A MONOPHONE SPEECH GENERATION SYSTEM.

G. Klompje and T.R. Niesler

Department of Electrical and Electronic Engineering, University of Stellenbosch, South Africa.

Abstract: Current speech synthesis systems generally require large and carefully annotated speech corpora for their development. However, for many languages these resources are not available. This paper describes a speech generation algorithm based on monophone subword units for minimal reliance on such databases. The system is based on the source-filter speech production framework, and includes a linear prediction based vocal tract model as well as an excitation model. An interpolation algorithm is presented to allow co-articulation between monophone units to be modelled. The excitation model includes a method for dealing with voiced and partially-voiced sounds based on a Gaussianity measure applied to the excitation spectrum. Promising first results were obtained when evaluating the intelligibility of the developed system's South African English speech output using the modified rhyme test and semantically unpredictable sentences.

Keywords: Text-to-speech, monophone synthesis, multilingual speech synthesis

1. INTRODUCTION

This paper describes the development of a speech generation system based on monophone subword units. The aim is to provide a means of generating intelligible synthetic speech with minimal reliance on specially-prepared and annotated databases. In the longer term, this represents an effort to develop a class of speech synthesis systems that can easily be applied to new languages and accents. Language and accent portability of this nature is particularly important to the development and application of speech technology in a strongly multilingual society such as South Africa.

Currently, concatenative speech synthesis systems are state-of-the-art. However, due to their dependency on an associated carefully-annotated speech corpus, this class of systems has certain inherent disadvantages [4]. 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 and/or prosodic contexts. Recording and annotating a speech database for use in a concatenative system is a very time-consuming and laborious process which may not be suitable in situations where resources are scarce.

Our system falls within the source-filter speech production framework, and makes use of an explicit vocal tract and excitation model. This has the advantage of allowing direct control over parameters such as pitch, duration and emphasis. Furthermore, our system makes use of an inventory of monophone units. This strongly reduces the number of sounds that must be modelled, and guarantees full coverage of the range of sounds that will be required during synthesis. Such guarantees are much harder to achieve using the more common diphone or triphone based systems, for which much larger and carefully-annotated speech databases are required.

The following section describes the structure of our speech generation system. Since we make use of monophones, which are context independent, the co-articulatory effects occurring at the transition between phones must be accounted for explicitly. Section 3 describes an interpolation method which has been developed for this purpose. Section 4 deals with the excitation model, which incorporates an original method for dealing with voiced and partially-voiced sounds. Finally, the effectiveness of our system is evaluated by means of perceptual tests in Section 5.

2. SYSTEM ARCHITECTURE

The general structure of a text-to-speech (TTS) system is shown in Figure 1. In this paper we focus only on the speech generation component of such a system. We assume that the phonetic transcription of the speech to be generated, as well as the prosodic information (pitch and magnitude contours as well as phone durations) are known. The text and linguistic analysis required to derive this prosodic information from a textual representation does not fall within the scope of this work.

The speech generation component shown in Figure 1 is represented in greater detail by Figure 2. The linear-prediction (LP) vocal tract model mimics the effect which the continually-changing geometry of the human vocal tract has on the resulting speech sound. Different sets of LP coefficients (LPCs) represent different geometries and hence different sounds. The excitation model accounts for the action of the vocal chords. For voiced sounds, the vocal chords vibrate and hence the excitation has a dominant periodic component. For unvoiced sounds, there is no vibration and the excitation becomes a stochastic process, usually approximated by Gaussian noise. Sounds with partial voicing can be represented by a combination of these two extremes.

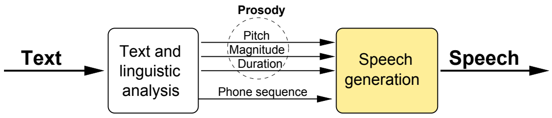

Figure 1: The general structure of a TTS system.

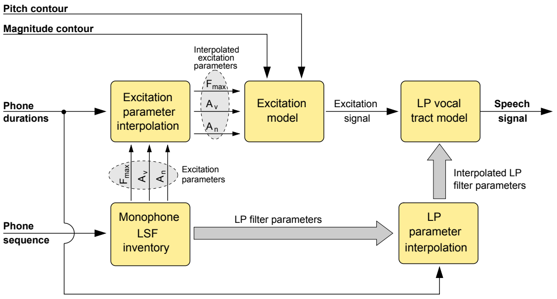

Figure 2: Structure of the speech generation system.

In our current system, each monophone is represented by a single set of 30 LPCs, which model the vocal tract during the production of the sound, as well as set of three parameters describing the accompanying excitation. These parameters are all estimated automatically from recordings of individual monophones. An utterance, specified as a sequence of monophones, is synthesised by generating smooth parameter trajectories between consecutive monophone vectors using an interpolation algorithm.

3. VOCAL TRACT FILTER MODELLING

An LP-based vocal-tract filter model was adopted for the speech generation system. Line spectral frequencies (LSFs) [7] were chosen as parameterisation for the monophone LP filters because their relatively consistent and stable behaviour in the transition regions between phones makes them well-suited for interpolation [8]. For example, Figure 3 shows the trajectories of the first 10 LSFs as calculated during the production of the phonetic sequence /ep/→/f/→/sw/→/s/→/sw/ as occurs within the word "*deficit*". A table of phone symbols appears in Appendix A.

The monophone units used by our speech generation system model only the approximately stationary characteristics of each phone. While systems based on context-dependent units such as diphones model the transitions between successive phones implicitly, our reliance upon monophones

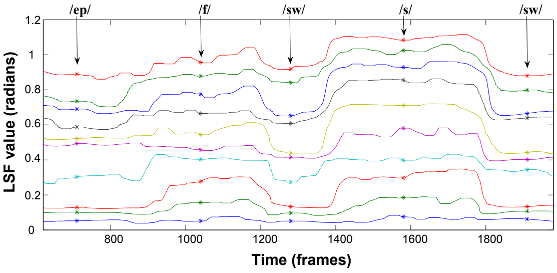

Figure 3: Trajectories of the first 10 line spectral frequencies (LSFs) measured within the word "deficit". One frame is extracted approximately every 0.5ms, each frame has a length of 50ms, and the sampling rate is 24kHz.

requires these co-articulatory effects to be accounted for explicitly. To accomplish this, an algorithm based on modified B-spline curves has been developed. This algorithm estimates the values of LSFs in the transition regions between monophones of the synthetic utterance.

A B-spline curve is a method that calculates interpolated values from a set of K *target points*, denoted by M_k such that $k = 0 \dots K-1$. 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 $M_0 \dots M_{K-1}$. 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 the LSF 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 $W(d)$ (known as 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. A polynomial approximation of the Gaussian function is normally used for this purpose, and spans $-2 < d < 2$ thereby limiting the number of contributing target points to 2 or 3. However, we have used the following sigmoidal basis function $W_s(d, \alpha)$ to ensure that the interpolated curve gradient is close to zero at each target point. This allows the duration of the transition regions to be scaled without introducing audible discontinuities at the target points.

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

Two further modifications to the standard B-spline interpolation algorithm were made. Firstly, "phantom" target points were introduced at $k = -1$ and $k = K$ to ensure that the initial and final gradients of $p(t)$ are zero. Secondly, a transformation was applied to the entire set of target points to ensure that $p(t)$ passes through each M_k. These two steps are summarised by Equation 2.

$$\hat{\mathbf{M}} = \mathbf{Q}^{-1}\mathbf{M} \tag{2}$$

Here $\mathbf{Q}^{-1}$ denotes the inverse of the $(K+2) \times (K+2)$ transformation matrix $\mathbf{Q}$, which is defined as:

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

where $w_x = W_s(x, \alpha)$ and $\mathbf{M}$ is a vector including the K target points:

$$\mathbf{M} = \begin{bmatrix} 0 \\ M_0 \\ \vdots \\ M_{K-1} \\ 0 \end{bmatrix}$$

while $\hat{\mathbf{M}}$ is a vector containing the $K+2$ transformed target points.

$$\hat{\mathbf{M}} = \begin{bmatrix} \hat{M}_{-1} \\ \vdots \\ \hat{M}_K \end{bmatrix}$$

Figure 4 illustrates the modified B-spline interpolation when applied to a sequence of five hypothetical target points, representing the values of the LSF for five successive phones. The figure also demonstrates how larger values of the constant α introduced in Equation 1 lead to more abrupt transitions. Values greater than approximately 7.5 ensure that the gradient of the interpolated curve is almost zero at each target point. Consequently, a value of $\alpha = 7.5$ was used in all further experiments.

The duration of the inter-phone transitions can be controlled by sampling $p(t)$ at N_k evenly spaced values of t in the transition between a predecessor phone k (occurring at $t = k$) and its successor $k+1$ (occurring at $t = k+1$). Hence the developed LSF interpolation algorithm requires only the sequence of LSF vectors corresponding to the phones to be specified. The interpolated curves are then obtained by sampling $p(t)$ according to the desired durations of the phones and the inter-phone transitions. Figure 5 shows the time-varying measurements of the 3rd LSF within the phonetic sequence /ep/→/f/→/sw/→/s/→/sw/ together with the interpolated curve passing through the target values.

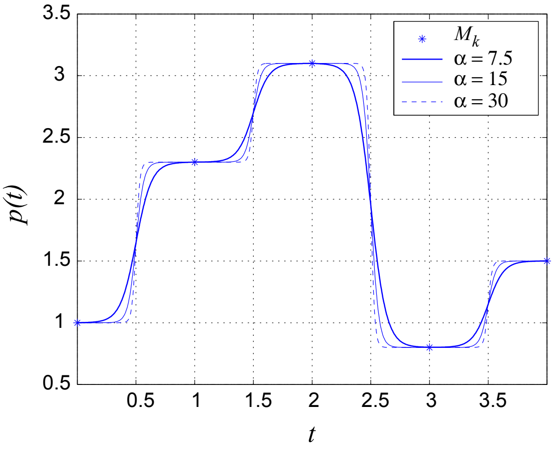

Figure 4: B-spline interpolation using five target points for various values of α, as defined in Equation 1.

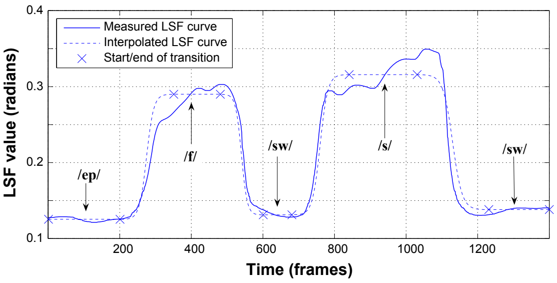

Figure 5: Measured and interpolated 3rd LSF angles within the word "deficit". One frame is extracted approximately every 0.5ms, each frame has a length of 50ms, and the sampling rate is $24kHz$. The boundaries of the transition regions are indicated by "×" markers.

Overall, the synthetic transitions are fairly similar to the measured transitions, which indicates that the presented B-spline interpolation algorithm is suitable for modelling the LSF inter-phone transitions.

4. EXCITATION MODELLING

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 [12]. While the HNM is typically used as a complete speech signal model, we apply it only to the excitation signal. Figure 6 illustrates the measured excitation spectra of three different sounds: the first voiced, the second unvoiced and the third with mixed voicing. In each case the excitation signals were obtained by inverse LP filtering of the recorded waveform. The periodic nature of the voiced excitation is immediately evident in the form of a clear band of harmonics between 0Hz and approximately 4kHz. For mixed-voicing, this band is narrower, and for the unvoiced sound absent altogether.

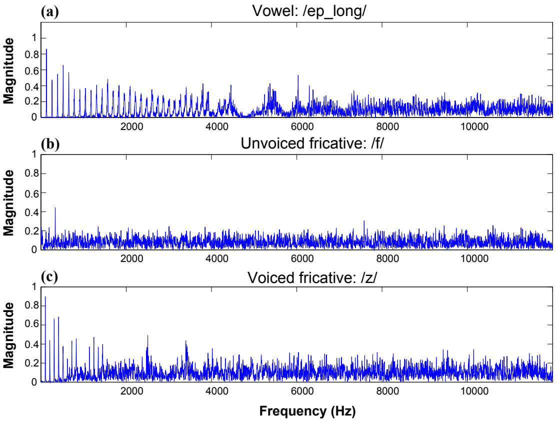

Figure 6: Excitation spectra of three speech sounds: (a) the vowel /ep_long/, (b) the unvoiced fricative /f/ and (c) the voiced fricative /z/.

All spectra in Figure 6 also exhibit a reasonably flat band which is characteristic of the stochastic component present to a certain degree in all speech signals. For the unvoiced sound, this stochastic band extends over the entire spectrum. For the other two signals, the spectrum may be approximated as flat beyond a certain frequency, which we will term the *harmonic cutoff frequency* and denote by F_{max}. Hence our our model divides the excitation spectrum into a lower *voiced band* and an upper *unvoiced band*, which meet at the frequency F_{max}. The model is defined as follows.

$$e(t) = h(t) + A_n(t) \cdot w(t) \tag{3}$$

The quantity $h(t)$ denotes the harmonic component of the excitation, and is expressed as:

$$h(t) = \sum_{k=1}^{K(t)} A_k(t) \cos\Big(\phi_k(t)\Big) \tag{4}$$

where $\phi_k(t)$ is the phase of the k^{th} harmonic and is given by:

$$\phi_k(t) = 2\pi k \int_0^t F_0(\tau) d\tau$$

Here $K(t)$ is the time-varying number of harmonics as determined by the pitch $F_0(t)$ and the frequency $F_{max}(t)$, while $A_k(t)$ is the amplitude of the k^{th} harmonic. In our system, the additive noise component $w(t)$ is modelled simply by Gaussian white noise and is therefore not strictly limited to the upper band. A future improvement may be to include high-pass filtering to account for the dampening effect which inverse LP filtering appears to have on the noise component in the voiced band.

The sinusoidal voicing model allows us to control the frequency envelope of the excitation signal's harmonics by specifying the $A_k(t)$, thereby finding a representation which is spectrally similar to the recorded speech units. It was however considered impractical to estimate all the harmonic amplitudes $A_k(t)$ due to their large number and their dependence on the time-varying $F_0(t)$ and $F_{max}(t)$. Instead, we approximate the voiced portion of the excitation spectrum by introducing for each monophone a linear *harmonic frequency envelope*. This envelope decays linearly from $A_v = A_1(t)$ at frequency $F_0(t)$ to $A_{K(t)}(t) = 0$ at frequency $F_{max}(t)$ and thereby approximates the amplitude of each harmonic in the voiced band by means of the two parameters A_v and F_{max}.

4.1. Estimating F_{max}

We now require a method for determining the harmonic cutoff frequency F_{max} for a particular monophone. In order to estimate the balance between periodic and stochastic components in a speech signal, the autocorrelation sequence of the estimated excitation signal is often used. The autocorrelation is a measure of a signal's periodicity, and hence can be used to estimate a speech signal's degree of voicing.

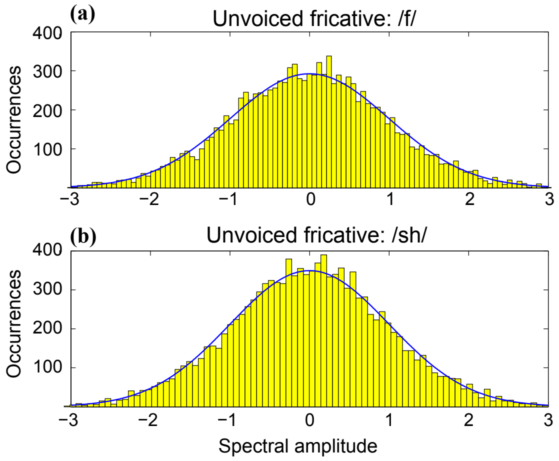

Figure 7: Histograms for the measured excitation spectra of the two unvoiced phones /f/ and /sh/.

Other proposed methods include the use of spectral tilt [10] and correlation-based estimates within different frequency bands [13].

Our estimation procedure is based on the observation that the samples of an excitation signal's spectrum have an approximately Gaussian distribution within the unvoiced band, but are non-Gaussian withing the voiced band. This is illustrated in Figure 7, which shows the distribution of the spectral values for two unvoiced sounds, and in Figure 8, which compares the distribution of the spectral samples from within the voiced and the unvoiced bands respectively of the same voiced fricative. Within the voicing band, the distribution is more highly peaked and may be described as super-Gaussian.

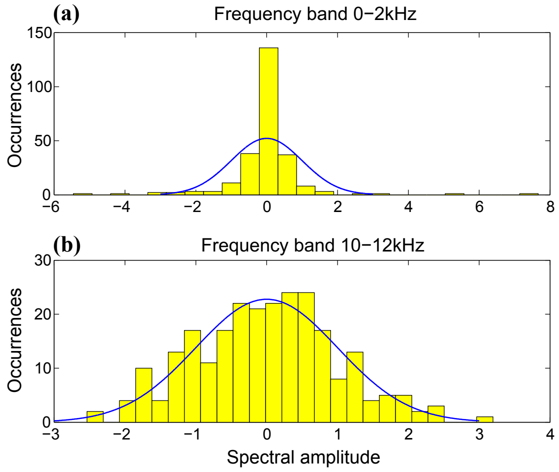

Figure 8: Histograms for the measured excitation spectra from within the (a) voiced and (b) unvoiced bands of the voiced phone /dh/ (as used in "this").

Hence, in order to estimate the cutoff frequency F_{max} we will measure the "Gaussianity"of the excitation spectrum, by which we mean the degree to which this signal's distribution approaches that of Gaussian data.

To measure Gaussianity, we have developed an expectation-based measure which we have found to be more computationally efficient than alternatives such as entropy or kurtosis [6, 11]. Consider a Gaussian random variable x with mean $\mu_X = 0$ and variance $\sigma_X^2 = 1$. The probability density function (PDF) of x then has the form:

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

The expected value of this PDF for a zero mean, unity variance Gaussian random variable may therefore be calculated as:

$$\begin{aligned} \mathcal{E}[f_X(x)] &= \int_{-\infty}^{\infty} (\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2})^2 dx \\ &= \frac{1}{2\sqrt{\pi}} \end{aligned} \qquad (5)$$

where we have used 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 an excitation signal $\varepsilon(n)$ by estimating the expected value of $f_X(x)$:

$$m_{f_\varepsilon} = \frac{1}{N} \sum_{n=0}^{N-1} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\hat{\varepsilon}(n)^2} \qquad (6)$$

where $\hat{\varepsilon}(n) = \frac{\varepsilon(n) - \mu_\varepsilon}{\sigma_\varepsilon}$ are the normalised excitation spectrum samples. Because samples drawn from the unvoiced band of an excitation spectrum have an approximately Gaussian distribution, we expect m_{f_ε} to be close to $\frac{1}{2\sqrt{\pi}}$. In contrast, higher values of m_{f_ε} should be observed for samples drawn from voiced or partially-voiced bands of an excitation spectrum, since their distribution is more highly peaked than Gaussian.

In order to estimate F_{max} for a given excitation signal, we will measure how its Gaussianity changes with frequency. This will allow us to define a threshold F_{max} beyond which we can assume the harmonic content to be zero. Figure 9 shows the variation of m_{f_ε} as a function of frequency for the excitation signals obtained for of the nasal sound /nj/ (as found in "thing") and the voiced fricative /zh/ (as found in "genre"). Each m_{f_ε} value was computed by averaging the estimates obtained for the real and imaginary spectral components within a $1kHz$ window of the spectrum centered at the corresponding frequency. A median filter was subsequently applied to the sequence of m_{f_ε} values. The median filter removes outliers in the m_{f_ε} curve that occur due to the stochastic nature of the spectral estimates. The initial and final segments of the m_{f_ε} curves are not shown, since these fall outside of the $1kHz$ window.

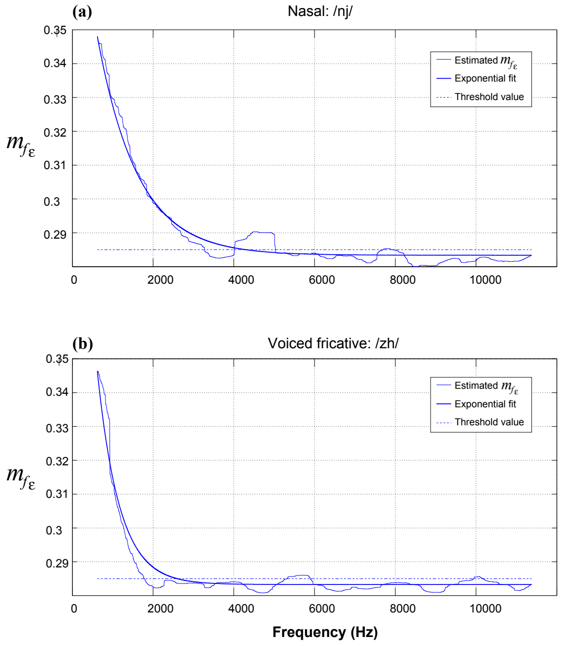

Figure 9: Estimated degree of Gaussianity (m_{f_ε}) as a function of frequency for (a) the nasal /nj/ and (b) the voiced fricative /zh/.

As shown in Figure 9, the value of m_{f_ε} for an excitation signal of a sound that contains voicing exhibits an approximately exponential decay over frequency. In particular, the graphs tend to the theoretical value found in Equation 5 of $\frac{1}{2\sqrt{\pi}}$ (≈ 0.2821) for Gaussian data associated with the parameter m_{f_ε}. Hence we have chosen to fit a decaying exponential curve to the graph, as also shown in Figure 9. Experimentally, it was found that the intersection of this curve with the threshold value of 0.285 gives a good estimate of the boundary between the voiced and unvoiced bands F_{max}. As illustrated in Figure 9, using this threshold results in $F_{max} \approx 4.3KHz$ for /nj/ and $F_{max} \approx 2.6KHz$ for /zh/. These values correspond approximately to the frequencies at which the harmonics are no longer discernible in the spectra of these sounds, as shown in Figure 6. Figure 10 shows examples of corresponding excitation spectra synthesised using the methods described in this section and application of Equations 3 and 4.

4.2. Excitation Parameter Interpolation

Initial experiments applied the same interpolation scheme described in Section 3 to the excitation signal parameters. However, co-articulation appears to be more relevant to the vocal tract movements than to the excitation signal, because the movement of the oral cavity is more physically limited than that of the glottis. For this reason the source signal parameters were allowed to vary more abruptly than the vocal tract parameters. According to [1], $40ms$ is a suitable duration for the source signal parameter transitions of most phonetic combinations. For more natural speech output, however, more refined duration modelling will be necessary [1].

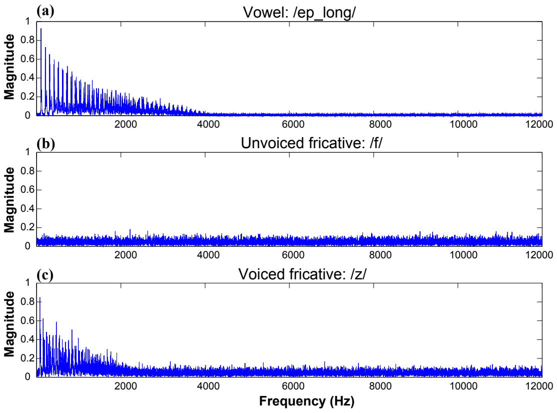

Figure 10: Synthetic excitation spectra for the three speech sounds analysed in Figure 6.

In the event of a filter parameter transition which is shorter than $40ms$, the source signal parameter transition duration is limited so as to not exceed that of the filter parameters.

The particular interpolation algorithm used to generate the source signal parameter transitions was found during informal listening tests to be of little perceptual importance. It is suspected that the $40ms$ window is short enough to make it difficult to discern slight variations in the source signal parameters. Interpolation was accomplished using a scaled and offset half period of a cosine function to ensure the smooth excitation parameter transition. Figure 11 shows the interpolated curves of the parameters A_n (unvoiced magnitude) and A_v (voiced magnitude) for the Afrikaans name “Hansie”. For clarity, F_{max} (which is interpolated in the same way) is not shown.

5. EVALUATION AND RESULTS

A set of high quality recordings was made for each of the 47 South African English monophones described in Appendix A. Recording took place in a quiet room using a high quality microphone connected to a personal computer via an external audio capture device sampling at 24kHz. A total of 30 LPC’s were estimated for each sound using the Levinson-Durbin algorithm. The frequency F_{max} was estimated for each sound according to the procedure described in Section 4.1.

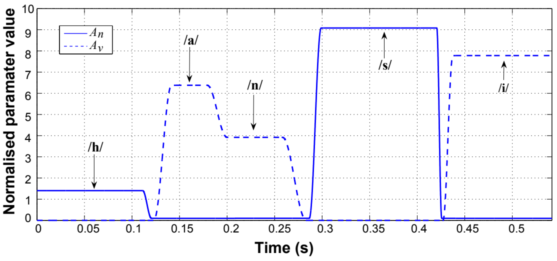

Figure 11: The interpolated excitation parameters A_v and A_n for the Afrikaans name “Hansie”.

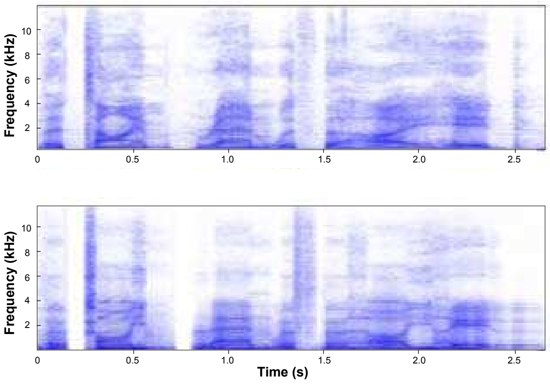

Figure 12: Spectrograms for the sentence “The time ran with the high head.”. The natural speech is shown above and the synthetic reproduction below.

Figure 12 shows the spectrograms of a recorded sentence and its synthetic reproduction as used during the system’s evaluation in Section 5.2. The prosodic contours used in the synthetic sentence were estimated from the original recording, but the monophone models were obtained independently. The degree of voicing and the 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. Voiced plosive sounds were found to be modelled poorly by our current system, due to their highly non-stationary nature, and an improved strategy for such sounds is an important aspect of ongoing work. The formant trajectories of the two spectrograms are similar, although it appears that the durations were not accurately estimated and consequently the filter transitions are shorter in the case of the synthetic utterance than for the natural reference.

The primary objective of the speech generation system developed in this paper has been the production of intelligible speech, and little attention has been given to the modelling of prosody. Hence, only intelligibility tests have been carried out, and naturalness (which is influenced strongly by prosody) has not been measured. We used the modified rhyme test (MRT) and semantically unpredictable sentence (SUS) speech intelligibility tests for this purpose [2, 5]. The incorporation of prosodic modelling for improved naturalness will include the development of a text- and linguistic analysis front-end as shown in Figure 1, and remains the subject of ongoing work.

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 via an external audio interface. All audio files used linear 16-bit PCM encoding and were played at a sampling rate of $24kHz$. Among the 25 listeners, 23 were male and 2 were female, with ages ranging between 20 and 40. There were 19 non-native South African English speakers among the listeners, and only 3 indicated that they hear synthetic speech on a regular basis.

A simple graphical user interface (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, and then the SUS test. More detailed instructions were shown immediately before each test.

5.1. MRT Results

The modified rhyme test is designed to quantify the intelligibility of a speech synthesiser by measuring the confuseability of word-initial and word-final consonants in synthetic speech [5]. The test consists of 50 sets each containing 6 words. In 25 of these sets, only the word-initial consonant varies, as for example in the set "*kit, bit, fit, hit, wit, sit*". In the remaining 25 sets only the word-final consonant varies. During execution of the MRT, the 50 sets are processed in random order. For each set, a single word is played to the test subject, who must identify it from a list of all six words.

In order to generate the complete set of 300 MRT words, each word was manually transcribed to a phonetic representation. A duration model was defined by hand for the common phonetic segments in each of the 50 ensembles. For the initial and final consonants, durations were set according to a simple set of rules. This procedure was not optimised and therefore possibly not ideal. All words were synthesised at the same magnitude level and a pitch curve was randomly selected for each instance from a set of ten curves previously estimated and manually corrected from recordings of MRT words. The pitch curves were chosen to be similar in an effort to prevent listener bias toward any particular one. The procedure described in [5] was followed for the test presentation, except that the listeners' responses were logged automatically, rather than making use of answer sheets.

Table I: System performance for various word subsets according to MRT scores.

Subset of MRT words	**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 I summarises the results of the MRT, and shows that an overall accuracy of 67.67% was achieved. This result can be viewed in the context of a previous comparative MRT evaluation, in which 7 formant synthesisers, 1 LPC synthesiser, 1 segment concatenation synthesiser as well as natural speech were considered [9]. As one might expect, the study shows that natural speech yielded the highest word accuracies. Although the average accuracy of the formant synthesisers was 88.55%, their individual accuracies ranged from 62.56% to 96.75%. In particular, the performance of our system compares favourably with that of the tested LPC synthesizer, which exhibited an accuracy of 64.44%.

Also shown in Table I is the percentage of correct words when all words containing voiced plosives are omitted from the scoring set. These sounds include /b/ (as used in "baby"), /d/ (as used in "death") and /g/ (as in "gun"). 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 the total number) occurred in words containing voiced plosives, 131 (33.76%) of which are the result of a voiced plosive being classified incorrectly.

Some listeners commented that the approximants /rt/ (as in "red") and /l/ (as in "legs") 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 the total number) occurred in words containing /rt/ or /l/, of which 38 (9.79%) were caused by an incorrect classification of one of these sounds. Other sound classes were selectively omitted from the scoring, but resulted in only small changes in the word accuracy, and are therefore not shown in Table I. The final row in the table shows that the accuracy rises to 75% when all words that contain any of the problematic sounds (/b/, /d/, /g/, /rt/ and /l/) are omitted from the test.

5.2. SUS Results

The semantically unpredictable sentence (SUS) test measures speech intelligibility by presenting the user with synthesised sentences that are syntactically correct but semantically meaningless [2]. An example of a SUS is "*Start the trial and the fund*". On hearing the synthetic utterance, the user is asked to transcribe it, and the accuracy of this transcription is then used as a measure of intelligibility. Since the sentences are meaningless, the listener is unable to make use of contextual information to identify the words.

For our SUS test, only 15 sentences (3 from each syntactic structure) were synthesised due to the difficulty associated with the manual definition of the prosodic information. A list of these sentences appears in Appendix B. The sentences were first recorded, after which each was phonetically transcribed by hand. Phone and transition durations as well as pitch curves were extracted from each recording for use in the generation of the corresponding synthetic utterance. A smoothed magnitude envelope was extracted from the recording and used to normalise the energy envelope of the corresponding synthetic sentence.

Since the SUS test was performed directly after the MRT, and because only 15 sentences were generated, it was decided that no listener training is necessary prior to the SUS tests. Instead, it was assumed that a listener was already accustomed to the acoustic quality of the synthetic speech. However in order to prepare the listener to the linguistic abnormalities in SUS, five example sentences (one from each syntactic structure) which were unrelated to the test sentences at a word-level were displayed before commencement of the test. The 15 test sentences were then presented according to the procedure detailed in [2].

Table II: System performance according to SUS test scores as word and at phonetic level.

Test set	Accuracy
Words	41.25%
Words (article “the” omitted)	29.14%
Phonetic transcriptions	47.97%

The results of the SUS test are shown in Table II. At first glance, the accuracies appear to be very poor. However, other studies show that SUS test scores are typically low even for natural speech [2]. This is attributed to the cognitive difficulty associated with transcribing semantically unpredictable sentences.

Table II 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. By considering the phonetic transcriptions of the sentences as well as of the listeners’ responses, we find that the overall phonetic accuracy is higher than the word-level accuracy. This indicates that many of the incorrect word transcriptions were phonetically similar to the true words.

A French SUS test evaluation comparing natural speech, 3 diphone-based synthesisers and 3 unit selection synthesisers is presented in [3]. The results indicate word accuracies of between 60% and 75% and phone accuracies of between 70% and 85% for the synthesizers. Natural speech yielded the best results, with word and phone accuracies both approximately 90%.

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 as defined in [2]), the word-level accuracy climbs to 48.77% and the phonetic accuracy to 55.34%.

6. SUMMARY AND CONCLUSIONS

The speech generation system that has been described in this paper is intended to allow rapid development of speech synthesis systems in under-resourced languages. The system is based on monophone units, and makes use of a novel interpolation algorithm to synthesis realistic transitions between these context independent units. In addition, a simple excitation model is developed, which allows estimation and modelling of the harmonic and stochastic content of an excitation signal. Compared to concatenative synthesisers, which represent the current commercial state-of-the-art, the amount of transcribed speech data required for the proposed system is minimal. Added associated advantages are a small memory requirement and fairly low computational load.

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. Further attention will have to be given to highly non-stationary sounds, especially voiced plosives, which were found to be synthesised poorly. Since only the speech generation section has been tested, however, it is difficult to directly compare these results with other synthesisers.

7. REFERENCES

[1] J. Allen, S. Hunnicutt, and D.H. Klatt. *From text to speech: The MITalk system*. Cambridge University Press, 1987.

[2] 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*, 18(4):381–392, 1996.

[3] 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 *Proc. LREC*, Genoa, Italy, 2006.

[4] M. Edgington. “Investigating the limitations of concatenative synthesis”. In *Proc. Eurospeech*, Rhodes, Greece, 1997.

[5] A. S. House, C. Williams, M. H. L. Hecker, and K. D. Kryter. “Psychoacoustic speech tests: A Modified Rhyme Test”. *J. Acoust. Soc. America*, 35:1899, 1963.

[6] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. John Wiley, New York, 2001.

[7] F. Itakura. “Line spectrum representation of linear predictive coefficients of speech signals”. *J. Acoust. Soc. America*, 57:35, 1975.

[8] W. B. Kleijn and K. K. Paliwal, editors. *Speech coding and synthesis*. Elsevier Science, Amsterdam, 1998.

[9] J. S. Logan, B. G. Greene, and D. B. Pisoni. “Segmental intelligibility of synthetic speech produced by rule”. *J. Acoust. Soc. America*, 86(2):566–581, August 1989.

[10] P. J. Murphy. “Spectral tilt as a perturbation-free measurement of noise levels in voice signals”. In *Proc. Eurospeech*, Aalborg, Denmark, 2001.

[11] C. E. Shannon. “A mathematical theory of communication”. *The Bell System Technical Journal*, 27:379–423, 623–656, 1948.

[12] Y. Stylianou, T. Dutoit, and J. Schroeter. "Diphone concatenation using a harmonic plus noise model of speech". In *Proc. Eurospeech*, Rhodes, Greece, 1997.

[13] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura. "Mixed excitation for HMM-based speech synthesis". In *Proc. Eurospeech*, Aalborg, Denmark, 2001.

APPENDIX A: PHONE INVENTORY

The following table describes the South African English phones used by the speech generation system. An example of a word in which the phoneme occurs in each language is given in each case. If this example appears in italics, it indicates that the word has been borrowed from a different language.

DESCRIPTION	LABEL	EXAMPLE
Stops		
Voiceless Bilabial Plosive	p	pit
Voiced Bilabial Plosive	b	baby
Voiceless Alveolar Plosive	t	total
Voiced Alveolar Plosive	d	death
Voiceless Velar Plosive	k	kick
Voiced Velar Plosive	g	gun
Fricatives		
Voiceless Labiodental Fricative	f	four
Voiced Labiodental Fricative	v	vat
Voiceless Dental Fricative	th	thing
Voiced Dental Fricative	dh	this
Voiceless Alveolar Fricative	s	some
Voiced Alveolar Fricative	z	zero
Voiceless Post-Alveolar Fricative	sh	shine
Voiced Post-Alveolar Fricative	zh	genre
Voiceless Velar Fricative	x	*Gauteng*
Voiceless Glottal Fricative	h	hand
Voiced Glottal Fricative	hht	*Johannes*
Approximants		
Alveolar Approximant	rt	red
Alveolar Lateral Approximant	l	legs
Palatal Approximant	j	yes
Voiced labio-velar Approximant	w	west
Nasals		
Bilabial Nasal	m	man
Alveolar Nasal	n	not
Velar Nasal	nj	thing
Vowels		
High Front Vowel	i	*Piet*
High Front Vowel with duration	i_long	keep
Lax Front Vowel	ic	him
High Back Vowel	u	*Kapkaroord*
High Back Vowel with duration	u_long	blue
Lax Back Vowel	hs	push
Mid-high Front Vowel with duration	e_long	*Vrede*
Rounded Mid-high Back Vowel	o_long	*Sibongile*
Mid-low Front Vowel	ep	nest
Mid-low Front Vowel with duration	ep_long	fairy
Rounded Mid-low Front Vowel	oe	nurse
Rounded Mid-low Front Vowel with duration	oe_long	burst
Central Vowel with duration	epr_long	turn
Rounded Mid-low Back Vowel	ct	*Hartenbos*
Rounded Mid-low Back Vowel with duration	ct_long	bore
Low Back Vowel	ab	hot
Lax Mid-low Vowel	vt	hut
Low Central Vowel	a	*Garsfontein*
Low Central Vowel with duration	a_long	*Klerksdorp*
Low Back Vowel with duration	as_long	harp
Central Vowel (Schwa)	sw	the
Mid-low Front Vowel	ae	average
Mid-low Front Vowel with duration	ae_long	dad

APPENDIX B: SUS SENTENCES

The following table lists the 15 semantically unpredictable sentences (SUS) that were employed during the perceptual tests.

Structure	Sentence
(1)	The state went at the hot point. The school stayed for the new tube. The time ran with the high head.
(2)	The poor sense hit the tax. The short field found the step. The thin job got the voice.
(3)	Start the trial and the fund. Call the game and the front. Live the plant or the test.
(4)	When does the dog lead the hard set? Why does the sign learn the green bed? How does the chance plan the cold roof?
(5)	The song marked the branch that burned. The truth helped the leg that failed. The top drew the pool that died.