

A perceptual evaluation of corpus-based speech synthesis techniques in under-resourced environments

D.R. van Niekerk, E. Barnard & G. Schlünz

Human Language Technologies Research Group
Meraka Institute, CSIR, Pretoria, South Africa

dvniekerk@csir.co.za, ebarnard@csir.co.za, gschlunz@csir.co.za

Abstract

With the increasing prominence and maturity of corpus-based techniques for speech synthesis, the process of system development has in some ways been simplified considerably. However, the dependence on sufficient amounts of relevant speech data of high quality remains a central challenge in under-resourced environments. In this paper we investigate the quality implications when building baseline synthesis systems with reduced amounts of speech data. This is done through a perceptual evaluation of synthesis systems based on unit-selection and statistical parametric synthesis techniques. We show that – although it is possible to build an acceptable unit-selection synthesizer with as little as 27 minutes of carefully recorded speech data – synthesis quality obtainable from Hidden Markov Model-based synthesis is more consistent and requires significantly less speech data.

1. Introduction

By employing corpus-based techniques towards speech synthesis, the development of basic synthesizers for new languages has become more feasible. These techniques allow engineers and speech technologists to rely on speech data to render more intelligible and natural speech than before, whilst requiring less linguistic knowledge and language-specific processing. Although this has greatly aided in the development of systems for under-resourced languages (by reducing the number of language-specific technical challenges and linguistic resources required to develop new systems), the increased reliance on data quality and quantity still present significant challenges. This represents a significant obstacle when developing systems in under-resourced languages. This is true for a number of reasons including: a lack of skills (of technical nature) and basic resources (e.g. text corpora) which results in the construction of large corpora being either a costly or entirely infeasible endeavour.

Two corpus-based synthesis approaches have in the past been successfully employed towards developing systems in under-resourced contexts, namely unit-selection [1] and statistical parametric synthesis based on Hidden Markov Models (HMMs) [2].

The advantage of using the unit-selection approach is that it makes use of the data directly by building an acoustic inventory and resynthesising unique utterances by selecting and concatenating these unit samples directly without signal modification. This results in speech quality which resembles the original audio quality of the recordings, having a highly natural quality due to the absence of signal distortion, if units from appropriate linguistic contexts are available in the acoustic inventory. Although this implies a high level of sensitivity on data quality

and reliance on large quantities of speech data to cover relatively sparse linguistic contexts, it has been demonstrated that acceptable synthesis systems can be constructed with relatively small speech corpora if recording conditions are carefully controlled to limit the amount of variation in the speech corpus [3].

Alternatively, by constructing acoustic models from the data, rather than using acoustic samples directly, a level of robustness against extreme variations in the data can be achieved by the averaging process of maximum likelihood model estimation, while context-specific models ensure that unique features of acoustic units in different linguistic contexts are captured if enough training data exists. The main disadvantage of this approach is that resynthesising speech from these models (usually achieved by means of a vocoder) almost always results in speech with a slightly synthetic or unnatural quality (e.g. “buzziness”), even when a large amount of speech data is available to estimate models.

Considering the varying properties of these synthesis techniques and the fact that the development of large amounts of quality speech data is difficult and expensive, an interesting and important question involves the perceived quality differences of these synthesis techniques when speech data is limited. If questions such as the general suitability and data dependence of these techniques in the context and the minimum amount of data required to build robust synthesizers can be answered, this could have a significant impact on the design of synthesis systems for under-resourced languages.

In this paper, we thus construct a number of baseline synthesis systems using these two techniques with varying amounts of speech data and report on perceptual experiments comparing synthesis examples. These experiments are based on simple questions of preference where we attempt to establish the relative quality of the synthesis samples. We also employ the Dynamic Time Warping (DTW) algorithm to compare synthesis samples with natural speech instances to determine the utility of such a process for the purpose of measuring relative synthesis quality.

The following section describes the details regarding the experimental setup. This is followed by a section presenting the results for each experiment, a brief discussion section and a section with conclusions.

2. Experimental setup

In order to perform perceptual experiments as mentioned, we construct complete synthesis systems in two South African languages, namely South African English and Afrikaans. In the following two sections we describe the design and properties of the corpora that were used and the details regarding the im-

plementations of our synthesis system. This is followed by a section describing the nature of perceptual experiments and the implementation of the DTW-based quality measure.

2.1. Speech corpora

Two small corpora, designed to ensure phonetic coverage and carefully recorded to limit excessive variation in voice conditions (including intonation) were employed to develop synthesis systems (see Table 1).

Language	Gender	Utterances	Duration	Phonemes
Afrikaans	Female	522	33 mins.	22218
English	Male	447	31 mins.	18564

Table 1: Complete corpora properties.

For the purpose of the perceptual experiments, we partitioned the corpora into subsets of training data by removing some test utterances (for the application of DTW), leaving 400 utterances for each language from which we defined 4 sets of data containing 100, 200, 300 and 400 utterances respectively. These selections were made to ensure phonetic coverage in each case. The properties of each of these sets are summarized in Table 2.

Subset	Utterances	Duration	Phonemes
Afr100	100	7 mins.	4656
Afr200	200	14 mins.	9049
Afr300	300	21 mins.	13647
Afr400	400	27 mins.	17833
Eng100	100	7 mins.	4317
Eng200	200	13 mins.	8544
Eng300	300	20 mins.	12968
Eng400	400	26 mins.	16991

Table 2: Partitioned corpora properties.

2.2. System development

For both languages, text analysis and basic linguistic modules were developed (e.g. phoneme sets, grapheme-to-phoneme rules and syllabification routines) for the Speect speech synthesis framework [4], which is based on the architecture of the Festival Speech Synthesis System [5]. Thus, the synthesis backends both rely on an identical text analysis process.

Phonetic alignments were performed automatically using an HMM-based forced-alignment process bootstrapped with data from the TIMIT corpus [6] as described in [7].

In the following two sub-sections details regarding systems using the different techniques are briefly described.

2.2.1. Unit-selection synthesis

The process of constructing a unit-selection voice in Speect closely follows the method implemented in the Festvox software package [8], extracting Linear Predictive Coefficients (LPCs) and residuals for residual excited LPC resynthesis and Mel Frequency Cepstral Coefficient (MFCC) vectors for the calculation of join costs. The unit-selection algorithm implementation closely follows the Multisyn implementation [1].

One enhancement implemented in addition to the standard Multisyn approach (which is especially helpful when synthesising from small corpora) is the fact that the constructed acoustic

database consists of halfphone units instead of diphones and employs a hierarchical selection process where halfphones in context (essentially diphones and larger units) are considered first before relying on smaller units if larger units do not exist in sufficient quantities.

2.2.2. Statistical parametric synthesis

Training of HMM models was done via the standard demonstration script provided for the HTS synthesis engine [9]. Phonetically aligned utterance files were obtained from a text-analysis front-end, using the alignment procedure described above.

For the model tying decision tree, questions relating to sentence and phrase context were based on the English demo while further questions were generated based on phonetic categories defined in the phone set (e.g. categories such as plosives, nasals and vowels and voicing etc.). The only other customisation to the training procedure involved changing the frequency ranges for pitch extraction of the male voice.

2.3. Perceptual experiments

Three distinct experiments were defined for each of the languages:

1. A comparison of synthesis samples by unit-selection with varying amounts of data.
2. A comparison of synthesis samples by HMM-based synthesis with varying amounts of data.
3. A comparison between synthesis samples of HMM-based and unit-selection synthesis.

The number of respondents who took part in the experiments was 10 per language. The method of delivery was a website which adhered to the following protocol for each experiment:

- The sample comparisons were randomly ordered in pairs which covered all the possible combinations of different amounts of data.
- Each pair of samples, or particular combination, was synthesised from the same sentence for 5 different sentences from the test set of utterances.
- For each pair, the transcription of the sentence was displayed to the respondent.
- The respondent would listen to both samples in the pair.
- The respondent would then choose one of the samples or indicate no preference, according to the open-ended question “Which sample do you prefer?”.

The preference criteria was made non-specific intentionally to limit the time required and complexity of the evaluation process. This simplifies the task of the listener, making it more manageable when listening to a large number of comparisons. For similar reasons, the number of comparisons were reduced in the case of the unit-selection comparison, as a number of samples from the 200- and 300-utterance unit selection databases are identical since the system often selects units from the same base of 200 utterances. On the other hand, the HMM-based approach averages over the number of utterances in its parameter estimation, so differences might be perceivable, even between the 200- and 300-utterance syntheses.

2.4. Dynamic Time Warping

In addition to having human respondents rate the test utterances, we also calculate a similarity metric based on the Dynamic Time Warping algorithm (see Algorithm 1).

<p>Data: A synthetic and natural speech example of a matching utterance.</p> <p>Result: The mean frame distance over the best path found.</p> <p>Initialise the accumulated distance matrix $D(1, 1) = d(1, 1)$; Initialise the most likely path $B(1, 1) = 1$;</p> <p>for $i \leftarrow 2$ to N do</p> <table border="1" style="margin-left: 20px;"> <tr> <td>for $j \leftarrow 1$ to M do</td> <td>$D(i, j) = \min_{1 \leq p \leq M} D(i-1, p) + d(p, j)$;</td> </tr> <tr> <td></td> <td>$B(i, j) = \operatorname{argmin}_{1 \leq p \leq M} D(i-1, p) + d(p, j)$;</td> </tr> </table> <p>end</p> <p>end</p> <p>Backtrack over B to find the most likely path through the matrix;</p> <p>Calculate the mean frame distance over the path found;</p>	for $j \leftarrow 1$ to M do	$D(i, j) = \min_{1 \leq p \leq M} D(i-1, p) + d(p, j)$;		$B(i, j) = \operatorname{argmin}_{1 \leq p \leq M} D(i-1, p) + d(p, j)$;
for $j \leftarrow 1$ to M do	$D(i, j) = \min_{1 \leq p \leq M} D(i-1, p) + d(p, j)$;			
	$B(i, j) = \operatorname{argmin}_{1 \leq p \leq M} D(i-1, p) + d(p, j)$;			

Algorithm 1: Dynamic Time Warping [10]

The implementation of this procedure is based on the DTW phonetic alignment procedure implemented in the Festvox software package, using the feature extraction parameters commonly used in this context, namely MFCCs with 12 coefficients and their first order derivatives (24 coefficients) calculated from Hamming windowed frames with length 25ms and 5ms frame shift. The frames are compared by calculating the Euclidean distance between frames of feature vectors extracted for both signals.

3. Results

3.1. Perceptual experiments

We present the outcomes of the perceptual evaluations in this section by simply counting all the votes for a particular instance in each comparison category over all respondents. In each comparison category, an overall level of equivalence is also quantified, calculated as follows:

$$\text{equivalence} = 1 - \frac{|A - B|}{A + B + E} \quad (1)$$

where A , B and E represent the number of votes for “sample A”, “sample B” and “no preference” respectively.

For experiment 1, where we compared samples by the unit-selection approach with different amounts of data, we obtain the results for Afrikaans and English presented in Table 3.

These results indicate in each case a very distinct preference for the synthesizer built with more data.

Experiment 2 is of similar nature to experiment 1; however, in this case we compare HMM-based synthesizers (results presented in Table 4).

It is evident from these results that a much larger percentage of votes identify these samples to be equivalent (compared to the unit-selection samples) and that on average the difference between votes for each comparison is less than in the case of unit-selection.

In the final perceptual experiment, we compared the different techniques at the lower and upper extremes of data quantities (see Table 5).

Comparison	A	B	E	Equivalence
Afr100 vs. Afr200	9	32	9	0.54
Afr100 vs. Afr400	0	50	0	0.00
Afr200 vs. Afr400	6	34	10	0.44
Eng100 vs. Eng200	10	35	5	0.50
Eng100 vs. Eng400	2	48	0	0.08
Eng200 vs. Eng400	0	32	18	0.36
100 vs. 200	19	67	14	0.52
100 vs. 400	2	98	0	0.04
200 vs. 400	6	66	28	0.40

Table 3: Experiment 1: Unit-selection data

Comparison	A	B	E	Equivalence
Afr100 vs. Afr200	6	22	22	0.68
Afr100 vs. Afr300	7	23	20	0.68
Afr100 vs. Afr400	2	32	16	0.40
Afr200 vs. Afr300	5	17	28	0.76
Afr200 vs. Afr400	6	18	26	0.76
Afr300 vs. Afr400	9	10	31	0.98
Eng100 vs. Eng200	10	16	24	0.88
Eng100 vs. Eng300	7	17	26	0.80
Eng100 vs. Eng400	4	21	25	0.66
Eng200 vs. Eng300	11	20	19	0.82
Eng200 vs. Eng400	7	21	22	0.72
Eng300 vs. Eng400	8	12	30	0.92
100 vs. 200	16	38	46	0.84
100 vs. 300	14	40	46	0.74
100 vs. 400	6	53	41	0.53
200 vs. 300	16	37	47	0.79
200 vs. 400	13	39	48	0.74
300 vs. 400	17	22	61	0.95

Table 4: Experiment 2: HMM-based data

This shows that when minimal data is used, HMM-based voices are clearly preferable, however at 400 utterances (26-27 minutes of speech data) the unit-selection approach starts becoming somewhat more popular, especially for the English voice.

3.2. Dynamic Time Warping

We also compared each of the synthesised test sentences with natural speech instances via the DTW algorithm, determining the average frame distances.

In Tables 6 and 7 we present the results representing the average frame distance at each distinct data quantity level over all utterances per language for each technique respectively.

Comparison	A	B	E	Equivalence
HTS-Afr100 vs. US-Afr100	42	6	2	0.28
HTS-Afr400 vs. US-Afr400	26	24	0	0.96
HTS-Eng100 vs. US-Eng100	38	11	1	0.46
HTS-Eng400 vs. US-Eng400	19	30	1	0.78
HTS-100 vs. US-100	80	17	3	0.37
HTS-400 vs. US-400	45	54	1	0.91

Table 5: Experiment 3: Unit-selection (US) versus HMM-based (HTS) synthesis

Language	100	200	400
Afrikaans	4.11524	3.79383	3.55794
English	3.57096	3.88060	3.82922

Table 6: DTW average frame distances (US)

Language	100	200	300	400
Afrikaans	3.41558	3.24804	3.16199	3.09201
English	3.06342	3.00894	2.96880	2.94408

Table 7: DTW average frame distances (HTS)

These results show that by comparing even a few utterances with instances of natural speech (and only one natural sentence per utterance), one can detect an increase an improvement in synthesis quality indicated by a reduction in the mean frame distance for samples synthesised with more speech data. This pattern seems quite consistent in the case of HMM-based sample comparisons. In the case of the unit-selection samples, the trend was somewhat less consistent, with the English voice using 100 sentences displaying an unexpectedly low average frame distance.

4. Discussion

In this paper we investigated the perceptual consequences when constructing corpus-based speech synthesizers with minimal amounts of data, and evaluated the use of DTW as a measure of relative synthesis quality. For the presentation of perceptual test results, we found it meaningful to combine all respondents' votes into a single pool of votes as no clear bias by any individual respondent for samples by a specific technique was observed.

In the following sections we briefly discuss the results presented in the previous sections.

4.1. Experiment 1

When comparing samples by the unit-selection approach, the preferences noted at the various data levels follow a distinct trend. Synthesised speech quality clearly drops when the data is reduced, despite the fact that the speech corpora were carefully recorded to minimise variation to which this technique is sensitive. The clearly distinguishable improvements with every increase in the number of utterances suggest that synthesis quality has not yet stabilised—more data would still be beneficial. Thus it seems that we are at the lower threshold of data requirements for this technique to achieve acceptable quality synthesis.

4.2. Experiment 2

The comparison of samples by the HMM-based approach stands in contrast. Although there is a steady improvement in the synthesis quality of samples synthesised from more data, as indicated by the consistent preference bias, there are also much larger percentages of instances where the samples are judged to be of similar quality. Therefore, the HMM-based samples are less differentiable among one another than their unit selection counterparts, even between samples with large differences in the number of training utterances employed. This was confirmed by a number of the respondents who commented that their ear grew accustomed to the HMM-based syntheses over time, to the point where they could not distinguish the qual-

ity degradation between samples anymore. It suggests that this degradation associated with limiting training data might be acceptable when resources are truly scarce.

4.3. Experiment 3

The comparison between the two synthesis methods shows that HTS is clearly preferable when data is severely limited. When more data is available, unit-selection is more competitive and at 400 utterances (26-27 minutes), it is slightly preferred over the HTS samples on average. However, the observation that even at this level HMM-based synthesis is preferred in the case of Afrikaans, emphasizes the fact that synthesis quality is more variable (sensitive to recording conditions) in the case of unit-selection. This result also seems to agree with an earlier observation we made when building unit-selection systems, namely that male voices are more likely to lead to acceptable quality synthesizers when data is highly limited.

4.4. DTW

Using DTW to obtain a measure of synthesis quality by comparing samples with a test set of natural speech samples from the same speaker seems to provide ratings which correlate with perceived quality (especially in the case of the HTS samples). However, we found a greater amount of variance in the mean frame distance when comparing unit-selection samples. It is possible that the reliability of this measure is adversely affected when gross errors occur during synthesis (causing misalignment of the path for comparison or resulting in large frame distances), which occurred occasionally especially in the case of unit-selection where bad joins were present when data was limited. It is nevertheless possible that one might be able to obtain useful ratings if a larger number of test samples are compared, for instance, by a K-fold cross-validation procedure. Such an approach could also be combined with an analysis of the path found, in order to rule out misalignments as a source of noise in the measurements.

5. Conclusion

In the given scenario where speech data is limited, it is generally agreed that it is difficult to construct a good quality unit-selection synthesizer. Therefore, the speech corpora used here were very carefully recorded to minimise unnecessary variation, resulting in a rather monotonous speech quality. Under these conditions unit-selection only really becomes feasible at around 400 utterances (26-27 minutes). In contrast, we have shown that it is possible to use corpora of this nature to construct acceptable HMM-based synthesis systems with as few as 100 utterances (7 minutes) of data. Given these results, it might even be possible to build prosodically more natural voices with 400 utterances using HTS, by relaxing the recording constraints on variation slightly.

The advantage of using these corpus-based techniques for synthesis compared to rule-based approaches (such as formant synthesis) is that prosodic variation can be learned implicitly from the speech corpus itself, reducing the need for explicit definition. However, when data is limited, this might not be feasible. The use of statistical parametric synthesis techniques seems to be a prudent choice in this context, clearly outperforming unit-selection at the lower end of data availability. In addition, such techniques may benefit from their modelling flexibility if more sophisticated prosodic and other models become available, thus eliminating the dependency on large amounts of

data to cover many linguistic contexts. Finally, the statistical techniques are also likely to be important in multilingual environments, where the same flexibility may support (for example) code switching by combining speech from speakers of different languages.

6. References

- [1] 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.
- [2] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, Nov. 2009.
- [3] 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.
- [4] J.A. Louw, "Speect: a multilingual text-to-speech system," in *Proceedings of PRASA*, Cape Town, South Africa, November 2008, pp. 165–168.
- [5] P. Taylor, A. W. Black, and R. Caley, "The Architecture of the Festival Speech Synthesis System," in *The Third ESCA/COCOSDA Workshop (ETRW) on Speech Synthesis*, Jenolan Caves, Blue Mountains, NSW, Australia, 1998, pp. 147–151, ISCA.
- [6] 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 Corpus CD-ROM*, US Dept. of Commerce, National Institute of Standards and Technology, 1993.
- [7] D.R. van Niekerk and E. Barnard, "Phonetic alignment for speech synthesis in under-resourced languages," in *INTERSPEECH*, Brighton, UK, 2009, pp. 880–883.
- [8] A. W. Black and K. A. Lenzo, *Building Synthetic Voices*, <http://www.festvox.org/bsv/>, 2007.
- [9] H. Zen et al, *HTS speaker dependent training demo*, http://hts.sp.nitech.ac.jp/archives/2.1/HTS-demo_CMU-ARCTIC-SLT.tar.bz2.
- [10] A. Acero X. Huang and H. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm and System Development*, Prentice Hall, 2001.