corpora have often been minimally designed [1].

Systems such as that described in [1] demonstrate the possibility of successfully employing corpus-based synthesis techniques such as the unit-selection approach described in [2] with surprisingly small amounts of data achieving suitable levels of intelligibility. However, such development presents a number of challenges that make its widespread adoption unlikely. Such challenges involve corpus development with regards to both the recording and annotation processes and have a significant impact on the quality of resulting systems. One of these challenges, which we will consider in this paper, involves obtaining high quality phonetic alignments.

Two approaches have been successfully applied to text-dependent alignment when developing large TTS corpora:

1. Dynamic time warping (DTW), and
2. Hidden Markov Model-based (HMM) Viterbi forced-alignment.

The use of DTW to perform automatic phonetic alignment of a single speaker TTS corpus was first advocated by [3]. The idea is that an existing synthesiser is used to not only predict the pronunciation, but also synthesise the template signal which is aligned to the input signal. Alignment using Viterbi forced-alignment requires the training of HMMs, modelling each phoneme in the language individually. This has been attempted in a number of ways (including the application of speaker-independent models and speaker-independent models that have undergone speaker adaptation). However, an approach that has become common when developing TTS corpora simply involves training a speaker specific set of models on the same data to be segmented [2].

Although these techniques have been compared on a number of occasions [4, 5, 6], such comparisons have often been under ideal conditions without conclusions on practical issues such as aligning data in new languages and circumstances under which one or the other method might be preferred (e.g. considering data scarcity). Furthermore, results are not fully in agreement, with [6] concluding that HMMs outperform DTW conclusively and [4, 5] finding DTW more effective in fine alignment accuracy, including suggestions that alignments from a DTW stage be used to initialise the HMM based process [4].

With regard to achieving accurate alignments, a large amount of work has been done and a comprehensive study of general phonetic alignment with HMMs, including an extensive summary of approaches can be found in [7]. Other researchers have noted that local refinement techniques can be employed to improve alignments resulting from either HMM or DTW procedures [8, 9, 10]. The approaches proposed are, however, generally resource intensive and in some cases require extensive manual development.

In this paper we thus focus on evaluating the applicability of baseline text-dependent alignment techniques in the context presented and consider options toward applying these methods in a way which limits manual interaction and maximises alignment accuracy when aligning small corpora in new languages. In the following section we briefly describe our experimental setup, including the nature of corpora and manual alignment quality in the scenario presented. In Section 3 the application of baseline techniques is described and evaluated. Finally we consider options for improving the estimation of HMMs when aligning small prototypical corpora in Section 4, followed by a brief set of conclusions.

2. Experimental setup

Three sets of speech recordings used in the construction of prototypical TTS systems in South African languages are employed. These data sets represent minimally designed single speaker speech corpora, where text is selected carefully in order to cover all the appropriate phonetic constituents (diphones) of each language. The languages represented constitute three of South Africa's eleven official languages and come from distinct family groups (see Table 1 for specific details of each corpus).

Language	Group	Gender	Utts.	Dur.	Phones
Afrikaans	Germanic	Male	134	21 min.	12341
isiZulu	Nguni	Male	150	20 min.	8559
Setswana	Sotho	Female	332	46 min.	26010

Table 1: *Properties of the reference corpora.*

The corpora listed here were developed by manually correcting phonetic alignments based on baseline text-dependent techniques as these baseline techniques did not result in sufficiently accurate alignments to support intelligible concatenative synthesis systems. This work was largely performed by undergraduate students with limited experience presented with a short training session on correcting phonetic alignments.

We evaluate alignment approaches by comparing alignment results obtained on the reference corpora presented here. Two measures of comparison between automatic and reference labels are used: firstly, the traditional boundary accuracy (where boundaries falling within a certain threshold of the reference are considered as correct) and secondly the "overlap rate" (OR), which involves calculating to what degree segments overlap in time, in a duration-independent way [11]:

$$OR = \frac{D_{com}}{D_{max}} \quad (1)$$

$$= \frac{D_{com}}{D_{ref} + D_{auto} - D_{com}} \quad (2)$$

where D_{com}, D_{max}, D_{ref} and D_{auto} are the *common*, *maximum*, *reference* and *automatic* durations respectively.

Due to the limited level of experience involved in manual verification of the local corpora, it can be expected that the consistency and accuracy of the reference alignments be somewhat less ideal than generally encountered in cases where expert transcribers are employed. We thus determine the level of confidence in these alignments by independently manually aligning subsets of each corpus and calculating the level of agreement based on the measures of comparison used. These results can be found item 5 of Table 2.

Boundary comparison results can be directly compared with similar results from other studies mentioned above, as expected, the inter-transcriber discrepancies are somewhat higher than cases reported for expert transcribers [6, 7]. This is especially the case in the lower tolerance ranges (e.g. $<$ 5ms where figures close to 70% should be possible).

3. Baseline alignment approaches

In this section we evaluate the two baseline text-dependent segmentation approaches mentioned in Section 1 in this context and characterise typical difficulties and the relationship between accuracy and corpus properties.

These two approaches are based on similar dynamic programming algorithms, with the difference being the reference representation which is either a relevant synthetic speech signal, or a model describing the phonetic sequence. Thus an important factor determining the accuracy of these techniques involves the construction of these reference representations or templates.

When developing small prototypical corpora in new languages, one has neither significant amounts of speech data in the target language (for the effective training of HMMs) nor an appropriate synthesiser (possibly not even in a closely related language) for the application of DTW. Thus the following concerns exist with regard to alignment using these techniques in this context:

- The implications of using a widely available English synthesiser to synthesise template signals for different languages on DTW performance,
- The implications of training and applying HMMs from minimally designed corpora where some phone occurrences are extremely limited, and
- The general suitability of these approaches with respect to accuracy, robustness and practical implementability in the given context.

We experimented here by employing the DTW procedure implemented in the *Festvox* toolkit [12] and a generic HMM training procedure using *HTK* described in [13]. In both cases we used the suggested default parameters with regard to feature extraction, that is 24-dimensional MFCCs (Mel Frequency Cepstral Coefficients) including base and delta coefficients extracted every 5ms using 25ms windows for the DTW process and 39-dimensional MFCCs, including base, delta and delta delta coefficients extracted every 10ms using 20ms windows for the HMM procedure. For the HMM-based process, three-state left-to-right models with a single Gaussian mixture per state are initialised using the flat start approach.

In Figure 1 a plot of the boundary accuracy values for each system compared to the reference segments are presented over a range of thresholds. From the information present in this figure one can assess the relative ability of each system to place phonetic boundaries within a small region around the reference boundaries (i.e. fine placements or accuracy) as well as an estimate of the nature of large discrepancies between automatic and reference alignments (i.e. gross errors). It is clear in this case that the alignments resulting from the HMM-based procedure are consistently closer to the reference alignments, meaning that this system results in both more accurate fine placements and fewer gross misplacements compared to the DTW technique.

Since the performance of HMM-based alignment is dependent on the size of the corpus while this is not the case for DTW, an experiment was performed where utterances are segmented with both methods for subsets of each corpus ranging from only one utterance to the full size. For each outcome the mean OR is calculated. Figure 2 shows these values for each data set size. While the stability of these results are dependent on the phone distributions (e.g. in the case of isiZulu there are more phones with few occurrences), it is surprising to see that the HMM-based procedure performs consistently better on average than the DTW procedure with as few as 20 utterances to be aligned.

4. HMM estimation for alignment

Motivated by the outcome of the experiments described in the previous section, we investigated efficient methods towards improving the accuracy and consistency of HMM-based alignments (baseline results can be found under item 1 of Table 2).

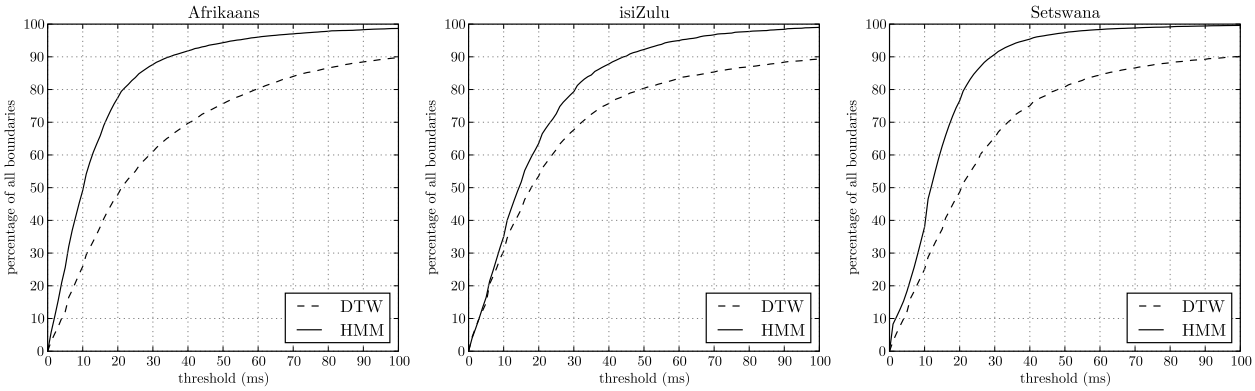

Figure 1: *A comparison of boundaries in agreement with the reference sets for a range of thresholds.*

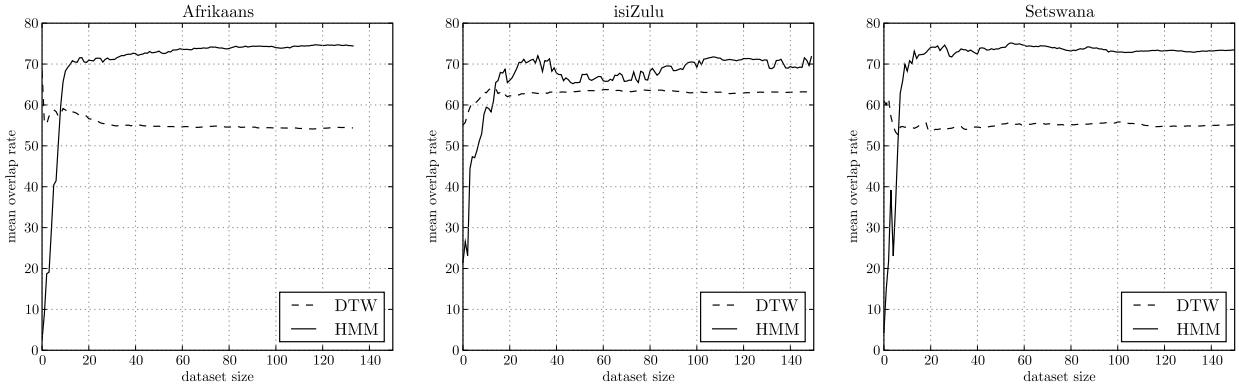

Figure 2: *Mean OR with the reference sets for data set sizes ranging from 1 to 150 utterances.*

At this point we chose to split plosives and affricates into closure and burst segments as these are generally distinguished in labeled concatenative TTS corpora.

Closer inspection of the baseline alignment results (specifically the overlap rates achieved by each phone category - e.g. plosives, vowels and nasals) showed that shorter segments exhibited significantly lower overlap rates [14]. One possible reason for this is convergence problems during training based on the flat start procedure where initial model parameters are dominated by frames belonging to longer segments.

In order to investigate the possibility of bootstrapping the training process, we selected minimal subsets of utterances from each of the corpora, with each of the selected subsets containing at least three occurrences of each phone in the specific language (this amounted to between 10 and 20 utterances for each language). These utterances were removed from the main corpus and used during the initialisation process only. This involved using the manually aligned input transcriptions in order to perform an iterative Viterbi alignment as implemented in *HTK*. The subsequent embedded re-estimation process and alignment was completed on the remaining utterances as before. The results are contained in item 2 of Table 2. Although this approach resulted in a significant increase in alignment accuracy, the necessity of manually aligning subsets of data does increase development time significantly.

We thus experimented with bootstrapping the training process by using the well known DARPA TIMIT [15] corpus. For these experiments the same model initialisation process described above was used, however, each phone model was initialised by mapping all phones in the TIMIT corpus to their broad phonetic categories before bootstrapping, i.e. each phone belonging to a specific broad category is initialised with a model identical to all other phones in the same category, based on the acoustic properties of all phones in the same category in the bootstrap set. The only form of normalisation of the acoustic features used during this experiment is Cepstral Mean Normalisation (CMN), which was indeed used during all the experiments thus far. Item 3 in Table 2 suggests that this is comparable to bootstrapping with minimal data with the added advantage of being more cost-effective.

Comparing the boundary accuracies between items 3 and 5 in Table 2, one can see that the HMM-based alignment accuracy is particularly lacking for lower tolerances. Further investigation revealed that a significant number of segments had durations less than the minimum duration imposed by the three-state HMMs. At this stage we investigated more appropriate system parameters in order to make the system more suitable to the goal of alignment. We considered model topology, feature extraction rate, context dependence of models and state distribution complexity. We considered the general effect of these parameters on alignment accuracy on the above-mentioned corpora (i.e. without tuning of parameters on individual corpora) by comparing to manual alignments. The conclusion is that the choice of five-state left-to-right context-dependent (triphone) models using a

single Gaussian mixture per state with 39-dimensional MFCCs extracted every 5ms for 10ms windows yields good results in this context (see item 4 in Table 2). Experimenting with distinct models containing fewer states for modeling shorter segments such as closures and plosives proved unsuccessful.

Language	Boundary comparisons			OR	
	< 5ms	< 10ms	< 20ms	μ	σ
1. HMM baseline					
Afrikaans	20.22%	39.25%	63.81%	62.17%	24.18
isiZulu	16.71%	35.56%	62.14%	66.26%	21.80
Setswana	14.83%	30.55%	62.29%	59.78%	25.08
2. HMM with bootstrapped models					
Afrikaans	27.51%	55.27%	83.28%	71.98%	19.65
isiZulu	31.12%	57.19%	83.86%	77.33%	18.18
Setswana	31.58%	57.46%	87.15%	75.62%	16.56
3. HMM with cross-language models					
Afrikaans	29.95%	59.57%	84.94%	73.38%	19.38
isiZulu	29.49%	55.58%	82.57%	76.32%	18.90
Setswana	28.52%	54.78%	86.87%	74.61%	16.69
4. HMM with bootstrapped models and improved parameters					
Afrikaans	45.13%	69.01%	86.91%	77.15%	19.15
isiZulu	42.57%	66.24%	86.51%	80.40%	17.34
Setswana	45.84%	71.52%	88.81%	79.34%	17.03
5. Inter-transcriber agreement					
Afrikaans	54.58%	73.35%	88.84%	79.41%	18.90
isiZulu	49.33%	74.35%	89.49%	81.16%	17.82
Setswana	58.05%	77.85%	90.64%	82.18%	16.54

Table 2: *Alignment results achieved through the approaches described here, compared to baseline flat started alignments and inter-transcriber agreement.*

5. Conclusions

Based on the investigation and results presented here, we make the following conclusions:

- Baseline HMM-based segmentation is more convenient, robust and accurate than DTW when aligning data in new languages for all practical scenarios, however, neither of the baseline systems result in alignment accuracy acceptable for system development.
- Practical model initialisation options exist in this context, allowing significant improvements in alignment accuracy without extensive resources or manual interaction.
- An HMM-based approach should be tuned to the task of alignment, with the selection of appropriate parameters involving a trade-off between modeling precision and system output resolution (affected by the selection of model topology and feature extraction properties).
- The quality of manual alignments in the context described are significantly lower than agreement levels between expert transcribers. As a result, the accuracy and consistency of automated alignments obtained here are comparable to manual alignment results and might prove to be more appropriate for system building, considering the conclusions found in [16].

This work demonstrates how the problem of automated phonetic alignment can be approached for the rapid development of TTS corpora for under-resourced languages and should also be applicable when considering efficient development of larger corpora. The use of higher resolution HMM parameters here is especially important due to the fact that plosives and affricates are subdivided here. The effect of this on synthesis quality is still to be determined.

Our initial experiences with a concatenative TTS system for Afrikaans built using the techniques described here were very positive: highly intelligible synthesis was achieved with minimal manual intervention. We are in the process of repeating the process with several other languages, and intend to do formal perceptual evaluations of those systems.

6. References

[1] J. A. Louw, M. Davel, and E. Barnard, "A general-purpose IsiZulu speech synthesizer," *South African journal of African languages*, vol. 2, pp. 1–9, 2006.

[2] 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.

[3] F. Malfrère and T. Dutoit, "High-quality Speech Synthesis for Phonetic Segmentation," in *EUROSPEECH*, Rhodes, Greece, 1997, pp. 2631–2634.

[4] F. Malfrère, O. Deroo, T. Dutoit, and C. Ris, "Phonetic alignment: speech synthesis-based vs. viterbi-based," *Speech Communication*, vol. 40, no. 4, pp. 503–515, 2003.

[5] J. Kominek, C.L. Bennett, and A.W. Black, "Evaluating and Correcting Phoneme Segmentation for Unit Selection Synthesis," in *EUROSPEECH*, Geneva, Switzerland, September 2003, pp. 313–316.

[6] J. Adell, A. Bonafonte, L.A. Hernndez Gmez, and M. J. Castro, "Comparative study of Automatic Phone Segmentation methods for TTS," in *Proceedings of ICASSP*, Philadelphia, Pennsylvania, USA, 2005, vol. 1, pp. 309–312.

[7] D.T. Toledano, L.A. Hernández Gómez, and L.V. Grande, "Automatic Phonetic Segmentation," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 617–625, 2003.

[8] A. Sethy and S. Narayanan, "Refined Speech Segmentation for Concatenative Speech Synthesis," in *Proceedings of ICSLP*, Denver, Colorado, USA, September 2002, pp. 149–152.

[9] Y. Kim and A. Conkie, "Automatic Segmentation Combining an HMM-Based Approach and Spectral Boundary Correction," in *Proceedings of ICSLP*, Denver, Colorado, USA, September 2002, pp. 145–148.

[10] S.S. Park, J.W. Shin, and N.S. Kim, "Automatic speech segmentation with multiple statistical models," in *INTERSPEECH*, Pittsburgh, Pennsylvania, USA, September 2006, pp. 2066–2069.

[11] S. Paulo and L.C. Oliveira, *Advances in Natural Language Processing*, Springer Berlin / Heidelberg, 2004.

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

[13] 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)*, Cambridge University Engineering Department, http://htk.eng.cam.ac.uk/, 2005.

[14] D.R. van Niekerk and E. Barnard, "Important factors in HMM-based phonetic segmentation," in *Proceedings of PRASA*, Pietermaritzburg, South Africa, November 2007, pp. 25–30.

[15] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, *Darpa Timit: Acoustic-phonetic Continuous Speech Corps CD-ROM*, US Dept. of Commerce, National Institute of Standards and Technology, 1993.

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