PAN ARCHIVES OF ACOUSTICS
Vol. 43, No. 4, pp. 593–602 (2018)

DOI: 10.24425/aoa.2018.125153

Speech and Music – Nonlinear Acoustical Decoding in Neurocognitive Scenario

Susmita BHADURI, Dipak GHOSH

Deepa Ghosh Research Foundation
Kolkata, West Bengal 700031, India; e-mail: susmita.sbhaduri@dgfoundation.in, deegee111@gmail.com

(*received May 24, 2017; accepted July 26, 2018*)

Speech and music signals are multifractal phenomena. The time displacement profile of speech and music signal show strikingly different scaling behaviour. However, a full complexity analysis of their frequency and amplitude has not been made so far. We propose a novel complex network based approach (*Visibility Graph*) to study the scaling behaviour of frequency wise amplitude variation of speech and music signals over time and then extract their PSVG (*Power of Scale freeness of Visibility Graph*). From this analysis it emerges that the scaling behaviour of amplitude-profile of music varies a lot from frequency to frequency whereas it's almost consistent for the speech signal. Our left auditory cortical areas are proposed to be neurocognitively specialised in speech perception and right ones in music. Hence we can conclude that human brain might have adapted to the distinctly different scaling behaviour of speech and music signals and developed different decoding mechanisms, as if following the so called *Fractal Darwinism*. Using this method, we can capture all non-stationary aspects of the acoustic properties of the source signal to the deepest level, which has huge neurocognitive significance. Further, we propose a novel non-invasive application to detect neurological illness (here autism spectrum disorder, ASD), using the quantitative parameters deduced from the variation of scaling behaviour for speech and music.

Keywords: speech signal; multifractality; Visibility Graph; Fractal Darwinism; neurocognitive disorders.

1. Introduction

Music and speech are the most cognitively complex phenomena created by sound for human beings. We can easily differentiate between speech and music just by listening to the signals for a few seconds. Speech contains a large variety of complex sounds with varying temporal grain, periodic and aperiodic components, noise, frequency, and amplitude modulations etc. On the contrary music is much more difficult to be decoded with respect to its acoustic features as it has richer frequency content than speech. Spectral envelope, duration, fundamental frequency, etc. are the main acoustic features of both speech and music, with respect to the perception. The formants of the spectral envelope are critical cues for speech, whereas spectral envelope itself is the main feature identifying the timbre of musical sound (Wolfe, 2002). As for analysis of the acoustic features from the neurocognitive perspective, it has been established that relatively good speech comprehension could be acquired with even two spectral channels, demonstrating that the temporal variation contained within these two noise bands were enough to allow the speech decoding mechanism to function sufficiently (Shannon *et al.*, 1995). Further research has proved that various aspects of speech decoding depend largely on the left auditory cortical regions (Hickok, Poeppel, 2000). However, extracting significant acoustic cues from music is difficult for human beings, as music is much richer acoustically. Samson and Zatorre (1994) have shown that musical timbre perception depends on systems of neural structure inside the right temporal lobe. Patients with damages of the right temporal cortex showed deficiency in discriminating musical timbre. Unlike speech, music does not have a fixed semantic system and it may convey meaning through emotional appraisal and associative memories (Trost *et al.*, 2012). Zatorre *et al.* (1992; 1994) have confirmed that comparative specialisation within the right auditory regions for tonal processing is substantiated by functional imaging data from a wide variety of melodies in musical compositions. Most of the frequency domain features are extracted from a sound spectrogram and as per Joos (1948).

The Acoustic Uncertainty Principle specifies that one cannot make a precise simultaneous measurement of an auditory event in both the time and frequency domains. So there always exists a spectral-temporal approximation in a spectrogram of sound signal. Zatorre *et al.* (2002) have argued that to address this *acoustic uncertainty*, the auditory cortices in the two hemispheres might have become comparatively specialised, such that temporal resolution is better in the left auditory cortical areas and spectral resolution is better in the right auditory cortical areas. They proposed that these cortical asymmetries might have developed as a general solution to the requirement to optimise processing of the acoustic environment in both temporal and frequency domains.

It has already been proved that the human brain is a complex and chaotic system constructed over multiple scales of space and time and the signals generated from the various lobes of brain are nonlinear and non-stationary (Babloyantz *et al.*, 1985; Bullmore, Bassett, 2011). All the organs of human body behave nonlinearly due to their inherent complex dynamic nature. The processes of speech production and cognition by human beings are complex phenomena (Proctor, Van Zandt, 2008). Musical compositions are also complex systems (Vaggione, 2001). The theory of complexity is rooted in chaos theory (Poincaré, 1889) and has various parameters whose combined behaviour refers to the border between order and randomness, termed as *the edge of chaos* (Horgan, 1995). As per chaos theory, a chaotic system is extremely sensitive to initial conditions, does not repeat itself, however, it is deterministic. The chaos-based complexity theory attempts to decode behaviour of dynamic nonlinear systems (Gallagher, Appenzeller, 1999; Mikulecky, 2001; Higgins, 2002). To provide order or definite properties to a structural form inherent in the chaotic system, *fractal geometry* has been evolved (Peitgen *et al.*, 2004). According to Mandelbrot (1967; 1983), *fractal* is a geometric scheme which repeats itself at smaller or larger scales to generate *self-similar*, irregular shapes, or surfaces that cannot be represented by Euclidean geometry. *Fractal* systems can extend to infinitely large values of their coordinates, in all directions from the centre towards the outside. The principal feature of *fractals* is their *self-similarity*. It is a phenomenon where smaller and bigger fragments of a system look very alike to but not necessarily exactly the same as the whole *fractal* system. *Power law* (as per statistics, a *power law* is a functional relationship between two quantities where one quantity varies as a power of another) is applied to represent the self-similarity of the large and small fragments of a *fractal* system. This *power law* exponent is defined as the *scaling exponent* of the self-similarity or the *fractal dimension* of the system. *Fractals* are of two types: monofractals and multifractals. Scaling properties of the monofractals are the same in different regions of the system, whereas scaling properties of multifractals are different in different regions of the systems (Chen *et al.*, 2002).

If the time series is long range correlated, its DFA function shows a *power law* relationship with its scale parameter. If we denote the DFA function of the time series by $F(s)$ and its scale parameter by s, $F(s)$ will vary with a power of s as per the equation $F(s) \propto s^H$, where the exponent H is termed as *Hurst exponent*. If D_F is the *fractal dimension*, it is related with H-Hurst exponent as per the equation $D_F = 2 - H$ (Kantelhardt *et al.*, 2001). MF-DFA (Kantelhardt *et al.*, 2002) method has the highest precision in the scaling analysis. Results obtained by DFA and MFDFA methods are proved to be more reliable compared to the methods like Wavelet Analysis, Discrete Wavelet Transform, Wavelet Transform Modulus Maxima, Detrending Moving Average, Band Moving Average, Modified Detrended Fluctuation Analysis etc. (Oświcimka *et al.*, 2006; Serrano, Figliola, 2009; Huang *et al.*, 2011). We have applied MF-DFA method successfully for analysing various kinds of time series formed from natural signals like speech signals (Bhaduri *et al.*, 2016) and biological signals like EEG and ECG signals (Bhaduri, Ghosh, 2015; 2016a; Nilanjana *et al.*, 2016; Bhaduri *et al.*, 2017).

The speech production process exhibits *fractal* characteristics. The quasi-static oscillations of the vocal folds and the adaptation process of the vocal tract are both nonlinear processes (Levelt, 1999). Multifractal nature of speech has been explored for automatic speech recognition (Maragos, Potamianos, 1991), speaker recognition (Gonzalez *et al.*, 2012), speech decomposition (Langi *et al.*, 1997), speech segmentation, representation, and characterisation (Kinsner, Grieder, 2008). Music is traditionally defined as an ordered arrangement of sounds of varying acoustic frequencies (pitches, tones) in succession (melody), of sounds in combination (harmony), and of sounds spaced in temporal succession (rhythm) (Hsü, Hsü, 1990). Mandelbrot (1983) defined *scaling noise* as a certain kind of sound whose quality stays unaltered even with changing play speed. *White noise* is the most simple *scaling noise*. The power spectral density, say denoted by $S(f)$, of a time series produced in agreement with the temporal variation of *white noise* varies with frequency content, say denoted by f, as per the equation $S(f) \propto f^{\beta}$, where β is the scaling exponent (Oświęcimka *et al.*, 2011). Voss and Clarke (1975) were the first to do *fractal* analysis of music and showed that it is *pink noise* or $1/f$ noise. Tricot (1988) implemented *fractal* theories on self-affine functions, and found a *power law* relationship between the power spectra and the *fractal* dimension. Recently some work about multifractal analysis of music was reported by Su and Wu (2006)

and Jafari *et al.* (2007). Hence, fractality and multifractality of speech and music have already been established.

Considerable amount of work has been done to devise automatic speech-music signal classification system using the conventional acoustic features. Most of these methods deal with time domain features like Zero Crossing Rate (Panagiotakis, Tziritas, 2005), Short Time Energy (El-Maleh *et al.*, 2000), and frequency domain features like signal bandwidth, spectral centroid, signal energy (Cohen *et al.*, 1995; Mckay, Fujinaga, 2004), fundamental frequency (Wold *et al.*, 1996), Mel-Frequency Cepstral Co-efficients (MFCC) (Harb, Chen, 2003). Most of these conventional stationary techniques involve Fourier spectral analysis which is based on linear superpositions of trigonometric functions. Secondary harmonic components, which are common in natural non-stationary time series, may generate a distorted wave outline for these natural signals. These distortions are the consequence of nonlinear contributions which are not normally extracted from the non-stationary signals, when analysed using these stationary techniques.

We should define a speech-music classification system by analysing speech and music as complex system using state of the art methods in *fractal* domain, in contrast to the conventional stationary techniques. This way all aspects of speech and music signal can be understood at the deepest level. In our earlier work (Bhaduri, Ghosh, 2016b), we have applied MFDFA method to the time-displacement profile of speech (non-musical), drone (periodically musical), and Indian art music samples with different musicality and showed that the value of the width of the multifractal spectrum is substantially different for speech and music signals. In another work (Bhaduri *et al.*, 2016), we have applied the same approach over speech signal and proposed a quantitative parameter for categorising various emotions by analysing the non-stationary details of the dynamics of speech signal, generated out of differing emotions. A non-invasive system has been proposed using this parameter for early detection of Alzheimer's disease. However, both DFA and MFDFA methods mandate that the data series in question should be of *infinite* length, which is a difficult scenario in most of the real-life situations, hence we have adopted an absolutely different, meticulous method – *Visibility Graph analysis* (Lacasa *et al.*, 2008; 2009), discussed in detail in Subsec. 2.1 in (Bhaduri *et al.*, 2016) and implemented a modified version of this method to analyse time displacement profile of speech signal generated out of contrasting emotions of anger and sadness, effectively classified them according to their emotional content and proposed the framework for assessing suicidal tendency of the subjects of experiment. However, these approaches do not explore the frequency properties of speech and music audio signals.

Considering the advantages and disadvantages of our earlier attempts as well as the drawbacks of DFA, MFDFA methods, we have approached both frequency and power properties of the speech and music signal from a totally different perspective of visibility network analysis in this work. Sporns *et al.* (2005) have suggested the concept of *connectome* to define the network of anatomical connections linking the neuronal elements of the human brain. Various approaches based on graph theory have been developed to investigate the human brain *connectome*, either in normal or diseased state. Here, we have implemented the most rigorous and state-of-the-art method of *Visibility Graph* to analyse the audio signals of speech and music (drone) and proposed finer-level acoustic cues for differentiating speech and music signals. These quantitative cues eventually establish the models proposed many times from the neurocognitive perspective that speech and music are decoded differently in two hemispheres of the human brain, as they are found to be completely different in terms of their acoustic contents as well as complexity, in this experiment. Our left auditory cortical areas are comparatively specialised in speech perception and right ones are in music (Zatorre *et al.*, 2002). Binnig *et al.* (2002) have proposed the concept of *Fractal Darwinism* which states how fractality of multiple complex systems adapt with each other according to their degree of self-similarity. In this work, we have attempted to establish how different perception mechanism of speech and music by human brain, as proposed by Zatorre *et al.* (2002), might have evolved from the different scaling pattern inherent in speech and music signals, as if following the so called *Fractal Darwinism*.

Neurocognitive disorders involve deterioration of cognitive abilities like memory, problem solving, perception, judgement, singing, speech, etc. These disorders result from temporary or permanent damage to the brain, degenerative processes like Alzheimer's or Parkinson's disease, dementia, and also from affective disorders like depression, pathological anxiety, and even bipolar disorder, autism, and dyslexia (Ganguli *et al.*, 2011). Based on the parametric cues found for all aspects of speech and music signals in this work and earlier ones (Bhaduri, Ghosh, 2016b; Bhaduri *et al.*, 2016a; 2016), we can model non-invasive applications for assessment of various neurocognitive disorders. During the last decade, there is increasing interest in applying music as a therapeutic tool in neurocognitive rehabilitation. Varnet *et al.* (2015) have analysed various effects of music over brain and how musical training imparts better cognitive abilities. Using our parametric cues, we can also implement a quantitative basis for existing music-therapeutic approaches for neurocognitive disorders.

Here, we have experimented with various kinds of speech signals generated from both male and female voiced speech and the drone signal which can be generated by effortlessly playing drone, as drone signal is the most basic and the simplest form of music signal (Van der Merwe, 1989). Based on findings of our analysis we have proposed the roadmap of an exemplary and novel, non-invasive and real-time application for early detection and monitoring of autism spectrum disorder or ASD.

The rest of the paper is organised as follows. The method of *Visibility Graph* technique, the details of data, our analysis, and the inferences from the test results are presented in Sec. 2. The inferences are discussed and an example of application for detecting and monitoring autism spectrum disorder has been elaborated in Sec. 3.

2. Methods

2.1. Visibility Graph Algorithm

Lacasa *et al.* (2008; 2009) have used fractional Brownian motion (fBm) and fractional Gaussian noises (fGn) series as their theoretical framework to study real time series in various scientific fields. They showed how a conventional method of complex network analysis can be implemented to measure long-range dependence and *fractality* of a time series.

The algorithm is a one-to-one mapping from the domain of time series X to its *Visibility Graph*. Let X_i be the i-th point of the time series. This way all the input data points are mapped to their corresponding nodes or vertices (according to their value or magnitude). In this node series, two nodes, say X_m and X_n, corresponding to m-th and n-th points in the time series, are said to be connected via a bidirectional edge if and only if Eq. (1) is valid. This way, the Visibility Graph is constructed out of a time series X

$$X_{m+j} < X_n + \left(\frac{n-(m+j)}{n-m}\right)\cdot(X_m - X_n), \qquad (1)$$

where $\forall j \in Z^+$ and $j < (n-m)$.

As shown in Fig. 1, the nodes X_m and X_n, with $m = i$ and $n = i+6$, can see each other, if Eq. (1) is satisfied for them. With this logic two sequential points of the time series can always see each other hence all sequential nodes are connected together. The time series should be converted to positive planes as the above algorithm is valid only for positive X values in the time series.

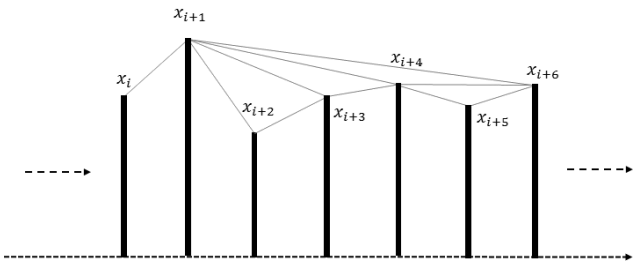

Fig. 1. *Visibility Graph* for time series X.

2.1.1. Power of scale-freeness of VG-PSVG

As per the graph theory, the definition of the degree of a node is the number of connections or edges that the node has with other nodes. The degree distribution, say $P(k)$, of a network formed from the time series is defined as the fraction of nodes with degree k in the network. Thus, if there are n nodes in total in a network and n_k of them have degree k, then $P(k) = n_k/n$.

The *scale-freeness* property of *Visibility Graph* states that the degree distribution of its nodes satisfies a *power-law*, i.e. $P(k) \sim k^{-\lambda_p}$, where λ_p is a constant and it is known as *Power of the Scale-freeness in Visibility Graph – PSVG*, which is denoted by λ_p and is calculated as the gradient of $\log_2[P(k)]$ *versus* $\log_2[1/k]$ plot. λ_p corresponds to the amount of complexity and *fractal* nature of the time series indicating the *Fractal Dimension* of the signal (Lacasa *et al.*, 2008; 2009; Ahmadlou *et al.*, 2012). It is also proved that there exists a linear relationship between *PSVG*-λ_p and *Hurst exponent*-H of the associated time series (Lacasa *et al.*, 2009). This method has recently been applied widely over time series with *finite* number of data points, even with 400 data points (Jiang *et al.*, 2013), and achieved authentic results in various domains of science.

2.2. Data description

Audio clips used in our previous work (Bhaduri, Ghosh, 2016b) are used in this experiment. The speech, drone music samples of duration of 160 seconds, are in .wav format. Sampling frequency for the data is 44.1 KHz. Samples are encoded by 16 bit-stream and of type mono. The amplitude waveform is taken for the testing. We have used the empirical mode decomposition method as per Huang *et al.* (1998) for noise removal from the original signal. Speech signals consist of both male and female voices, the language spoken in the samples is English. The drone signal used here is purely instrumental, of classical genre, and played by multiple classical artists. Here we have taken drone signal as the music signal because it is the most basic and the simplest form of music signal (Van der Merwe, 1989).

2.3. Data analysis

Following are the steps of our method.

1) First we calculate the power spectrum for each of the audio clip. As per Wiener-Khinchin theorem, power spectrum of a signal is the Fourier trans-

form of its autocorrelation function. For deterministic signals, the power spectrum is the magnitude squared of its Fourier transform. If we denote power spectrum by $S(f)$, then $S(f) = |X(f)|^2$, where $X(f)$ is the Fourier transform of the time-displacement profile of the signal $x(t)$ as per the below equation

$$X(f) = \int_{-\infty}^{\infty} x(t)\, \mathrm{e}^{-2\pi i f t}\, \mathrm{d}t. \tag{2}$$

For each of the samples of speech and drone signals, power spectral components for the range of frequencies 0.02–20 kHz, which is the audible range of frequencies for human beings (Rosen, Howell, 2010), are extracted.

2) After this we extract first 20 high strength frequencies according to their power in the power spectrum, for both speech and drone audio samples. Then spectrogram is generated for each audio sample, by computing 1024 point FFT with 50% overlap and using a Hamming window (Fulop, Fitz, 2006). As we know, the spectrogram is based on the Short-Time Fourier Transform, where the input signal is broken into chunks and on each chunk Fourier Transform is applied. If we extract the information for a particular frequency from the spectrogram over time, we get its magnitude in each chunk over time. Here we assume that in a particular chunk in the spectrogram, the amplitude of the specific frequency is constant.
3) For each of the speech and drone audio files, we extract the amplitude variation of the first 20 strongest frequencies extracted from its power spectrum, over time of progression of the audio file. The amplitude variation for each of these 20 frequencies is extracted from the spectrogram generated from the corresponding audio file. Figures 2a and 2b show the variation of amplitude over time for particular frequency for speech and drone signals, respectively.
4) Then for each of these amplitude profiles, we have constructed *Visibility Graphs* as per the method described in Subsec. 2.1. Then the values of k *versus* $P(k)$ are calculated for the *Visibility Graphs* corresponding to each of the 40 time series (20 for the speech + 20 for the drone). The k *versus* $P(k)$ plots for the time series for a sample, each from speech and drone signals, are shown in Figs. 3a and 3b, and the *power law* relationship is evident here.
5) *Power of Scale freeness in Visibility Graph* (*PSVG*) – λ_p value is calculated from the slope of $\log_2[1/k]$ *versus* $\log_2[P(k)]$ for each audio file, as per the method in Subsec. 2.1. Plot of $\log_2[1/k]$ *versus* $\log_2[P(k)]$ for the same k *versus* $P(k)$ series is shown in Fig. 3c for the speech with $\lambda_p = 3.23$ and Fig. 3d for the drone with $\lambda_p = 2.92$.

2.4. Results

After calculation of all the PSVG-λ_p values for 40 samples (20 for the speech + 20 for the drone), we have plotted their trend over the 20 high-strength frequencies, as shown in Figs. 2c and 2d for the speech and drone, respectively. It is very interesting to observe that the scaling exponent of the dominant frequencies

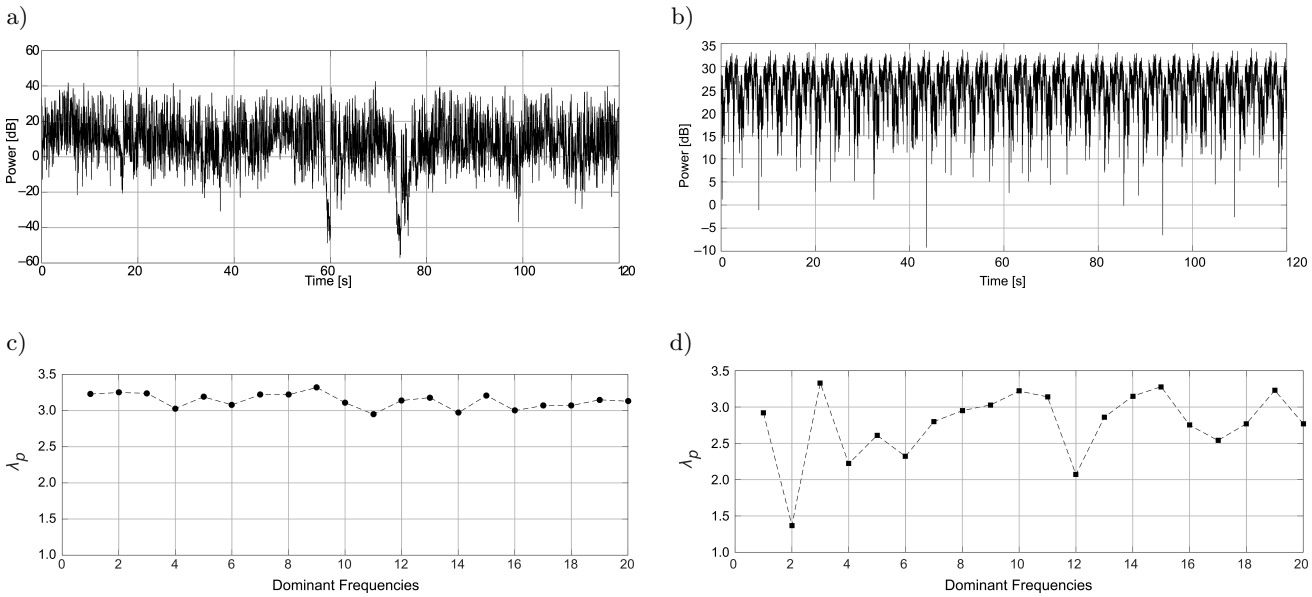

Fig. 2. a) Amplitude variation over time for speech signal for a particular frequency, b) amplitude variation over time for drone signal for a particular frequency, c) trend of λ_p values for first 20 dominant frequencies for a speech audio sample, d) trend of λ_p values for first 20 dominant frequencies for a drone audio sample.

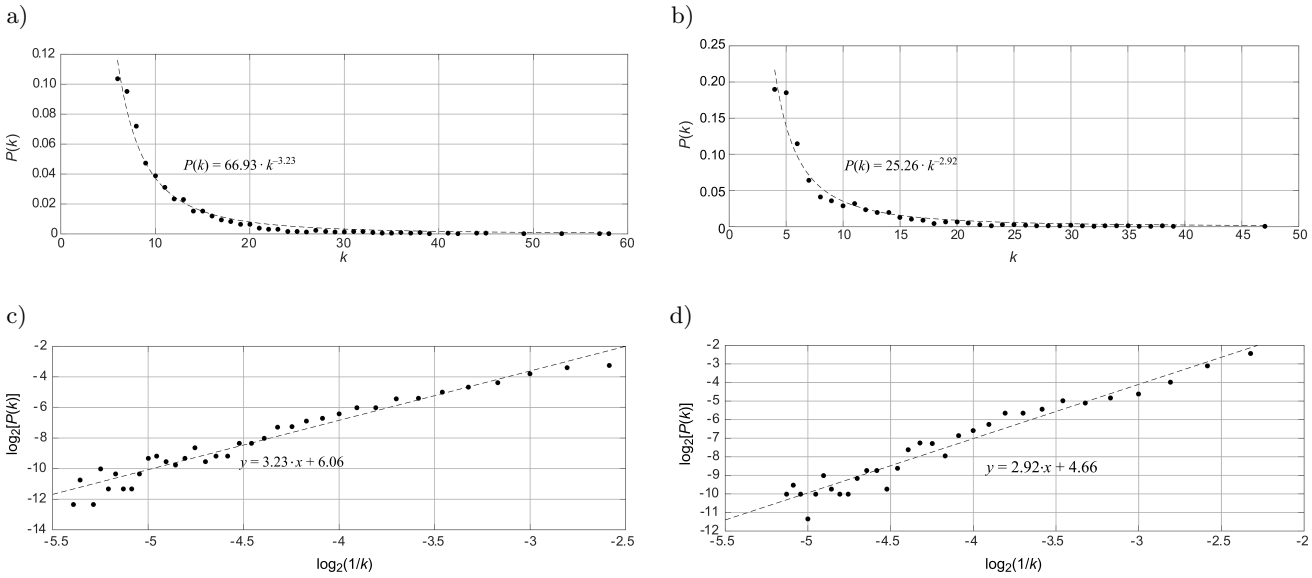

Fig. 3. a) k *versus* $P(k)$ trend for the *Visibility Graph* of amplitude profile for the speech signal for a particular frequency, b) k *versus* $P(k)$ trend for the *Visibility Graph* of amplitude profile for the drone signal for a particular frequency, c) $\log_2[1/k]$ *versus* $\log_2[P(k)]$ calculated for the amplitude profile for the same frequency of the speech signal, d) $\log_2[1/k]$ *versus* $\log_2[P(k)]$ calculated for the amplitude profile for the same frequency of the drone signal.

vary within a comparatively smaller range of values for the speech signal than that of drone or the most basic music signal. It is evident from the figures that speech and music can be clearly segregated by the variance of their scaling behaviour from frequency to frequency.

3. Discussion

It is evident from Figs. 2a and 2b that the amplitude profile for a dominant frequency in the speech signal is steadily irregular over time. Whereas, for the drone signal, the amplitude varies almost pseudo-periodically over time. This observation is almost consistent for all 20 dominant or high-strength frequencies for both speech and music samples.

Good scaling behaviour is evident from *power law* fitting for the k *versus* $P(k)$ plots calculated for the *Visibility Graphs*, for the time series of both speech and drone signals, are shown in Figs. 3a and 3b. The same is also seen from the straight line fitting calculated in $\log_2[1/k]$ *versus* $\log_2[P(k)]$ series in Figs. 3c and 3d. Hence, we can confirm that for each dominant frequency, the amplitude variation over time obeys the scaling law, for both speech and music.

Finally, in Figs. 2c for the speech and 2d for the drone, we can see that for the drone signal, the scaling behaviour of amplitude varies a lot from frequency to frequency, whereas it's almost consistent for the speech signal. Hence, we can conclude that the human brain might have been adapted to the different scaling behaviour of the acoustic signal of speech and music, as if by following the theory of *Fractal Darwinism* (Binnig *et al.*, 2002) and developed different decoding mechanisms for speech and music – left auditory cortical areas for specialised speech perception and right ones for music (Zatorre *et al.*, 2002).

From the above observations it can be summarised that if a naturally generated audio signal shows a lot of variation in scaling behaviour from one dominant frequency to the next one, and if this variation is greater than certain threshold (say, denoted by δ), then we can confirm that the signal is of music and not speech. In other words, an audio signal needs to have this variation greater than δ to be qualified as music or to be converted to music. This threshold (δ) should be defined by analysing the multifractal properties of time displacement profile using the method in (Bhaduri, Ghosh, 2016b) as well as variation of scaling pattern of amplitude (frequency wise) using the proposed method, for large number of speech and drone samples. Using this threshold (δ), various neurocognitive applications for detecting neurological illness, autism spectrum disorder, disorder of consciousness etc., can be devised. As an example, we have broadly outlined a framework for one such application for detection of autism spectrum disorder in the Subsec. 3.1.

3.1. Proposed exemplary application

As already mentioned in the Sec. 1, neurocognitive disorders involve cognitive impairment restricting proper emotional expression in speech and singing, memory problems, issues involving problem solving, perception, judgment etc. and these disorders result from temporary or permanent damage to the brain,

degenerative processes like Alzheimer's or Parkinson's disease, dementia, as well as from affective disorders like depression, pathological anxiety, and even from bipolar disorder, autism, and dyslexia (Ganguli *et al.*, 2011).

In this work, we propose a quantitative framework to capture the change in intricate dynamics of speech produced by a normal subject or drone played by the same subject and the same signals analysed for a subject suffering from autism spectrum disorder (ASD). Here for music signal, the drone is chosen as a musical instrument as it's the most effortlessly played instrument that generates a very basic form of music signal (Van der Merwe, 1989). According to the steps elaborated below, three types of threshold (δ) values are calculated. One is for normal subjects, the second type for subjects who have already been diagnosed with ASD, and the third one is for any subject to be diagnosed for ASD. Then, depending upon the proximity of the third one to the first and second ones, proneness or onset of ASD can be decided.

- First, speech and drone signals generated by a large number of normal subjects, are to be collected. Then, after doing the multifractal analysis of time displacement profile using the method in (Bhaduri, Ghosh, 2016b) and analysis of the scaling pattern of amplitude profile for high strength frequencies as per proposed here method, for both speech and drone signals, the threshold δ for normal subjects, say denoted by δ_{norm}, can be base-lined. δ_{norm} reflects the degree to which the speech and music signals generated by normal subjects can be discriminated. This would be the first control element for this application for detection and monitoring of ASD.
- Similarly, audio clips of the speech and drone signals generated by the subjects already diagnosed with ASD, are to be recorded. People with ASD process information differently than normal people in their brain. It's already been shown in Sec. 1 that human brain acts as a complex system and that music and speech are complex phenomena. Also they all have fractal characteristics. Hence, the speech produced and the simplest music generated by an autistic subject would definitely display different scaling behaviour in all acoustic aspects, than those of normal subjects. Hence, using the same method of Step 1, the second control element, say denoted by δ_{ASD}, for diseased subjects having ASD, can be base lined.

  Different ranges of parameters between δ_{norm} and δ_{ASD} may be defined, to reflect the proneness or the severity of ASD. One sample set of ranges is given below.

  1) One range for deciding whether the subject to be diagnosed is at all prone to ASD or not.
  2) Second range for deciding the onset of ASD.
  3) Third for prognosis of ASD.
- Finally, the speech and the drone sample generated by the subject to be assessed for ASD, would be collected, and then similar scaling analysis for both kinds of signal as elaborated in Step 1, is to be done. This way, the threshold for the subject of the experiment, say denoted by δ_{exp}, is calculated. As per the range (defined in Step 2), the proximity of δ_{exp} towards δ_{norm} and δ_{ASD} would be checked. According to the range where δ_{exp} falls, the absence, onset, or the severity of ASD can be assessed, also in a non-invasive manner.

We can frame an uncomplicated, lightweight application for routine check-up, where we can locally save the control elements and calculate δ_{exp} on a real time basis. As we propose this to be a routine check-up model, this procedure will be an ongoing one which can set an alarm parameter for any deviation of δ_{exp} from its predefined normal range and we can monitor and accordingly forecast the onset of ASD at early stages. Eventually we propose to validate the conjectures using a larger database of speech and drone signals collected from normal as well as diseased subjects and devise a simple android application for routine check-up for non-invasive self-assessment as well as monitoring of ASD or other neurocognitive disorders.

Acknowledgments

We thank the Rabindra Bharati University and Department of Higher Education, Govt. of West Bengal, India for logistic support of the computational analysis.

References

1. Ahmadlou M., Adeli H., Adeli A. (2012), *Improved visibility graph fractality with application for the diagnosis of autism spectrum disorder*, Physica A: Statistical Mechanics and its Applications, **391**, 20, 4720–4726.
2. Babloyantz A., Salazar J.M., Nicolis C. (1985), *Evidence of chaotic dynamics of brain activity during the sleep cycle*, Physics Letters A, **111**, 3, 152–156, doi: 10.1016/0375-9601(85)90444-X.
3. Bhaduri A., Bhaduri S., Ghosh D. (2017), *Visibility graph analysis of heart rate time series and biomarker of congestive heart failure*, Physica A: Statistical Mechanics and its Applications, **482**, 786–795, doi: 10.1016/j.physa.2017.04.091.
4. Bhaduri A., Ghosh D. (2016a), *Quantitative assessment of heart rate dynamics during meditation: An ECG based study with multi-fractality and visibility graph*, Frontiers in Physiology, **7**, 44, doi: 10.3389/fphys.2016.00044.

5. Bhaduri S., Chakraborty A., Ghosh D. (2016), *Speech emotion quantification with chaos-based modified visibility graph-possible precursor of suicidal tendency*, Journal of Neurology and Neuroscience, **7**, 3, 100, doi: 10.21767/2171-6625.1000100.

6. Bhaduri S., Das R., Ghosh D. (2016), *Non-invasive detection of Alzheimer's disease – multifractality of emotional speech*, Journal of Neurology and Neuroscience, **7**, 2, 84, doi: 10.21767/2171-6625.100084.

7. Bhaduri S., Ghosh D. (2015), *Electroencephalographic data analysis with visibility graph technique for quantitative assessment of brain dysfunction*, Clinical EEG and Neuroscience, **46**, 3, 218–223, doi: 10.1177/1550059414526186.

8. Bhaduri S., Ghosh D. (2016b), *Speech, music and multifractality*, Current Science (00113891), **110**, 9, 1817–1822, doi: 10.18520/cs/v110/i9/1817-1822.

9. Binnig G., Baatz M., Klenk J., Schmidt G. (2002), *Will machines start to think like humans? Artificial versus natural Intelligence*, Europhysics News, **33**, 2, 44–47, doi: 10.1051/epn:2002202.

10. Bullmore E.T., Bassett D.S. (2011), *Brain graphs: graphical models of the human brain connectome*, Annual Review of Clinical Psychology, **7**, 113–140, doi: 10.1146/annurev-clinpsy-040510-143934.

11. Chen Z., Ivanov P.C., Hu K., Stanley H.E. (2002), *Effect of nonstationarities on detrended fluctuation analysis*, Physical Review E: Statistical, Nonlinear, and Soft Matter Physics, **65**, 4, 041107–041122, doi: 10.1103/PhysRevE.65.041107.

12. Cohen M.A., Grossberg S., Wyse L.L. (1995), *A spectral network model of pitch perception*, The Journal of the Acoustical Society of America, **98**, 2, 862–879.

13. El-Maleh K., Klein M., Petrucci G., Kabal P. (2000), *Speech/music discrimination for multimedia applications*, [in:] 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 00CH37100), Istanbul, Turkey, Vol. 4, pp. 2445–2448, doi: 10.1109/ICASSP.2000.859336.

14. Fulop S.A., Fitz K. (2006), *Algorithms for computing the time-corrected instantaneous frequency (reassigned) spectrogram, with applications*, The Journal of the Acoustical Society of America, **119**, 1, 360–371, doi: 10.1121/1.2133000.

15. Gallagher R., Appenzeller T. (1999), *Beyond reductionism*, Science, **284**, 5411, 79, doi: 10.1126/science.284.5411.79.

16. Ganguli M. *et al.* (2011), *Classification of neurocognitive disorders in DSM-5: a work in progress*, The American Journal of Geriatric Psychiatry: Official Journal of the American Association for Geriatric Psychiatry, **19**, 3, 205–210.

17. González D.C., Ling L.L., Violaro F. (2012), *Analysis of the multifractal nature of speech signals*, [in:] Alvarez L., Mejail M., Gomez L., Jacobo J. [Eds], *Progress in pattern recognition, image analysis, computer vision, and applications, CIARP 2012, Lecture Notes in Computer Science*, Vol. 7441, pp. 740–748, Springer, Berlin, Heidelberg, doi: 10.1007/978-3-642-33275-3_91.

18. Harb H., Chen L. (2003), *Robust speech music discrimination using spectrum's first order statistics and neural networks*, [in:] Proceedings of Seventh International Symposium on Signal Processing and Its Applications, Vol. 2, pp. 125–128, doi: 10.1109/ISSPA.2003.1224831.

19. Hickok G., Poeppel D. (2000), *Towards a functional neuroanatomy of speech perception*, Trends in Cognitive Sciences, **4**, 4, 131–138, doi: 10.1016/S1364-6613(00)01463-7.

20. Higgins J.P. (2002), *Nonlinear systems in medicine*, Yale Journal of Biology and Medicine, **75**, 5–6, 247–260.

21. Horgan J. (1995), *From complexity to perplexity*, Scientific American, **272**, 6, 104–109.

22. Hsü K.J., Hsü A.J. (1990), *Fractal geometry of music*, Proceedings of the National Academy of Sciences of the United States of America, **87**, 3, 938–941.

23. Huang N.E. *et al.* (1998), *The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis*, Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences, **454**, 1971, 903–995, doi: 10.1098/rspa.1998.0193.

24. Huang Y.X., Schmitt F.G., Hermand J.P., Gagne Y., Lu Z.M., Liu Y.L. (2011), *Arbitrary-order Hilbert spectral analysis for time series possessing scaling statistics: Comparison study with detrended fluctuation analysis and wavelet leaders*, Physical Review E: Statistical, Nonlinear, and Soft Matter Physics, **84**, 1, 016208, doi: 10.1103/PhysRevE.84.016208.

25. Jafari G.R., Pedram P., Hedayatifar L. (2007), *Long-range correlation and multifractality in Bach's Inventions pitches*, Journal of Statistical Mechanics: Theory and Experiment, **2007**, 04, P04012.

26. Jiang S., Bian C., Ning X., Ma Q.D. (2013), *Visibility graph analysis on heartbeat dynamics of meditation training*, Applied Physics Letters, **102**, 25, 253702, doi: 10.1063/1.4812645.

27. Joos M. (1948), *Acoustic phonetics*, Language, **24**, 2, 5–136, doi:10.2307/522229.

28. Kantelhardt J.W., Koscielny-Bunde E., Rego H.H., Havlin S., Bunde A. (2001), *Detecting long-range correlations with detrended fluctuation analysis*, Physica A: Statistical Mechanics and its Applications, **295**, 3–4, 441–454, doi: 10.1016/S0378-4371(01)00144-3.

29. Kantelhardt J.W., Zschiegner S.A., Koscielny-Bunde E., Havlin S., Bunde A., Stanley H.E.

(2002), *Multifractal detrended fluctuation analysis of nonstationary time series*, Physica A: Statistical Mechanics and its Applications, **316**, 1–4, 87–114, doi: 10.1016/S0378-4371(02)01383-3.

30. Kinsner W., Grieder W. (2008), *Speech segmentation using multifractal measures and amplification of signal features*, Proceedings of the 7th IEEE International Conference on Cognitive Informatics, ICCI 2008, Vol. 1, pp. 351–356.

31. Lacasa L., Luque B., Ballesteros F., Luque J., Nuño J.C. (2008), *From time series to complex networks: The visibility graph*, Proceedings of the National Academy of Sciences, **105**, 13, 4972–4975, doi: 10.1073/pnas.0709247105.

32. Lacasa L., Luque B., Luque J., Nuño J.C. (2009), *The visibility graph: A new method for estimating the Hurst exponent of fractional Brownian motion*, EPL (Europhysics Letters), **86**, 3, 30001, http://stacks.iop.org/0295-5075/86/i=3/a=30001.

33. Langi A.Z.R., Soemintapura K., Kinsner W. (1997), *Multifractal processing of speech signals*, Proceedings of ICICS, 1997 International Conference on Information, Communications and Signal Processing. Theme: Trends in Information Systems Engineering and Wireless Multimedia Communications (Cat. No. 97TH8237), Vol. 1, pp. 527–531, doi: 10.1109/ICICS.1997.647154.

34. Levelt W.J.M. (1999), *Models of word production*, Trends in Cognitive Sciences, **3**, 6, 223–232.

35. Mandelbrot B.B. (1967), *How long is the coast of Britain? Statistical self-similarity and fractional dimension*, Science, **156**, 3775, 636–638, doi: 10.1126/science.156.3775.636.

36. Mandelbrot B.B. (1983), *The fractal geometry of nature*, American Journal of Physics, **51**, 286, doi: 10.1119/1.13295.

37. Maragos P., Potamianos A. (1999), *Fractal dimensions of speech sounds: Computation and application to automatic speech recognition*, The Journal of the Acoustical Society of America, **105**, 3, 223–232.

38. Mckay C., Fujinaga I. (2004), *Automatic genre classification using large high-level musical feature sets*, Proceedings of the International Society of Music Information Retrieval Conference, ISMIR 2004, Vol. 1, pp. 525–530.

39. Mikulecky D.C. (2001), *The emergence of complexity: Science coming of age or science growing old?*, Computers and Chemistry, **25**, 4, 341–348.

40. Nilanjana P., Anirban B., Susmita B., Dipak G. (2016), *Non-invasive alarm generation for sudden cardiac arrest: a pilot study with visibility graph technique*, Translational Biomedicine, **7**, 3, doi: 10.21767/2172-0479.100079

41. Oświęcimka P., Kwapień J., Celińska I., Drożdż S., Rak R. (2011), *Computational approach to multifractal music*, arXiv preprint arXiv:1106.2902, http://arxiv.org/abs/1106.2902.

42. Oświęcimka P., Kwapień J., Drożdż S. (2006), *Wavelet versus detrended fluctuation analysis of multifractal structures*, Physical Review E: Statistical, Nonlinear, and Soft Matter Physics, **74**, 1, 016103, doi: 10.1103/PhysRevE.74.016103.

43. Panagiotakis C., Tziritas G. (2005), *A speech/music discriminator based on RMS and zero-crossings*, IEEE Transactions on Multimedia, **7**, 1, 155–166.

44. Peitgen H.-O., Jürgens H., Saupe D. (2004), *Chaos and fractals*, New York, NY: Springer.

45. Poincaré H. (1890), *On the problem of three bodies and equations of dynamics* [in French: *Sur le problème des trois corps et les équations de la dynamique*], Acta Mathematica, **13**, 1, A3–A270, doi: 10.1007/BF02392506.

46. Proctor R.W., Van Zandt T. (2008), *Human factors in simple and complex systems*, Taylor and Francis, CRC Press, Boca Raton.

47. Rosen S., Howell P. (2010), *Signals and systems for speech and hearing*, BRILL.

48. Samson S., Zatorre R.J. (1994), *Contribution of the right temporal lobe to musical timbre discrimination*, Neuropsychologia, **32**, 2, 231–240.

49. Serrano E., Figliola A. (2009), *Wavelet leaders: A new method to estimate the multifractal singularity spectra*, Physica A: Statistical Mechanics and its Applications, **388**, 14, 2793–2805.

50. Shannon R.V., Zeng F.G., Kamath V., Wygonski J., Ekelid M. (1995), *Speech recognition with primarily temporal cues*, Science, **270**, 5234, 303–304, doi: 10.1126/science.270.5234.303.

51. Sporns O., Tononi G., Kötter R. (2005), *The human connectome: a structural description of the human brain*, PLoS Computational Biology, **1**, 4, e42, doi: 10.1371/journal.pcbi.0010042.

52. Su Z.Y., Wu T. (2006), *Multifractal analyses of music sequences*, Physica D: Nonlinear Phenomena, **221**, 2, 188–194.

53. Tricot C. (1988), *Dimension fractale et spectre*, Journal De Chimie Physique, **85**, 379–384.

54. Trost W., Ethofer T., Zentner M., Vuilleumier P. (2012), *Mapping aesthetic musical emotions in the brain*, Cerebral Cortex, **22**, 12, 2769–2783.

55. Vaggione H. (2001), *Some ontological remarks about music composition processes*, Computer Music Journal, **25**, 1, 54–61, doi: 10.1162/014892601300126115.

56. Van der Merwe P. (1989), *Origins of the popular style : the antecedents of twentieth-century popular music*, Clarendon Press.

57. Varnet L., Wang T., Peter C., Meunie F., Hoen M. (2015), *How musical expertise shapes*

speech perception: evidence from auditory classification images, Scientific Reports, **5**, 14489, doi: 10.1038/srep14489.

58. Voss R.F., Clarke J. (1975), *'1/fnoise' in music and speech*, Nature, **258**, 317–318, doi: 10.1038/258317a0.

59. Wold E., Blum T., Keislar D., Wheaten J. (1996), *Content-based classification, search, and retrieval of audio*, IEEE Multimedia, **3**,3, 27–36.

60. Wolfe J. (2002), *Speech and music, acoustics and coding, and what music might be 'for'*, Proceedings of the 7th International Conference on Music Perception and Recognition, Sydney 2002, Vol. 6, pp. 10–13.

61. Zatorre R.J., Belin P., Penhune V.B. (2002), *Structure and function of auditory cortex: music and speech*, Trends in Cognitive Sciences, **6**, 1, 37–46.

62. Zatorre R.J., Evans A., Meyer E. (1994), *Neural mechanisms underlying melodic perception and memory for pitch*, The Journal of Neuroscience: The Official Journal of The Society for Neuroscience, **14**, 4, 1908–1919.

63. Zatorre R.J., Evans A.C., Meyer E., Gjedde A. (1992), *Lateralization of phonetic and pitch processing in speech perception*, Science, **256**, 846–849, doi: 10.1126/science.1589767.