A Parametric Monophone Speech Synthesis System

G. Klompje, T.R. Niesler

Digital Signal Processing Group, Department of Electrical and Electronic Engineering
University of Stellenbosch, Stellenbosch, South Africa
[gklompje, trn]@dsp.sun.ac.za

Abstract

Current TTS systems generally require large annotated speech corpora in the languages for which they are developed. For many languages these resources are not available. In their absence, a TTS system must generate synthetic speech by means of mathematical algorithms constrained by certain rules.

This paper describes a rule-based speech generation algorithm for use in a TTS system. The system allows phonetic and prosodic content as well as other parameters associated with a sound and its particular mode of articulation to be specified. Linear predictive (LP) models of monophone speech units are used, greatly reducing the amount of data required for development in a new language. A novel approach to the interpolation of monophone speech units is presented to allow realistic transitions between monophone units. Additionally, novel algorithms for estimation and modelling the harmonic and stochastic components of an excitation signal are developed.

Promising first results were obtained when evaluating the developed system's South African English speech output intelligibility using the modified rhyme test (MRT) and semantically unpredictable sentences (SUS).

1. Introduction

This paper describes the development of a flexible speech generation system that is not restricted to any specific language. Portability and extendability were considered important in order to allow the adapting to different devices. Language-independence is particularly important in a country such as South Africa, which has eleven official languages. Although project restrictions did not allow testing in multiple languages, the system is designed to synthesise speech in any target language if given a suitable phonetic description.

Concatenative TTS systems are currently state-of-the-art, but have certain inherent disadvantages due to their dependency on an associated speech corpus [1]. These include language, accent and pronunciation dependencies as well as the possibility that the data may not contain all synthesis units in all the desired phonetic contexts. Recording and annotating a speech database for use in a concatenative TTS system is also a time-consuming and laborious process. For these reasons it was decided that a synthesis system based on a speech production model may be better suited to the problem.

Monophones were chosen as the basic synthesis units, and linear prediction (LP) for speech modelling. In this way, each synthesis unit could be represented by a single LPC vector. This is in contrast to the use of diphones/triphones, of which there is a much larger number per language and modelling requires multiple parameter vectors for each unit. The use of monophones and parametric modelling also eases the adaptation of the system to a new language greatly, since the number of monophones

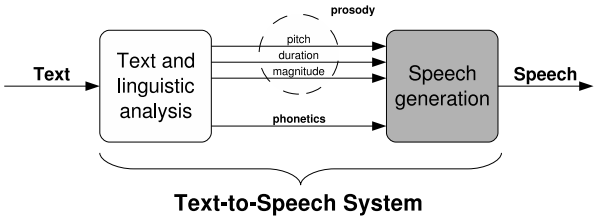


Figure 1: *The components of a TTS system.*

is always much smaller than the number of diphones or triphones, and many languages do not have comprehensive annotated databases from which the latter synthesis units can be extracted. Using monophones, however, requires us to model the inter-phone transitions which are otherwise implicitly modelled in the recorded units. It also requires a flexible parametric excitation signal model suited to such interpolation.

Within our system, each monophone is represented by a set of 30 LPC's and 3 excitation signal parameters, all of which are estimated automatically from recordings of individual monophones. For synthesis, the system uses the source-filter model of speech production such that an excitation signal is generated from the appropriate parameters and filtered by their corresponding LPC's. An utterance, specified as a sequence of monophones, is synthesised by generating smooth parameter trajectories between consecutive monophone vectors by interpolation.

The current system consists of only a speech generation module (shaded in Fig. 1). There is currently no front-end providing text and linguistic analysis. Hence no attempt is made to generate prosodic and other higher-level information from the text. Instead, this is assumed known. The system allows multiple prosodic contours to be overlaid within an utterance to ensure maximum flexibility [2]. Not all of the advantages of this approach are exploited in the current implementation, but remain the subject of future work.

2. Excitation Modelling

Various models that can be used for the excitation signal within a source-filter framework were considered for the system. Among these are Gaussian noise for the unvoiced component as well as impulse trains and *Rosenberg-Klatt* (RK, from [3]) waveforms for the voiced component. However, these models do not allow explicit control over the number of harmonics in the excitation or their frequency-domain envelope. For this reason, a mixed excitation model incorporating a sinusoidal voiced component and a Gaussian unvoiced component was developed.

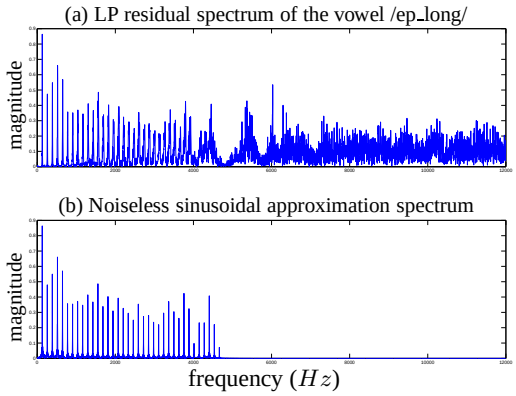


Figure 2: *Spectrum and sinusoidal model of a voiced residual.*

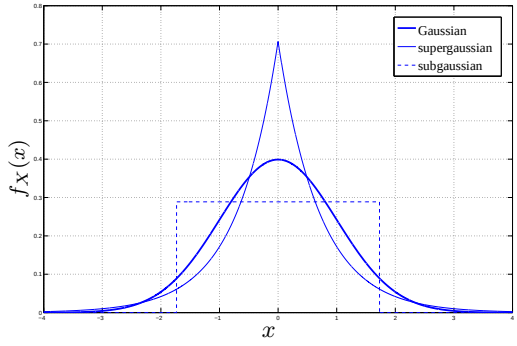


Figure 3: *Gaussian, supergaussian and subgaussian PDF's.*

2.1. Sinusoidal Excitation

Our excitation model is based on the Harmonic plus Noise Model (HNM) of speech, in which the speech signal is defined in terms of a sum of sinusoidal harmonics and a noise component, similar to the one described in [4]. The excitation spectrum is divided into a lower harmonic (voiced) band and an upper noise (unvoiced) band. The frequency point that separates these two bands is referred to here as the harmonic cutoff frequency, denoted F_{max}. In our case:

$$h(t) = \sum_{k=1}^{K(t)} A_k(t)\cos(k\theta(t) + \phi_k(t)) \tag{1}$$

with

$$\theta(t) = 2\pi \int_{-\infty}^{t} F_0(\lambda)(d)\lambda \tag{2}$$

where $h(t)$ denotes the harmonic component of the excitation and $K(t)$ is the time-varying number of harmonics determined by the pitch $F_0(t)$ and $F_{max}(t)$. $A_k(t)$ is the amplitude of the k^{th} harmonic and $\phi_k(t)$ its phase at time t. In our system, the additive noise component is modelled simply as Gaussian white noise and is therefore not limited to the upper band.

This model allows us to specify the excitation signal's frequency envelope by specifying $A_k(t)$, thereby finding a representation which is spectrally similar to the recorded speech units, as shown in Fig. 2. It is, however, impractical to use $A_k(t)$, $\theta(t)$ and $\phi_k(t)$ as the excitation model parameters due to their large number and their dependence on the time-varying F_0 and F_{max}. Instead, we introduce for each monophone a linear *harmonic frequency envelope*, which decays from $A_1(t)$ to $A_{K(t)}(t) = 0$ at F_{max}. This linear function specifies the amplitude of each harmonic in the voiced band, and alleviates the need to estimate each separately. The phase was found to be of little perceptual importance during informal listening tests, and hence not modelled.

2.2. F_{max} Estimation

The separation of voiced and unvoiced components within a single observation of a speech signal is not easy. In this section we present a method for determining the cutoff frequency F_{max} for a particular monophone, based on the sound's degree of voicing. We will employ the excitation signal's Gaussianity as a measure of this voicing. The term "Gaussianity" will be used to refer to the degree to which a signal's distribution approaches that of Gaussian data. This approach is based on the observation that the noise component in a speech LP residual signal (and its spectrum) has a Gaussian distribution.

Consider a Gaussian random variable x with mean $\mu_X = 0$ and variance $\sigma_X^2 = 1$. The PDF of x then has the form:

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \tag{3}$$

We now calculate the expected value of the PDF of a zero mean, unity variance Gaussian random variable as:

$$\begin{aligned} \mathcal{E}[f_X(x)] &= \int_{-\infty}^{\infty} (\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2})^2 dx \\ &= \frac{1}{2\sqrt{\pi}} \end{aligned} \tag{4}$$

using properties of the error function $\text{erf}(a) = \frac{2}{\sqrt{\pi}} \int_0^a e^{-t^2} dt$. We can now quantify the Gaussianity of a residual signal $\varepsilon(n)$ by estimating the expectation:

$$m_{f_\varepsilon} = \frac{1}{N} \sum_{n=0}^{N-1} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\hat{\varepsilon}(n)^2} \tag{5}$$

where $\hat{\varepsilon}(n) = \frac{\varepsilon(n) - \mu_\varepsilon}{\sigma_\varepsilon}$ is the normalised residual. Because unvoiced residual signals are approximately Gaussian, we expect m_{f_ε} to be close to $\frac{1}{2\sqrt{\pi}}$ for these sounds, whereas residual signals that contain some voicing should yield higher values because their true distribution is supergaussian (more peaked than a Gaussian, see Fig. 3).

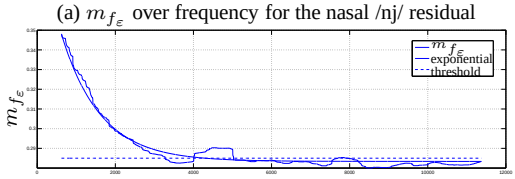
Table 1 lists the values of m_{f_ε} measured for some examples of the different sound classes. As expected, vowels have the highest values. Values for voiced sounds containing more unvoiced content (nasals, approximants and voiced fricatives) are lower, with unvoiced fricatives producing the lowest values of m_{f_ε}, which are very close to the theoretical value of $\frac{1}{2\sqrt{\pi}}$ (≈ 0.2821). Therefore Table 1 confirms that m_{f_ε} is able to estimate the degree of voicing for these sounds.

Plosives sounds, however, yielded erratic values and do not appear to be distinguishable as voiced or unvoiced using m_{f_ε}. This is due to the non-stationary behaviour of these sounds, which are produced by a short burst of energy. This apparently causes a plosive sound residual to be supergaussian, even when the sound itself is completely unvoiced. Furthermore, closer inspection of the values in Table 1 reveals that m_{f_ε} yields values

Table 1: *Gaussianity estimates for various sound classes.*

Class	Phone	m_{f_ε}
Vowel	/i_long/	0.307
	/ep_long/	0.322
	/a_long/	0.316
Nasal	/m/	0.292
	/n/	0.298
	/nj/	0.301
Approximant	/rt/	0.296
	/l/	0.294
	/w/	0.287
V. Fricative	/v/	0.292
	/z/	0.296
	/zh/	0.302
U. Fricative	/f/	0.284
	/s/	0.283
	/sh/	0.283

(a) m_{f_ε} over frequency for the nasal /nj/ residual



(b) m_{f_ε} over frequency for the voiced fricative /zh/ residual

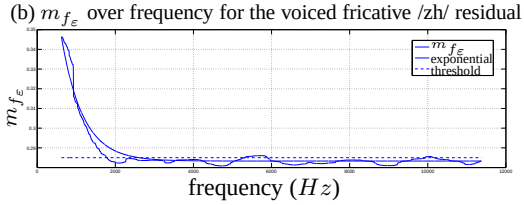


Figure 4: *Exponential curve fitting to residual Gaussianity.*

within similar ranges for nasals, approximants and voiced fricatives. During informal listening tests of synthesised sounds, this was found to result in unnatural nasals and approximants due to a large stochastic component.

Residual signal harmonics in the case of the nasals considered are discernible up to between $3.5kHz$ and $5.5kHz$. In contrast, residual harmonics of voiced fricatives appear to decay more rapidly with frequency and are discernible only up to between $1.5kHz$ and $3kHz$. Also, we found that voiced fricative residual harmonics are larger in magnitude than those of the nasal sounds. These two factors combine to cause the global measures of Gaussianity to be similar for these sounds, since they do not take into account the harmonic frequency envelope.

In order to estimate F_{max} for a given LP residual, we will measure how the residual's Gaussianity changes over frequency. This will allow us to define some threshold F_{max} beyond which we can assume the harmonic content to be zero. Let us therefore apply m_{f_ε} (which has thus far applied to time signals) to FFT spectra. It has been observed that, even though the FFT operation may alter the distribution of the signals, we still find that the FFT samples in the upper (unvoiced) frequencies appear more Gaussian than the lower (voiced) frequencies. We therefore proceed by assuming that the Gaussianity measures will yield values closer to that of ideal Gaussian noise at high frequencies than at low frequencies if voicing is present.

Fig. 4 shows the variation of m_{f_ε} as a function of frequency for the LP residuals of the nasal sound /nj/ (as in th*i*ng) and the voiced fricative /zh/ (as in *g*enre). Note that m_{f_ε} cannot be calculated directly from the magnitude spectrum. Instead, we average the estimates for the real and imaginary parts of the FFT.

The estimation of F_{max} from the graphs shown can be approached in a variety of ways. One possibility is to observe that the measure m_{f_ε} for an LP residual signal of a sound that contains voicing exhibits an approximately exponential decay over frequency. Notice that the graphs tend to the theoretical value of $\frac{1}{2\sqrt{\pi}}$ (≈ 0.2821) for Gaussian data associated with the parameter m_{f_ε}. Knowing this, we can fit a monotonically decreasing exponential curve to the graph to ensure that each point on the graph corresponds to a single frequency. Experimentally, it was found that a value of 0.285 is a suitable threshold for finding F_{max} when using m_{f_ε} as a measure of Gaussianity. The exponential approximations are shown in Fig. 4. Setting the threshold as indicated results in $F_{max} \approx 4.3KHz$ for /nj/ and $F_{max} \approx 2.6KHz$ for /zh/, which corresponds approximately to the frequencies at which the harmonics are no longer discernible in the FFT spectra of these sounds.

3. Filter Parameter Interpolation

LSF's were chosen as parametrisation for the monophone LP filters because their relatively consistent and stable behaviour in the transition regions between phones makes them well-suited for interpolation [5]. A simple, flexible algorithm for calculating the LSF vectors within the transition regions between monophones in a synthetic utterance was developed using modified B-spline curves.

A B-spline curve is an interpolation method which is calculated for K target points M_k such that $k = 0 \ldots K-1$. The driving concept behind the interpolation is that each interpolated point $p(t)$ for $0 \leq t \leq K-1$ is calculated as a weighted contribution of its surrounding target points. The nearest target point has the greatest effect on $p(t)$ to ensure that it follows the general trend of the sequence $M_0 \ldots M_{K-1}$. For the purpose of LSF interpolation between monophones, each LSF's trajectory is modelled by a single B-spline curve such that the k^{th} LSF value, which corresponds to the k^{th} phone in a particular sequence, is equal to M_k. This implies that the smooth curve $p(t)$ represents that LSF's transitions at each point where $t \neq k$. To ensure that $p(t)$ forms a smooth curve, the target point weights must be specified by a continuous function $W(d)$ (called the basis function) of the distance $d_k(t) = k - t$ so that W is a maximum when $d = 0$ and $W(d)$ decreases as $|d|$ increases. In practice, a polynomial approximation of the Gaussian function spanning $-2 < d < 2$, which limits the number of contributing target points to 2 or 3, is normally used. However, we have used a sigmoidal basis function that allows for scalable transition rates and ensures an interpolated curve gradient that is close to zero at its target points, and is defined as:

$$W_s(d,\alpha) = \begin{cases} \frac{c}{1+e^{\alpha(2|d|-1)}} & , 0 \leq |d| < 2; \\ 0 & \text{otherwise.} \end{cases} \quad (6)$$

The choice of c is of no consequence when transforming the target points according to (9).

Several modifications to the B-spline interpolation algorithm were made. These included introducing "phantom" target points at $k = -1$ and $k = K$ to ensure that the initial and final gradients of $p(t)$ are zero, as well as the transformation of the

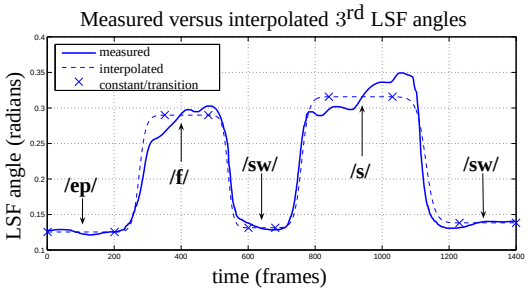


Figure 5: *Interpolation of an LSF within the word "**deficit**".*

entire set of target points to ensure that $p(t)$ passes through all the M_k. These two steps can be summarised as follows:

$$\mathbf{Q} = \begin{bmatrix} -\frac{1}{2} & 0 & \frac{1}{2} & \cdots & 0 & 0 & 0 \\ w_{-1} & w_0 & w_1 & \cdots & 0 & 0 & 0 \\ 0 & w_{-1} & w_0 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & w_0 & w_1 & 0 \\ 0 & 0 & 0 & \cdots & w_{-1} & w_0 & w_1 \\ 0 & 0 & 0 & \cdots & -\frac{1}{2} & 0 & \frac{1}{2} \end{bmatrix}, \quad (7)$$

$$\hat{\mathbf{M}} = \begin{bmatrix} \hat{M}_{-1} \\ \vdots \\ \hat{M}_K \end{bmatrix} \quad \text{and} \quad \mathbf{M} = \begin{bmatrix} 0 \\ M_0 \\ \vdots \\ M_{K-1} \\ 0 \end{bmatrix} \quad (8)$$

where $w_x = W_s(x, \alpha)$ and $\hat{M}$ denotes the transformed target points such that

$$\hat{\mathbf{M}} = \mathbf{Q}^{-1}\mathbf{M} \quad (9)$$

and $\mathbf{Q}^{-1}$ denotes the inverse of the transformation matrix $\mathbf{Q}$.

Further modifications added flexibility to the method in terms of duration control, since the above algorithm assumes uniform spacing between target points. In practice, the spacing is determined by the duration of the inter-phone transition. This can be achieved by choosing the set of values t in $p(t)$ such that each transition contains N_k evenly spaced values of t, where N_k is the number of desired samples in the transition between a predecessor phone k and its successor $k+1$.

The developed LSF interpolation algorithm requires only the sequence of specified phone LSF vectors, and is flexible in terms of the durations of the constant and transition regions of such a sequence. Fig. 5 shows the time-varying ($50ms$ frames, 10 sample increments) measurements of the 3rd LSF of a 30th order LSF vector within the phonetic sequence /ep/→/f/→/sw/→/s/→/sw/ of the word "d*eficit*" (recorded at $24kHz$) together with the interpolated curve through its target values, where constant and transition regions are separated by crosses. Note that it is difficult to measure the LSF trajectories accurately on a short time-frame basis, and therefore the variations in the measured LSF shown in Fig. 5 may be due to inaccuracies in the estimation process rather than actual vocal tract movements, as they are not consistently visible when varying the frame length. Overall, however, the synthetic transitions are fairly similar to the measured transitions, which indicates that the presented B-spline interpolation algorithm is suitable for modelling LSF transitions between phones.

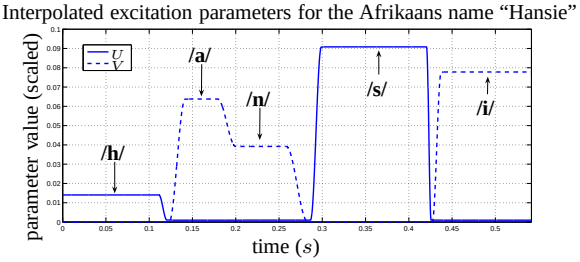


Figure 6: *Example of interpolated source signal parameters.*

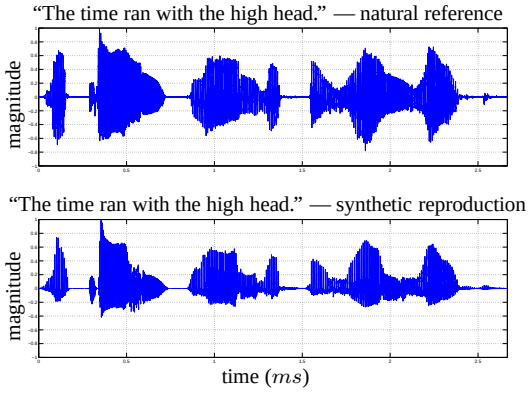


Figure 7: *Time waveforms of a natural and synthetic sentence.*

4. Excitation Parameter Interpolation

An interpolation algorithm for calculating the source signal parameter transitions between phones was developed. Initial experiments applied the same interpolation scheme described in the previous section to the excitation signal parameters, since these are included in each phone's parameter vector. However, it soon became clear that this was not ideal. Co-articulation appears to be more relevant to the vocal tract movements than to the excitation signal, because the oral cavity is physically limited to a certain movement rate, whereas the source signal can change much more abruptly at the glottis. It is for this reason that it was decided to allow the source signal parameters to vary more abruptly than the vocal tract parameters. According to [6], $40ms$ is a suitable duration for the source signal parameter transitions of most phonetic combinations. Since a preprocessor for duration modelling is not present, a default transition duration ($40ms$) had to be chosen. For more natural speech output, however, a rule set derived from speech data is required for effective duration modelling [6].

The particular interpolation algorithm used to generate the source signal parameter transitions was found to be of little perceptual importance during informal listening tests. It is suspected that the $40ms$ window is short enough to make it difficult to discern slight variations in the source signal parameters. It was decided to use a scaled and offset half period of a cosine function to ensure the smooth excitation parameter transition. Fig. 6 shows the interpolated curves of the parameters U (unvoiced magnitude) and V (voiced magnitude) for the Afrikaans name "Hansie". For clarity, F_{max} (which is interpolated in the same way) is not shown.

"The time ran with the high head."
natural reference (above) and synthetic reproduction (below)

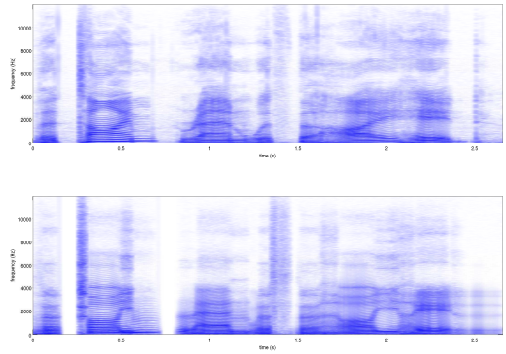


Figure 8: *Spectrogram of a natural and synthetic sentence.*

5. Results and Evaluation

Figures 7 and 8 show the time waveforms and spectrograms, respectively, of a recorded sentence and its synthetic reproduction as used during the system's evaluation. The prosodic contours used in the synthetic sentence were estimated from the natural reference (recording), but the monophone models were obtained independently. Note that the degree of voicing and the harmonic frequency band defined by F_{max}, of the synthetic sentence is very similar to that of the natural reference. The only clear exception to this is the final speech sound in the sentence, /d/, which is a voiced plosive and is known to be problematic. The formant trajectories are similar in shape, although it appears that the estimated duration model for this example did not allow for sufficiently large transition durations.

5.1. Evaluation Procedures

A set of high quality recordings ($F_s = 24kHz$) were made for each of 47 English monophones in a quiet room using a high quality microphone connected to a mixer and a high quality USB audio capture device. A total of 30 LPC's were estimated for each sound without pre-emphasis using the *Levinson-Durbin* algorithm. The parameter F_{max} and the voicing m_{f_c} were estimated for each sound according to the procedure described in section 2.2.

The synthesiser developed in this thesis has been aimed at producing intelligible speech, and little attention has yet been given to the modelling of prosody, which is one of the main factors influencing naturalness. Hence, only intelligibility tests have been carried out. We used the modified rhyme test (MRT, [7]) and semantically unpredictable sentences (SUS, [8]) speech intelligibility tests for this purpose.

The tests were administered in a laboratory where, on average, 10–15 people were present at any given time. Background noise was therefore not eliminated, but it was not sufficient to disrupt the tests. The tests were performed using high quality headphones connected to a personal computer (PC) via an external audio device. All audio files (PCM, 16 bits, mono) were played at a sampling rate of $24kHz$. Among the 25 listeners, 23 were male and 2 female with their ages ranging between 20 and 40. 19 listeners were non-native South African English speakers. Only 3 listeners indicated that they hear synthetic speech on a regular basis.

A simple GUI was developed in order to ease the test procedure. After some general information regarding the tests, each listener was presented first with the MRT, followed by the SUS test. More detailed instructions were shown immediately before each test.

5.2. MRT Results

In order to generate the complete set of 300 MRT words without a text- and linguistic analysis front-end, each word was manually transcribed to a phonetic representation. A duration model was manually defined for each of the 50 ensembles for the common phonetic segments within the ensemble. All words were synthesised at the same constant magnitude level and a pitch curve was randomly selected for each instance from a set of ten curves.

Table 2: *MRT scores for various word subsets.*

Test set	**Number of words**	**Correct**
All words	1200	67.67%
Voiced plosives omitted	817	73.81%
Approximants omitted	910	68.57%
Both of the above omitted	596	75.00%

Table 2 summarises the results of the MRT. Also shown is the percentage of correct words when all words containing voiced plosives are omitted from the scoring set. These sounds include /b/ (as in "*b*aby"), /d/ (as in "*d*eath") and /g/ (as in "*g*un"). These figures are shown because of the known difficulty the current synthesiser has in modelling voiced plosives. Further analysis showed that 174 errors (44.85% of all errors) occurred in words containing voiced plosives, 131 (33.76% of all errors) of which are the result of a voiced plosive being classified incorrectly.

Some listeners commented that the approximants /rt/ (as in "*r*ed") and /l/ (as in "*l*egs") were very unclear. The data seems to support this statement, as there is a slight increase in the word accuracy when these sounds are omitted from the scoring set, although it is less pronounced than in the case of voiced plosives. Analysis showed that 102 errors (26.29% of all errors) occurred in words containing /rt/ or /l/, of which 38 (9.79% of all errors) were caused by an incorrect classification of one of these sounds.

A comparative MRT evaluation of natural speech, 7 formant synthesisers, 1 LPC synthesiser and 1 segment concatenation synthesiser is presented in [9]. As one could expect, natural speech yielded the highest word accuracies. Although the formant synthesisers averaged 88.55%, their individual accuracies ranged from 62.56% to 96.75%. Comparing the overall accuracy of 67.67% obtained by our system, we find that it compares best with the LPC synthesizer at 64.44%.

5.3. SUS Results

For the SUS test, only 15 sentences (3 from each syntactic structure) were synthesised due to the difficulty associated with the manual definition of the phonetic and prosodic elements. These sentences were first recorded, after which each was manually transcribed phonetically. Phone and transition durations as well as pitch curves were manually extracted from each recording for use in the generation of the corresponding synthetic utterance.

Table 3: *Overall SUS test scores on sentence-, word- and phonetic level.*

Test set	**Accuracy**
Sentences	0.53%
Words	41.25%
Words (article "the" omitted)	29.14%
Phonetic transcriptions	47.97%

In order to prepare the listener to the linguistic abnormalities in SUS, the test instructions showed 5 example sentences (1 from each syntactic structure) which were unrelated to the test sentences at a word-level. The sentences were then presented according to the procedure detailed in [8].

The results of the SUS test are shown in Table 3. At first glance, a sentence accuracy of 0.53% appears to be very poor. However, other studies show that sentence-level SUS test scores are typically in the range 10–20% even for natural speech, which reflects the difficulty of the test [8]. This is attributed to the cognitive difficulty associated with transcribing semanticaly unpredictable sentences. Some listeners commented that the speaking rate of some sentences was too high for them to clearly discern the different words in the sentences. Together with the phonetic and prosodic constraints of the synthesis system described earlier, these facts provide some insights into the cause of the low sentence accuracy obtained.

Table 3 shows that the overall word-level accuracy is 41.25%. Also shown is the word-level accuracy for the sentences when the article "the" was not included in the scoring procedure. The decreased accuracy when doing so shows that this word is more easily identifiable than the other words occurring in the sentences. We find that the overall "phonetic accuracy" (refers here to the number of phonetic units, such as phones or diphthongs, classified correctly) is higher than the word-level accuracy. This indicates that many of the incorrect word transcriptions were phonetically similar to the actual words.

A comparative French SUS test evaluation of natural speech, 3 diphone-based synthesisers and 3 unit selection synthesisers is presented in [10]. The results indicate word accuracies of between 60% and 75% and phone accuracies of between 70% and 85% for the synthesizers, whereas natural speech yielded the best results (both word and phone accuracy of approximately 90%). However, the word-level accuracies were obtained by measuring the total number of correct *phonetic transcriptions* of words.

Word-level and phonetic accuracies for each of the sentences individually were also measured because some listeners commented that the longer sentences proved more difficult to remember during transcription. The results obtained seem to support this. When omitting the longer syntactic structures (4 and 5 in [8]), the word-level accuracy climbs to 48.77% (35.59% without articles) and the phonetic accuracy to 55.34%.

6. Conclusions

The speech generation system that has been described in this paper was aimed at the rapid deployment of TTS in new and under-resourced languages. Although time did not allow for formal testing in multiple languages, informal tests suggest that the system will be well suited for this purpose. Compared to concatenative synthesisers, which are the current commercial TTS standard, the amount of transcribed speech data required for a new target language is minimal. An added associated advantage is that the system has a very small memory footprint (120 kilobytes, including the parametric monophone database of 13 kilobytes) and fairly low computational requirements.

A novel approach to the interpolation of monophone speech units was presented to allow realistic transitions between monophone units. Additionally, novel algorithms were presented for the estimation and modelling of the harmonic and stochastic content of an excitation signal.

Within the current system, particular attention is given to certain sound classes, such as plosives, which require specific modelling techniques due to their non-stationary nature. However, a suitable representation for voiced plosives has not yet been found, as is evident from our results. If the system is to be applied to other languages in the future, additional modelling of certain new sound classes, such as the "click" sounds common to African languages, may become necessary.

Even without a text preprocessing front-end, results of the evaluation of the system in South African English show that the synthetic speech generated by this system is moderately intelligible. Since the system's implementation encompasses only the speech generation section of a full TTS synthesis system, these results however cannot be used for direct comparison with other synthesisers before the necessary modules of a full TTS system are added.

7. References

[1] M. Edgington, "Investigating the Limitations of Concatenative Synthesis," in *EUROSPEECH 1997*.

[2] M. Tatham, E. Lewis, and K. Morton, "An Advanced Intonation Model for Synthesis," in *EUROSPEECH 1999*.

[3] D. H. Klatt and L. C. Klatt, "Analysis, Synthesis and Perception for Voice Quality Variations Among Female and Male Talkers," *J. Acoust. Soc. America*, vol. 87, pp. 820–857, 1990.

[4] Y. Stylianou, T. Dutoit, and J. Schroeter, "Diphone Concatenation using a Harmonic plus Noise Model of Speech," in *EUROSPEECH 1997*.

[5] W. B. Kleijn and K. K. Paliwal, Eds., *Speech Coding and Synthesis*. Amsterdam: Elsevier Science, 1998.

[6] J. Allen, M. S. Hunnicutt, and D. Klatt, *From Text to Speech: The MITalk System*. Cambridge University Press, 1987.

[7] A. S. House, C. Williams, M. H. L. Hecker, and K. D. Kryter, "Psychoacoustic Speech Tests: A Modified Rhyme Test," *J. Acoust. Soc. America*, vol. 35, p. 1899, 1963.

[8] C. Benoît, M. Grice, and V. Hazan, "The SUS test: A Method for the Assessment of Text-to-Speech Synthesis Intelligibility Using Semantically Unpredictable Sentences," *Speech Communication*, vol. 18, no. 4, p. 381, 1996.

[9] J. S. Logan, B. G. Greene, and D. B. Pisoni, "Segmental Intelligibility of Synthetic Speech Produced by Rule," *J. Acoust. Soc. America*, vol. 86, no. 2, pp. 566–581, August 1989.

[10] P. B. De Mareüil, C. d'Alessandro, A. Raake, G. Bailly, M.-N. Garcia, and M. Morel, "A Joint Intelligibility Evaluation of French Text-to-Speech Synthesis Systems: the EvaSy SUS/ACR Campaign," in *LREC 2006*.