Article:
Elliott, J and Atwell, ES (2000) Is anybody out there? the detection of intelligent and generic language-like features. JBIS: Journal of the British Interplanetary Society, 53 (1/2). 13 - 22. ISSN: 0007-084X

Preprint of paper to appear in JBIS: Journal of the British Interplanetary Society, Vol. 53, No. 1/2, pp. 13-22, 2000

IS ANYBODY OUT THERE? THE DETECTION OF INTELLIGENT AND GENERIC LANGUAGE-LIKE FEATURES

John Elliott and Eric Atwell
School of Computer Studies, University of Leeds, lS2 9JT.
Email: jre@scs.leeds.ac.uk and eric@scs.leeds.ac.uk

Abstract

The authors present an overview of their language-detection research to date, along with considerations for further research. The research focuses on the unique structure of communication, seeking to identify whether a given signal has features within it that display intelligence or language-like characteristics, and comparing this with current methods used in searches for extra-terrestrial intelligence, in particular the SETI (Search for Extra Terrestrial Intelligence) Institute's Project Phoenix.

Project Phoenix looks for signals within a pre-defined bandwidth, on the basis that if they occur it would indicate a source of intelligence outside our own. In this active research area, the reported research looks beyond this for patterns in a signal which should indicate if intelligence is present by applying formulated algorithms and using tailor-made software which will sense if similar structures exist. The objective is therefore to investigate algorithms that will accomplish this goal. The research reported concentrates on ascertaining whether inter-species communication displays generic attributes that distinguish it from other sources, such as music and white noise. First contact may come from eavesdropping on radio broadcasts of their own natural language.

1. Introduction

Radio waves are easy to generate, easy to detect and even a very backward technological civilisation should stumble over radio relatively early in their exploration of the physical world [1]. Radio is therefore a natural candidate for any advanced civilisation to use as a deliberate beacon. It cannot be assumed that aliens would communicate using perfect American English (or even British English) so it is unlikely that an extra-terrestrial message could be deciphered should one be received. Therefore, as a starting point, a suite of programs needs to be developed to analyse digital input and to extrapolate whether or not language-like structures are present in aural (or written) media. Also, basic algorithms need to be constructed to cross-check and satisfy assessment criteria, including ruling out reflected signals from Earth. Once any ETI signal had been analysed, it would then seem sensible to treat it much as one would an email - return their message with one of our own introducing ourselves in our own language.

In the first instance, it would seem reasonable to look first at different varieties of human language - and also at other intelligent species that share our world and perform similar linguistic tasks with apparent equal ease - to ascertain if there are underlying patterns and structures common to all. This data could then form the basis for denoting intelligence and language-like features, which then could be applied to identifying any such signal, whether terrestrial, or extra terrestrial.

As inter-species comparators with human language, birds and dolphins have been chosen because of their perceived ability to learn language beyond the innate imperative cries they are born with. Dolphins and birds have individual signatures and variations between obviously close family groups [2], and also reflect our own developed social structures. Both birds and dolphins represent alternative advanced communicators on our own planet for

which comparisons can be drawn to represent independent development. Dolphins especially represent an alien intelligence on our own planet as they display such traits as advanced social behaviour, neotency (extended childhood) and self-awareness [3].

2. Baseline Assumptions

It has been assumed that:

- vocalised communication in alien life forms would, as with mammals, be subject to breathing rhythms which control wordlength and breaks;
- the sounds made would be such that the receiver (ear) could understand and cope with the sounds made having passed through an air-like medium;
- Zipf's principle of least effort applies to written and verbal communication, ie where the two forces of unification and diversification achieve a vocabulary balance which includes in-built redundancy for avoidance of misinterpretation;
- a small number of symbols (the average modern alphabet has 23) can be used to encode an infinite number of combinations;
- radio is our best chance of interstellar communication considering the vast distances involved, and that the waves travel at the speed of light, are easy to generate, easy to detect, and relatively free of the absorption and noise that plagues other areas of the spectrum.

3. Background

3.1 Human Language

Despite national and regional diversity, many conventions have evolved independently but consistently. For example, 87% of languages use either subject/object/verb or subject/verb/object ordering (from a survey based on 402 languages [4]) and most alphabets vary only marginally from the mean letter make-up of 23.

3.2 Bird Song

Like humans, although established well before we ever trod the Earth, birds developed their own form of communication. In bird song, individual notes are meaningless: it is the sequence, rhythm and intonation that are all important. Similarly, apart from one or two exceptions, in humans a single sound such as equates to a single letter utterance is meaningless. Therefore, sound segments (notes) fit into an overall rhythm and intonation pattern.

The language of bird song runs parallel to our own at a fundamental level and may well be evidence for structural universals. Birds have a double-barrelled system giving two distinct layers on their communication [2].

(i) Innate sounds - the calls for danger and congregation, which are in-built from birth.

(ii) Songs - these are, in comparison, far more complex, have form and rhythm and have to be learnt.

A direct comparison can be seen in our own language:

(iii) In-built sounds - cries of alarm and distress.

(iv) Speech (language) which is again far more complex and has to be learnt.

This double layering or duality places birds above animals which only display grunts and cries which do not display the structure of learnt systems. The need to communicate over distance without direct visual contact and a requirement to co-ordinate actions such as flocking and calls for mating may well underpin our similarities.

Birds also exhibit other language-like traits, such as regional dialects, which suggest a similar world view emergence to our own. Young birds develop a sub-song during the learning period - like our own children babble - and during this sensitive period of early critical learning they are subject to significant repercussions if separated from their instructors.

Bird vocabulary is limited to courtship, repelling trespassers etc, but shows how quite different

species can develop parallel systems of communication independently.

3.3 Dolphins

Humans share the Earth with at least two other intelligent species [5]. Number two in today's intelligence stakes is not the great apes but a group from a far older evolutionary branch - one which our first ancestors would have come but a poor second to. For approximately the past 35 million years, and up until the last two million years when modern humans began to emerge, dolphins and their relatives far exceeded the intelligence of all other animals.

It has recently been discovered that a huge impact occurred 35 million years ago at what is now known as Chesapeake Bay which wiped out many species and which also coincides with the rapid evolution of baleen and toothed whales, which include the forebears of dolphins. It also coincides with a sudden enormous increase in their brain size, something not seen again until humans.

Genetically we are no closer to dolphins than we are to big cats or rodents, and our intelligence evolved totally separately and much later. Dolphin brains are structurally very different to our own, in that the lobes that are used for language are constructed completely differently to those in humans, such that they could have evolved on a different planet [6]. Dolphins are therefore very useful as a comparator.

Their language - a form of clicks and whistles - conveys complex information, which can represent physical aspects, location and direction over great distances in some cases, and with no visual assistance. They are also capable of conveying many complex social communications but it is arguable how far this extends. It is apparent though that they use complex and truly social communication at many levels which displays a rhythm and structure akin to our own.

3.4 Evolutionary Imperatives

All vocalised communication in these life forms emanates from the mouth and is subject to breathing rhythms. The sounds made are such that a receiver (ear) can understand and cope with the sounds made having passed through an air-like medium. On that basis, it would seem reasonable to suppose that any alien would have developed parts of its body to sense the world, methods of communication to survive, exchanging and building on information in the short term but also passing on knowledge from one generation to the next.

3.5 Minimal Effort

Generally speaking, individuals will tend to use the least effort possible to achieve their goal. This is an obvious natural instinct, as its purpose is to optimise and save unnecessary effort. Communication is no different except that here a compromise has to be made due to at least two parties being involved.

The two forces are ***unification*** and ***diversification*** **[7]**. ***Unification*** is the force of the speaker's economy, who ideally would prefer to convey all meaning in a single word. The opposing force of ***diversification*** is that of the recipient or auditor, who in achieving total understanding of a communication would prefer to have every distinct meaning having a single distinct word. These two extremes of one endeavouring to reduce and the other expand the size of vocabulary are hypothetically latent in speech and serve to achieve a vocabulary balance which includes in-built redundancy for avoidance of misinterpretation.

This natural process is yet another, which serves to universalise the form communication, takes in its structure and is a prime move in the way we communicate in every day language. These forces are refined into a shared code-book.

3.6 A Shared Code-book

For a system to function effectively, the parties involved have to share a common but evolving code-book. This is where an internally held representation of an event is converted into semantic structures, then converted to an articulating code for uttering. Once received, these utterances are translated to an internal representation and decoded into a meaningful message, which can then result in a reply.

The content of these code-books increase with familiarity and become ways of reducing lines of communication, sometimes to mere gestures that can convey a whole message.

The idea that a shared code-book is a universal imperative for language is basic to this research. The existence of a shared code-book is further evidence that languages comprise a similar content for conveying information, have similar abilities for receiving and decoding, and in-built redundancy.

4. The Method

A method is required to extrapolate and analyse the content of a given signal, sound or digital sequence to ascertain if language-like features exist within.

However, our perceptions of patterns can be a barrier. Something, which has patterns according to its own rules, may not be understood unless at some point it displays patterns that we can interpret. Nevertheless, because language in all its developed and more advanced forms displays remarkably similar attributes, it was assumed that this holds universally across intelligent communications, such that the underlying rhythms and structures of communications will display patterns and signatures that closely correlate and will separate them from non-language-like phenomena. Nature is resplendent with patterns: it seems almost a pre-requisite that if something is natural it is comprised of patterns. Therefore language should display a common frequency signature which is found by varying the way in which a signal or communication is segmented to provide a key to further analysis.

To investigate whether language having particular patterns and an overall signature will indicate if intelligence is present, a variety of samples were analysed. These covered two distinct formats:

1 Digital representation of written language

2 Digital representation of sound waves

With the digital representation of written language, examples from a variety of sources were tested to see if the underlying patterns held. Such examples were taken from Teutonic, Romance and Slavonic languages. As comparators to these, non-language sources such as music and image files were analysed as controls.

Sound waves provide the opportunity to widen the scope to encompass other species who use forms of communication. Sound samples were compared of speech, bird song and dolphin clicks and whistles. Again, comparators of non-language-like sources such as music and white noise samples were analysed.

5. Detection of Symbolic Language

As the first of the two formats, written language was analysed as bit stream segments. The following processes are believed to provide a comprehensive test and basis for deriving a suitable algorithm indicating the presence of its symbolic representation.

5.1 Zipf's Law

Zipf, through a large body of statistical data, attempted to show that language is subject to the over-riding law which he called 'the principle of least effort' [7]. Morse, in developing his code, recognised this and assigned to the most frequent letter - 'e' - a single symbol, and to the least frequent - 'z' - the longest symbol sequence. Languages themselves have developed along these lines where the most frequently used words are typically the shortest.

This is reflected in Zipf's Curve (Figure 1) for word length against frequency derived from empirical evidence and should be a useful additional indicator for the evidence of language-like features.

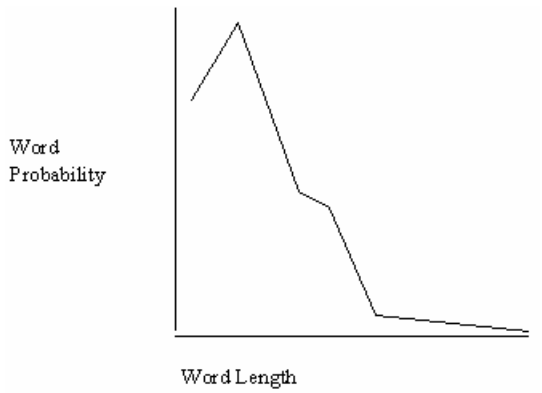

Figure 1: Zipf's Curve

5.2 Entropy

One useful measure of how much self-information there is in a set of symbols (or patterns) emitted from a source is its entropy.

Claude Shannon developed this as part of his mathematical Theory of Communication, which is now known as Information Theory. This entropy is measured in bits and in this case will be applied to the whole content of the messages analysed.

The lower the value of entropy, the more self-information is said to be present. The formula reads thus:

$$H = - \sum p(x_i) \log_2 p(x_i)$$

Where $\mathbf{x_i}$ is a particular pattern for which a probability measure is given, and

H is the average entropy value for the sample.

The summation gives the average value of self-information in the bit stream [8].

As linguists have empirically found that language is highly structured and communicates information, the entropy value of a bit stream when language is found should therefore be equal to its lowest value, or at least one which sharply drops off against the trend of surrounding values. This then - if true - should provide a useful indicator if a language-like structure is detected.

5.3 Compression

Another possible way of indicating if language-like structures are encoded in a digitised format is the degree to which it compresses. If a sample file is taken of unknown origin and compressed, the ratio of compressed to original file size should reflect the content of repeated patterns it contains.

In contrast to language data, digitally encoded images or just noise should compress at significantly different amounts.

From initial findings, language (text) files compress to approximately 50% of their original size. This is significantly different to that of noise and standard image formats which use 'lossy' compression algorithms for efficiency in transmission and storage, equating approximately to that of language redundancy in-built for purposes of communication so information is not lost. However, when using the file format 'rgb', which does not compress the image data, a similar result is obtained to that of language. Nevertheless, this does not detract from the validity of the testing, as such image data would be detected at later stages of analysis.

Therefore, as part of any algorithm for a general 'first pass' analysis to detect language-like content, a compression test is a worthwhile aid.

5.4 Initial Programs for Symbolic Communication Analysis

To analyse digitised input to ascertain if any language-like content is present in a written format, a number of programs were written to tackle differing aspects. The following are descriptions of the programs written to date, which form the basis of further research.

There are two main programs:

Program 1 calculates the frequency of pattern occurrences given a fixed string length. This length can be chosen at run-time to systematically search through all lengths required and deemed feasible. It does this by keeping a tally of how many times each encountered pattern of the bit length specified occurs. This data is then stored and also used for graphical representation.

Program 2 performs the same functions as above but also incorporates a 'sliding window' facility to cope with the real life situation of picking up a transmission after it has begun where, in addition to the problem of detecting if patterns exist that represent language, a legitimate 'take up' point mid-transmission needs to be found.

The program therefore takes information for the bit length and the offset from the file's first digit. By varying the two variables it will analyse all feasible ranges to see if language-like patterns exist. Obviously, given a bit length 'n' selected, all offsets analysed for this should not exceed 'n' and need only range from 0 up to 'n-1'. For example, if an 8-bit length is selected for

analysis and the offset is 1 displacing it the maximum number of places from all possible beginnings, 7 additional runs will only be required to sample these.

The secondary programs perform the following functions:

- Calculating notes (in binary) recorded by a particular instrument in a program written by W Towle [9] for later pattern analysis. This is to see if music can show similar patterns to human language-like findings from its representation in digital form.

- Calculating the number of symbol occurrences in a given text, for single or multiple patterns, with variable white space. This is to calculate individual (unigram) against n-gram probabilities within language and hypothesise to language-like frequency distributions. This program was developed for future use in continued research, as it may well contribute to areas pertaining to cognition and the possible 'alien' syllable.

- Applying Zipf's theory of word length frequency, as in [10]. This is done by taking the digital representation of a space, as identified from unknown sources by its predominant occurrences in comparison to any other pattern. It then looks at the patterns between the occurrences of the spaces, which should therefore equate to words and records the pattern frequencies against length. The statistics are then recorded and output to graph to see if they correlate to a Zipfian Curve.

- Extracting specific data, working with Program 2 above. A particular offset is selected as one of the several output by Program 2, then extracts the data from the pattern and frequency fields to display as histograms.

5.5 Chunking

The term 'chunking' was chosen to represent the process of segmenting a binary stream into discrete fixed lengths. These lengths can then be varied on subsequent passes for analysing all feasible encoding that the bit stream could contain, in order to ascertain if, at a given bit length or chunk, it displays language-like features.

ASCII character sets are generally encoded as 8-bit binary and therefore if a text file were received as a continuous bit stream, its language-like structure would be revealed when it was finally segmented into lengths of 8-bit.

However, any similar text-like communication of an unknown origin may be encoded at a completely different bit length, and so by systematically chunking the stream of digits of differing length the 'key' should be found.

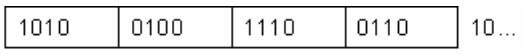

Figure 2: 4 bit chunking

The frequencies of all unique bit patterns that are found at a given length are then collated for statistical analysis.

5.6 Sliding Windows

As mentioned previously when describing the function of Program 2, any communication received cannot be guaranteed to be neatly captured from a suitable beginning. Using the first digit as a reference cannot be relied upon, and it will be necessary to vary the bit length in order to find the key to how the language is encoded - if such a key exists at all.

It is more likely that any transmission received will be captured from a completely meaningless point from which any reference will not provide a key. So, to compensate for this, a 'Sliding Window' system was developed. This, in addition to chunking at various given lengths as before, will also 'slide' or offset the reference origin from the first digit to all other possible origins given the chunking lengths specified. Using this method, any language-like structures may be detected that could have been missed

through the incorrect assumption that the first digit was a valid reference.

6. Initial Findings

The following are the main findings, detailed in [11]:

6.1 Entropy

Where language does occur, there is an uncharacteristic dip in entropy value against an otherwise linear upward trend. This ties in well with the notion that the lower the value of entropy, the more self-information is present with each unit.

6.2 Sliding Windows

This looks for all possible start points, as a signal is unlikely to be captured from the very beginning. Where language is present, a significant fluctuation in patterns occurs, with the anomaly that the number of patterns and their frequencies are identical at 0 and 1 offsets. This provides a key for identifying a suitable take-up point.

6.3 Music

Results on music where numerical values were given to notes with subsequent additions for octaves produced flat frequencies across patterns for each octave, which was a marked contrast to language-like results.

6.4 Images

The results show that for most of the frequencies recorded a near-flat distribution is found. Of the 120+ patterns, 90% of them lie in a frequency band which is only 35% of the total range, again displaying characteristics unlike those for language.

6.5 Latin-Based Language Samples

The program to segment patterns into varying bit lengths to search for a particular 'signature' for pattern distribution could be seen where language was detected. Histogram results confirmed a particular frequency curve or signature appears when language present in significant contrast to non-language samples. From this, it is easy to extrapolate where the spaces occur as they are always the most frequent pattern. This is a useful tool for later word segmentation and the use of Zipfian word distribution analysis. It can be concluded that the properties the curve displays should be a template for identifying if language is received in its symbolic form for a given input.

6.6 Algorithm

The following represents an initial algorithm for detecting symbolic communication (to be refined and developed in future research):

- Is sample compression rate c50%? If no compression, assign as 'noise' or compressed data which is characteristic of non-language like communication.
- Chunking text: does a particular chunk length display a language-like frequency signature after analysing all possible off-sets? If so, is there successive identical duplication on two offsets to establish reference?
- If candidate chunk length found, assign most frequent as space. Then apply Zipfian word length distribution analysis to intervening patterns.
- Does frequency distribution of pattern display type-token frequency which is out of character to trend?
- Does entropy value dip to low value out of step to general linear scale?
- If results show positive for most of the above assign as language-like communication.

7. Spoken Communication

This is the second and in some ways more likely of the two formats in which a signal would be detected. It is in this area that comparison can be made to other species and thereby analyse if common features or universals apply.

It should be noted that having received a candidate signal, unlike speech processing, no specific attempt is made to identify where words begin and end. The purpose is not to decipher but to identify the overall structure as language-like.

The intention is to look for breaks that occur between utterances; not their semantic implications, merely their structural ones.

7.1 Initial Programs for Sound-Wave Analysis

These programs perform the following functions:

- Analysing digitised data representing the waveform. Summary information envelopes are created of alternately positive and negative values with respect to the zero line, to provide the initial segmentation of waveform data. This is performed in a two-phase operation:

  Phase 1 looks at the amplitudes at the sampling rate provided, e.g. 10,000 per second (10 Khz), and performs the initial calculations for variance and segmentation into ranges which are either positive or negative. These are then committed to a temporary file due to high memory usage for such input as white noise.

  Phase 2 then calculates the duration and average amplitudes for the envelopes provided from Phase 1. This information is then stored in summary files.

- Merging the envelope values, which are concurrently, either above or below a given threshold value. It then creates the 'U' field data for the number of envelopes merged given the criteria. This then provides a time series of alternate unified envelopes for high amplitude activity and low-level amplitudes, which equate to pauses or background noise. Values are then re-calculated for all data in the preliminary tables to give the merged totals.

- Taking pre- or post-merged data from a summary file and extracting values for a given field. These values are then put in order for output to a graph file, then output to provide a histogram representation of the data.

- Storing calculations in temporary files for later use. In some cases, the need for memory exceeded maximum storage capacity, which caused the original program to fail, as the computer's memory was insufficient for coping with the volume of data. It therefore stores calculations in a temporary file from which they can then be taken, acting as an intermediary between input and output. One such example was when analysing white noise.

7.2 Analysing the waveform

The first step was to digitise the analogue waveform so that all pertinent quantifiable features within it could be extrapolated. Therefore:

(i) Any sampling below a given threshold should indicate pauses between significant activities, which should equate in language to phoneme-like segments.

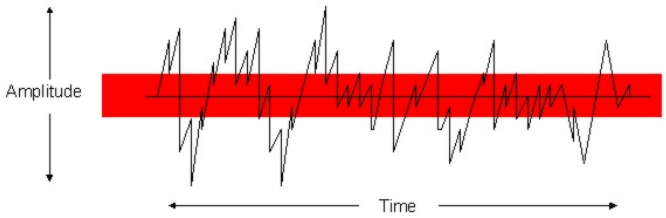

Figure 3: Vertical analysis

(ii) To catch the duration and rhythm of the soundwave; the internal structure of the waveform was analysed.

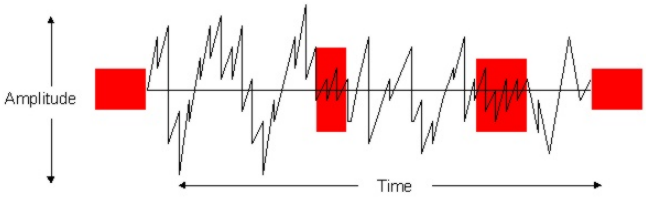

Figure 4: Horizontal analysis

Capturing and comparing the information in this way gives a comprehensive picture of the structure, which denotes if language-like features exist.

Measurements were made for: sampled amplitude at a given point; the number of samples in an envelope (the waveform period which remains constantly one side of zero); maximum amplitude in an envelope; and the approximate distance between samples.

From these initial measurements, additional calculations were made for the average amplitude of each envelope; the start and end sample of the envelope; the total amount of samples; the calculations made given the sampling rate used (e.g. 10,000 per second); and the variance or amount of distance (frequency shift) the wave travels in the envelope.

Preliminary statistics are taken of the sound wave so its precise structure can be adequately analysed when obtaining an overall picture.

Taking the information extrapolated from the digitised waveform when breaking it up, two additional features are introduced so as to build a clear picture. The first of these is the use of a threshold value. This threshold, which is set manually for the purposes of this project, is used as a mechanism to distinguish the Significant Activity Sections (SAS) from the periods of pause or comparative background noise occur which are essential to ascertaining overall structure.

The second is the by-product of thresholding, the value 'U' which is the unification of envelopes above and below the threshold according to the time frame.

8. Sound Sample Output Results

Detailed analyses are reported in [11]; the following are the main findings:

8.1 White Noise

Generally all values show a flat or near flat distribution equating to randomness and lack of structure. This lack of extremes was expected with such a source.

8.2 Human Speech

In direct comparison to white noise, analysis yields significant variations over a far greater range. Envelope durations mainly range from 600 - 2000 and, with the aid of histograms, show a rhythm and structure that are characteristic of bursts of sound with pauses, and can readily be identified as speech. The occurrences of amplitude show an 'A' shaped bi-symmetrical distribution around zero which is a marked difference to the flat distribution of noise.

8.3 Music

Durations of envelopes around zero were virtually non-existent. Where gaps occur between longer durations, these are of virtually no duration at all. The occurrences of amplitude show an opposite 'U' shaped distribution.

8.4 Dolphin Language

Statistics for both dolphins and humans show extremely similar 'A' shape graphs, symmetrical around zero. Durations of envelopes over time show a close correlation to humans, both in the regular structure of spikes to troughs and the frequency scale duration.

8.5 Bird Song

The occurrences of amplitude display a much simpler and less symmetrical 'A' shape distribution. Duration of envelopes found similar amounts of peak values seen over time showing similar rhythms to humans. Differences occurred where notes were held as part of the song. In human terms this would compare with singing.

8.6 Satellite Transmission

This sample is a good source of contrast within one transmission as speech occurs midway, but also before, after and to a lesser extent during mechanical clicks and tones. By looking at this and where they precisely occur in time it should be possible to separate them out and see their obvious structural differences.

Durations are now seen to reach far greater maximums. The mechanical nature of sounds created either side of speech are apparent as they occur at precise intervals and are of identical duration. Patterns that do not occur in natural phenomena are also observed.

However, where speech does occur a familiar rhythm of long and short durations within a limited range is seen. The clicks that occur during the time frame for speeds are obvious and are of durations far exceeding speech-like ones.

8.7 Initial Algorithm to Detect Speech

- Digitalise waveform received.
- Run amplitude occurrence analysis:
  If 'A' shape, this indicates that the sample is a language candidate;
  If 'U' shape, signal could possibly be music;
  If it is a flat distribution signal content is noise.
- Merge preliminary data with specified field and threshold.
- Select field 'd' for time-value analysis.
- Do 'd' values over time display values $n \leq x < m$?
- Apply 'difference' equation to window of 'd' values over time. If difference $n \leq x < m$ then
  $<$ = noise
  $>$ = non-language like

For segments which display candidate values, isolate for further analysis. This last stage is yet to be completed.

9. Future Considerations

The next steps will be to:

- Automate: implement the system at a receiver which will take in a signal, perform all stages automatically and state to the user if a candidate language-like signal is present.
- Devise a formula (equation of comparisons) for the comparison of segments over time to automatically recognise speech structure.
- Prosody: analyse the variation in pitch, within a significant activity of sound which equates to a phoneme-like segment, to see if patterns exist such as tone-tonic or particular pitch signatures over time.
- Investigate the possibility of a recurring pattern at several levels of language, using Natural Language Learning algorithms from Computational Linguistics, Artificial Intelligence, and Corpus Linguistics [12,13,14]. Rank frequency is a prime candidate for this fractal-like attribute.
- Ascertain if boundaries can be conclusively identified, even in novel utterances using such a low level structural analysis.

Although all the languages analysed and found to be consistent were all derived from the Latin alphabet, they do not all use the complete set of letters available and their combinations, structure and grammar are far from similar. As restrictions of available resources dictated such limitations, it is also worth stressing that languages - as with early tribes - evolve and eventually a global standard should develop.

The findings imply that if a language-like form of communication is present within a signal in its true and un-encoded format, then the algorithms above will provide a means for detecting its presence. SETI use the premise that any signal detected within a certain frequency and bandwidth will indicate a candidate for intelligence; this is likely to be an inadequate criterion, and I it would be better to widen the search and look also for structure.

This is only a beginning - a springboard: there are many more factors yet to consider in this area, some of which are outlined above. In tackling these, the intention is to move towards a fuller understanding of the generic structure of communication.

10. Conclusions

If a signal is received or intercepted which exhibits such traits as described, and whose source can be confirmed as outside our own, then these are good grounds for taking this as being indicative of intelligence 'alien' to human life.

Test results on language have shown that there is

a series of criteria by which its presence can be detected. This is due to a common requirement to use a finite set of symbols to represent an infinite variety of combinations for conveying information. In speech, physical limitations are part of the equation where our vocal apparatus restricts the set of possible outputs, their duration and amplitude. In written language, this set is similarly represented where letters generally equate to phonetic units and retain a similar overall structure due to previously discussed needs of minimal effort and a shared code-book.

Sound wave samples analysed for representative advanced communicators on this planet (birds, dolphins and humans) seems to display a common generic structure. This is most evident on results for occurrences of amplitude and values for durations over time which display statistics which are significantly different to the noise and music controls tested. It is concluded that any signal displaying such results should, if received, be recommended for further analysis on the basis of indicating intelligence and language-like structures.

Breaking the waveform up in such a way as to extract all the significant features provided an effective means for a fine-grained analysis. These, when presented graphically, illustrated some interesting and potentially significant results.

Language in its written format has proved to be a rich source for a variety of statistical analyses, some more conclusively than others, but when combined give a comprehensive algorithm for identifying the presence of language-like systems. Stages include compression, entropy, type-token distribution, word length Zipfian analysis and finding a frequency distribution signature by successive chunking.

Both dolphins and birds displayed characteristics, which indicate that a generic structure does exist and is therefore likely to for other advanced communicators. The structure also seems to refine itself as the complexity of the 'language' increases, whilst still retaining the overall signature.

It now seems most likely that if ever such a structure were to be detected it would be from an intelligent source, as:

(a) if natural, the criteria indicate the unique qualities of language-like structure; and

(b) if contrived, then an intelligent agent has created them.

In any case, in the event of a signal that shows language-like structures it is proposed to continue the dialogue by returning it with the addition of a simple message of our own, rather like using an e-mail where the sender's message is tagged on to the reply.

References

1. C. Sagan, 'Contact', Legend, London, 1988.

2. J. Aitchison, 'Seeds of Speech', Cambridge University Press, Cambridge 1996.

3. K. Marten, Project Delphis, Earthtrust, Hawaii, 1998 (letter to authors).

4. D. Crystal, 'Dictionary of Language and Languages', Penguin Books, London, 1992.

5. H. Couper and N. Henbest, 'Is Anybody Out There?', Dorling Kindersley, London, 1998.

6. H. Couper, Research findings sent via email, 1998.

7. G. K. Zipf, 'Human Behaviour and The Principle of Least Effort', Addison Wesley Press, New York, 1949 (1965 reprint).

8. C. E. Shannon and W. Weaver, 'The Mathematical Theory of Communication'. University of Illinois Press, Urbana, USA, 1964.

9. W. Towle, 'Ekaviel', Final year BSc project report, School of Computer Studies, University of Leeds, UK, 1995.

10. R. A. Sharman, 'An Introduction to the Theory of Language Models', IBM Lecture Notes, Winchester, July 1989.

11. J. Elliott, 'Is anybody out there? – The detection of intelligence and generic language-like features', Final year BSc project report, School of Computer Studies, University of Leeds, UK, 1998.

12. E. Atwell and N. Drakos, 'Pattern Recognition Applied to the Acquisition of a Grammatical Classification System from Unrestricted English Text' in B. Maegaard (ed), 'Proceedings of the Third Conference of European Chapter of the Association for Computational Linguistics', pp56-63, New Jersey, Association for Computational Linguistics, 1987.

13. J. Hughes and E. Atwell, 'The automated evaluation of inferred word classifications' in A. Cohn (ed), 'Proceedings of European Conference on Artificial Intelligence (ECAI)', pp535-539, Chichester, John Wiley. 1994.

14. E. Atwell, 'Machine Learning from Corpus Resources for Speech And Handwriting Recognition' in J. Thomas and M. Short (eds), 'Using Corpora for Language Research: Studies in the Honour of Geoffrey Leech', pp151-166, Longman, Harlow. 1996.