# A New Speech Synthesis Based on Fractal

S. Fekkai, M. Al-Akaidi
Faculty of Computing Sc. & Engineering,
De Montfort University, Leicester, LE1 9BH, UK.
Email: mma@dmu.ac.uk

**ABSTRACT**

Speech is the primary means of communication between people. Speech synthesis, automatic generation of speech waveforms has been under development for several decades [1, 2]. Recent progress in speech synthesis has produced synthesizers with very high intelligibility but the sound quality and naturalness still remain a major problem. The goal of this work is to develop a new speech synthesis system, which is based mainly on the fractal dimension to create natural sounding speech. Our initial work in this area showed that by careful use of the fractal dimension together with the phase of the speech signal to ensure consistent intonation contours, natural -sounding speech synthesis was achievable with word level speech. In order to extend the flexibility of this framework, we focused on the filtering and compression of the phase to maintain and produce natural sounding speech.

## 1 Introduction

In an ideal world, a speech synthesizer should be able to synthesize any arbitrary word sequence with complete intelligibility and naturalness. Figure 1 illustrates the trade-off of how current synthesizers have tended to strive for flexibility of vocabulary and sentences at the expense of naturalness (i.e., arbitrary words can be synthesized, but do not sound very natural). This applies to articulatory, rule-based and concatenative methods of speech synthesis [3, 4, 5, 6].

An alternative strategy is one, which seeks to maintain naturalness by operating in a constrained domain. There are potentially many applications where this mode of operation is perfectly suitable. In conversational systems for example, the domain of operation is often quite limited, and is known ahead of time [7]. Researchers in the past have examined how unit selection algorithms can be formulated, and what constraints must be maintained [3, 5, 6].

In this work, we have developed a frame work for natural-sounding speech synthesis using fractal dimension. The developmental philosophy that we have adhered to throughout the work, places naturalness as a paramount goal. In our preliminary work involving word fractal dimension, the vocabulary size is relatively small, but naturalness is very high. Our research follows the bottom curve of Figure 1 where we view naturalness as the highest priority.

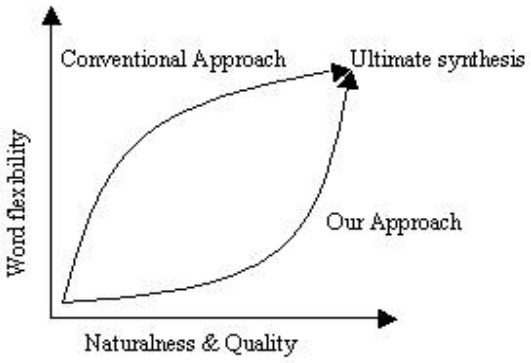


Figure 1: Synthesis development trade-off schematic.

## 2 Fractal Dimension

Fractal dimension as parameter is important because it can be defined in terms of real-world data, and can be measured approximately by means of experiment [8, 9, 10]. Fractal dimension is a real number that in general, falls between the limits of 1 and 5 and can be calculated in a number of ways. For the case of fractal speech signals and curves the fractal dimension lie between 1 and 2. The Power Spectrum Method (PSM) [11] has been used as an application of the Fourier power spectrum technique to calculate the fractal dimension of speech phonemes. The speech signal is Fourier Transformed by means of an FFT and the power spectrum is computed, $P_i = Re(k_i)^2 + Im(k_i)^2$. Assume that $P_i$ is the measured power spectrum then $\hat{P}_i$ is the expected form of the fractal power spectrum, $\hat{P}_i = c|k_i|^{-\beta}$, where $c$ is a positive constant and $\beta$ the positive spectral exponent [12].

Applying the least Square approach to calculate the spectral exponent $\beta$ and $c$ yields to the following equation:

$$\beta = \frac{N\sum_{i=1}^{N}(\ln P_i)(\ln|k_i| - (\sum_{i=1}^{N}\ln P_i)(\sum_{i=1}^{N}\ln|k_i|)}{\sum_{i=1}^{N}\ln|k_i)^2 - N\sum_{i=1}^{N}\ln|k_i)^2} \quad (1)$$

&

$$C = \frac{\sum_{i=1}^{N}\ln P_i - \beta\sum_{i=1}^{N}\ln|k_i}{N} \quad (2)$$

where $C = \ln c$. Using the relationship:

$$D = \frac{5-\beta}{2} \quad (3)$$

Provides a simple formula for computing the fractal dimension from the power spectrum of a signal.

The implementation of the PSM consists of applying the FFT to the speech signal in order to obtain a spectral representation of the phoneme. A pre-filter step is then used to adjust the estimated values of the fractal dimension to fit within the range 1 and 2. The power spectrum of the pre-filtered signal is computed then the least square approach is applied to calculate the power exponent $\beta$ (Eq.1). Hence the fractal dimension $D$ (Eq. 3) is obtained.

It is important to mention that without the pre-filtering step, the values of the fractal dimension were not satisfying the range of the fractal model. However the use of the Pre-filter ( $\frac{1}{w}$) has the effect of confirming the speech data to fit the range of the fractal dimension for speech signal which lies between the range 1 and 2.

## 3 Non-Stationary algorithms for speech Synthesis

Synthesized speech can be produced by several different methods. All of these have some benefits and deficiencies. The methods are usually classified into three groups:

1. Articulatory synthesis, which attempts to model the human speech production system directly.

2. Formant synthesis, which models the pole frequencies of speech signal or transfer function of vocal tract based on source-filter-model.

3. Concatenative synthesis, which uses different length prerecorded samples derived from natural speech.

The formant and concatenative methods are the most commonly used in present synthesis systems. The formant synthesis was dominant for long time, but today the concatenative method is becoming more and more popular. The articulatory method is still too complicated for high quality implementations, but may arise as a potential method in the future.

Articulatory synthesis typically involves models of the human articulators and vocal cords. The articulators are usually modeled with a set of area functions between glottis and mouth. The first articulatory model was based on a table of vocal tract area functions from larynx to lips for each phonetic segment[13]. For rule-based synthesis the articulatory control parameters may be for example lip aperture, lip protrusion, tongue tip height, tongue tip position, tongue height, tongue position and velic aperture. Phonatory or excitation parameters may be glottal aperture, cord tension, and lung pressure [14].

When speaking, the vocal tract muscles cause articulators to move and change shape of the vocal tract, which causes different sounds. The data for articulatory model is usually derived from X-ray analysis of natural speech. However, this data is usually only 2-D when the real vocal tract is naturally 3-D, so the rule-based articulatory synthesis is very difficult to optimize due to the unavailability of sufficient data of the motions of the articulators during speech. Other deficiency with articulatory synthesis is that X-ray data do not characterize the masses or degrees of freedom at the articulators [13]. The movements of tongue are so complicated that it is almost impossible to model them precisely. Advantages of articulatory synthesis are that the vocal tract models allow accurate modeling of transients due to abrupt area changes, whereas formant synthesis models only spectral behavior [15].

In the next section we will introduce the synthesis of speech using fractal. A new technique, which develop a framework for natural sounding speech synthesis.

## 4 Synthesising Speech with Fractals

In the previous section we have discussed how the power spectrum of a signal's Fourier transform can be used to extract the fractal dimension. This followed from assuming the power spectrum, $\hat{P}_i$, was related to the dimension in the following form,

$$\hat{P}_i = c|k_i|^{-\beta}, \quad \text{where} \quad \beta = 5 - 2D$$

To create a synthetic fractal is then the process of filtering white noise of the required size with a low pass filter, $q$ whose Fourier transform is:

$$Q(k) = |k|^{-\frac{\beta}{2}}, \quad \text{where} \beta = 5 - 2D$$

Using this principle we will start by creating a fractal signal. The process is consist of four stages explained bellow:

**Step 1:** Compute a random Gaussian distributed array $G_i, i = 0, 1, ..., N-1$ using a conventional Gaussian random number generator, with zero mean and unit variance. Compute a random number sequence of uniform distributed numbers $U_i, i = 0, 1, ..., N-1$ in the range zero to one.

**Step 2:** Calculate the real and imaginary parts; $N_i = G_i \cos 2\pi U_i$ and $M_i = G_i \sin 2\pi U_i$. This defines $G_i$ as the amplitude and $U_i$ as the phase.

**Step 3:** Filter $N_i$, $M_i$ with $W_i = \frac{1}{K_i^{\beta/2}}$ to create $N'$ and $M'$.

**Step 4:** Inverse DFT the result using a FFT to obtain

$$n_i = Re(\hat{F}^{-1} N' + iM')$$

The exponent is $\beta/2$ to ensure that the power spectrum,$P_k$ , satisfies

$$P_k = (N'_k)^2 + (M'_k)^2 \propto k^{-\beta}$$

By using the same random noise for $U$ and $G$, we can see how changes to $D$ affect the signal.

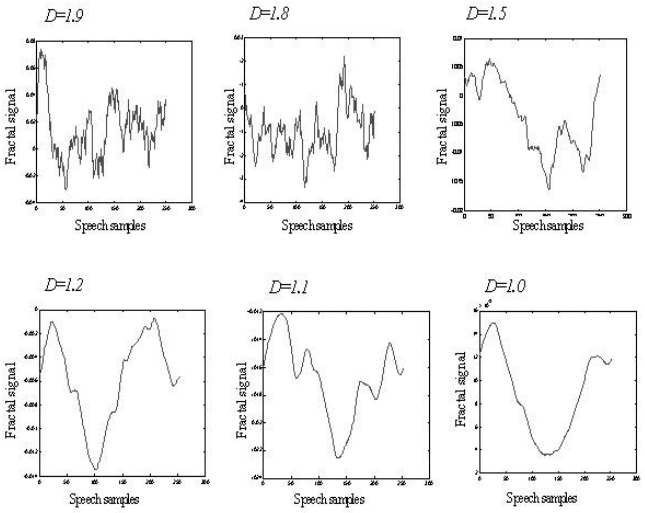


Figure 2: Fractal Signals .

The next section will illustrate the new algorithm used to reproduce a natural sounding speech synthesis system, which make use of the fractal dimension of the word and its unwrapped phase.

## 5 Algorithm Used

The main objective of this algorithm is to create natural sounding speech and to ensure consistent intonation contours. The work carried out was based on two hypotheses: First, that the phase of the speech signal carries important information, hence intelligibility of the synthesis speech. Second, that the fractal dimension characteristics, if used in the energy of the signal, will reproduce a natural sounding speech.

The algorithm steps are summarised in the block diagram given in Figure 3.

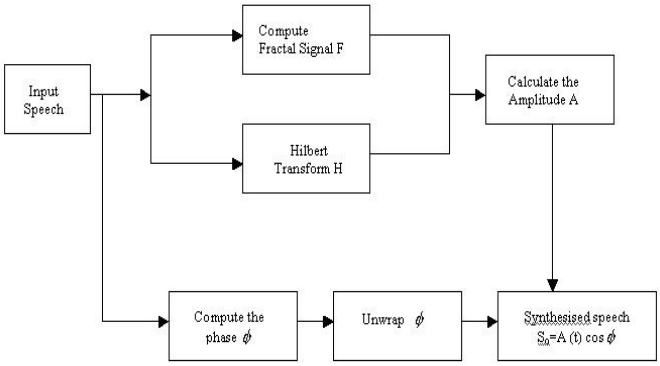


Figure 3: Block diagram of speech synthesis using Fractal.

## 6 Experiments & Discussion

Three experiments have been conducted in the simulation process involving four different words namely "test", "best", "Open" and "zone".

In the first experiment, only the unwrapped phase of the word has been used along with the fractal signal. In the second one the phase is then 65% compressed from the original. In the third experiment the phase is low pass filtered and the remaining signal is replaced by a white random noise.

The three experiments gave good quality and intelligibility of the synthesised words and they all sounded very natural, however, among the three the best natural sounding speech was enhanced when the phase of the speech signal was low pass filtered and white noise added.

It is important to mention here, that the use of the fractal signal in the energy of the reconstructed speech signal has the effect to control the naturalness sounding of speech synthesis.

Figure 4 shows the unwrapped phase of the word "best", the amplitude, as well as the original word and

its reconstructed one. We can clearly notice the similarities in the waveforms of the input speech word with the reconstructed one.

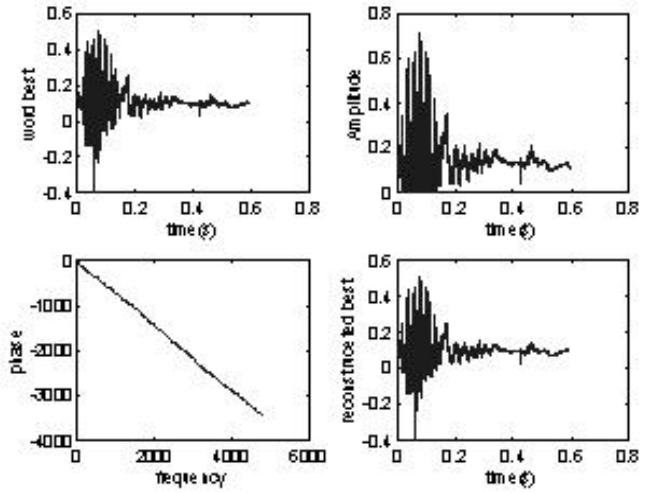


Figure 4: Synthesis results for word "best"

To test the validity of our results 10 people have been listening to the synthetic speech and their evaluation is elaborated in Figure 5.

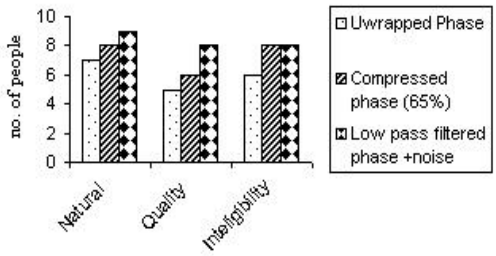


Figure 5: Evaluation of the synthesis

## 7 Conclusion

A new algorithm based on fractals has been used for the synthesis of speech words. The synthesis process involved three cases of experiments, which increased the quality and the intelligibility of the synthesised speech. The naturalness level we placed as paramount in our work was highly achieved as a result of the fractal characteristic used in the synthesis process. Despite the small size of vocabulary we used, the naturalness is very high and as the pursuit of naturalness dominates, human listening provided the best feedback.

## References


[1] Kleijn K., Paliwal K., "Speech coding and synthesis", Elsevier Science B.V., The Netherlands, 1998.

[2] Santen J., Sproat R., Olive J., Hirschberg J., "Progress in speech synthesis", Springer-Verlag, New York Inc., 1997.

[3] Campbell N., "CHATR: A high-definition speech re-sequencing system," Acoustical Society of America and Acoustical Society of Japan, Third Joint Meeting, Dec. 1996.

[4] Huang X., Acero A., Adcock J., Hon H., Goldsmith J., Liu J., and Plumpe M., "Whistler: A trainable text-to-speech system," in Proc. ICSLP, Philadelphia, PA, pp. 2387-2390, Oct. 1996.

[5] Hunt A. J. and Black A.W., "Unit selection in a concatenative speech synthesis system using a large speech database," in Proc. ICASSP, Atlanta, GA, pp. 373-376, May 1996.

[6] Sagisaka Y., "Speech synthesis by rule using an optimal selection of nonuniform synthesis units," in Proc. ICASSP, New York, NY, pp. 679-682, Apr. 1988.

[7] Jon R.W. Yi.and James R. Glass, "Natural-sounding speech synthesis using variable-length units", ICLSP98.

[8] McDowell P.S and Datta S. "A fractal approach to the characterisation of speech ", Acoustic Letters, Vol.17, No.1, 1993.

[9] Maragos P. "Fractal Aspects of Speech Signals: Dimension and Interpolation", Proceedings IEEE international Conference on Acoustic Speech and Signal Processing (ICASSP), Vol.1, pp417-424, 1991.

[10] Fekkai S. and Al-Akaidi M. "A Novel approach to measure fractal dimension of speech phonemes", Euromedia, Belgium, 1999.

[11] Fekkai S. and Al-Akaidi M. "Word recognition based on fractal techniques ", Proceeding of international conferences on image, science, system and technology, Las Vegas, USA, pp589-595, 1999.

[12] Srinivas J. "Fractals classification, generation and application ", University of Texas, IEEE, vol. 2, pp.1024-1027 (1992).

[13] Klatt D., "Review of Text-to-Speech Conversion for English", Journal of theAcoustical Society of America, JASA vol. 82 (3), pp.737-793, 1987.

[14] Krger B., "Minimal Rules for Articulatory Speech Synthesis", Proceedings of EUSIPCO92 (1): 331-334, 1992.

[15] O'Saughnessy D., "Speech Communication - Human and Machine", Addison-Wesley, 1987.