CLAIMS,

1. Multi-modal systems for voice-based mental health assessment with emotion stimulation, said system comprising:
   - a task construction module configured to construct tasks for capturing acoustic, linguistic, and affective characteristics of speech of a user;
   - a stimulus output module (202) configured to receive data from said task construction module, said stimulus output module (202) comprising one or more stimuli, basis said constructed tasks, to be presented to a user in order to elicit a trigger of one or more types of user behaviour, said triggers being in the form on input responses;
   - response intake module (204) configured to present, to a user, the one of more stimuli, basis said constructed tasks, from the stimulus output module (202), and, in response, receive corresponding responses in one or more formats;
   - a features' module comprising:
     - a feature constructor configured to define features, for each of said constructed task, said features being defined in terms of learnable heuristic weighted tasks, said feature constructor being a Geneva Minimalistic Acoustic Parameter Set (GeMAPS) based feature constructor, said feature constructor using Higher order spectra (HOSA) functions, said functions being functions of two or more component frequencies achieving bispectrum frequencies;
     - a feature extractor configured to extract one or more defined features from said receive corresponding responses correlative to said constructed tasks, with learnable heuristic weighted model

considering at least one selected from a group consisting of ranked constructed tasks;

- a feature fusion module configured to fuse two or more defined features in order to obtain fused features;

- an autoencoder to define relationship/s, using said fused features, between:
  - an audio modality, of said feature fusion module, working in consonance with said response intake module (204) extract high-level features, in said responses, so as to output extracted high-level text features; and
  - a text modality, of said feature fusion module, working in consonance with said response intake module (204) to extract high-level features, in said responses, so as to output extracted high-level audio features;

said autoencoder configured to receive extracted high-level text features and extracted high-level audio features, in parallel, from said audio modality and said text modality to output a shared representation feature data set for emotion classification correlative to said mental health assessment.

2. The system as claimed in claim 1 wherein, said constructed tasks being articulation tasks and / or written tasks.

3. The system as claimed in claim 1 wherein, said task construction module comprises a first order ranking module configured to rank said constructed tasks in order of difficulty; thereby, assigning a first order of weights to each constructed task.

4. The system as claimed in claim 1 wherein, said task construction module comprises a first order ranking module configured to rank said constructed tasks in order of difficulty; thereby, assigning a first order of weights to each constructed task, in that, said construed tasks being stimuli marked with a ranked valence level, correlative to analysed responses, selected from a group consisting of positive valence, negative valence, and neutral valence.

5. The system as claimed in claim 1 wherein, said task construction module comprises a second order ranking module configured to rank complexity of said constructed tasks in order of complexity; thereby, assigning a second order of weights to each constructed task.

6. The system as claimed in claim 1 wherein, said task construction module comprises a second order ranking module configured to rank complexity of constructed tasks and affective expectation in terms of response vectors, of said ranked constructed tasks in order to create a data collection pool.

7. The system as claimed in claim 1 wherein, said constructed tasks being selected from a group of tasks consisting of cognitive tasks of counting numbers for a pre-determined time duration, tasks correlating to pronouncing vowels for a pre-determined time duration, uttering words with voiced and unvoiced components for a pre-determined time duration, word reading tasks for a pre-determined time duration, paragraph reading tasks for a pre-determined time duration, tasks related to reading paragraphs with phoneme and affective complexity to open-ended questions with affective variation, and pre-determined open tasks for a for a pre-determined time duration.

8. The system as claimed in claim 1 wherein, said constructed tasks comprising one or more questions, as stimulus, each question being assigned a question embedding with a 0-N vector such that a question-specific feature extractor is trained in relation to determination of embeddings' extraction, from said question, correlative to word-embedding, phone-embedding, and syllabl-level embedding, said extracted embeddings being and forced aligned for a mid-level feature fusion.

9. The system as claimed in claim 1 wherein, said one or more stimulus being selected from a group of stimuli consisting of audio stimulus, video stimulus, combination of audio and video stimulus, text stimulus, multimedia stimulus, physiological stimulus, and its combinations.

10. The system as claimed in claim 1 wherein, said one or more stimulus comprising stimulus vectors calibrated to elicit textual response vectors and / or audio response vectors and / or video response vectors and / or multimedia response vectors and / or physiological response vectors in response to the stimulus vectors.

11. The system as claimed in claim 1 wherein, said one or more stimulus being parsed through a first vector engine (232) configured to determine its constituent vectors in order to determine a weighted base state in correlation to such stimulus vectors.

12. The system as claimed in claim 1 wherein, said one or more responses being selected from a group of responses consisting of audio responses, video

responses, combination of audio and video responses, text responses, multimedia responses, physiological responses, and its combinations.

13. The system as claimed in claim 1 wherein, said one or more stimulus comprising response vectors correlative to elicited audio response vectors and / or elicited video response vectors in response to stimulus vectors of said stimulus output module (202).

14. The system as claimed in claim 1 wherein, said response intake module (204) comprising a passage reading module configured to allow users to perform tasks correlative to reading passages for a pre-determine time.

15. The system as claimed in claim 1 wherein, said feature constructor being a Geneva Minimalistic Acoustic Parameter Set (GeMAPS) based feature constructor for analysis of audio responses.

16. The system as claimed in claim 1 wherein, said feature constructor being a Geneva Minimalistic Acoustic Parameter Set (GeMAPS) based feature constructor, said feature constructor:
    - using a set of 62 parameters to analyse speech;
    - providing a symmetric moving average filter, 3 frames long, to smooth over time, said smoothing being performed within voiced regions, of said responses, for pitch, jitter, and shimmer;
    - applying arithmetic mean and coefficient of variation as functionals to 18 low-level descriptors (LLDs), yielding 36 parameters;
    - applying 8 functions to loudness;

- applying 8 functions to pitch;
- determining arithmetic mean of Alpha Ratio;
- determining Hammarberg Index;
- determining spectral features vide spectral slopes from 0–500 Hz and 500–1500 Hz over all unvoiced segments;
- determining temporal features of continuously voiced and unvoiced regions from said responses; and
- determining Viterbi-based smoothing of a F0 contour, thereby, preventing single voiced frames which are missing by error.

17. The system as claimed in claim 1 wherein, said feature constructor being a Geneva Minimalistic Acoustic Parameter Set (GeMAPS) based feature constructor, said feature constructor being configured with a set of low-level descriptors (LLDs) for analysis of the spectral, pitch, and temporal properties of said responses being audio responses, said features being selected from a group of features consisting of:
    - Mel-Frequency Cepstral Coefficients (MFCCs) and their first and second derivatives;
    - Pitch and pitch variability;
    - Energy and energy entropy;
    - Spectral centroid, spread, and flatness;
    - Spectral slope;
    - Spectral roll-off;
    - Spectral variation;
    - Zero-crossing rate;
    - Shimmer, jitter; and harmonic-to-noise ratio;

- Voice-probability (based on pitch); and
- Temporal features like the rate of loudness peaks, and the mean length and standard deviation of continuously voiced and unvoiced regions

18. The system as claimed in claim 1 wherein, said feature constructor being a Geneva Minimalistic Acoustic Parameter Set (GeMAPS) based feature constructor, said feature constructor being configured with a set of frequency related parameters selected from a group of frequency-related parameters consisting of:
    - Pitch, logarithmic F0 on a semitone frequency scale, starting at 27.5 Hz (semitone 0);
    - Jitter, deviations in individual consecutive F0 period lengths;
    - Formant 1, 2, and 3 frequency, centre frequency of first, second, and third formant;
    - Formant 1, bandwidth of first formant;
    - Energy related parameters;
    - Amplitude related parameters;
    - Shimmer, difference of the peak amplitudes of consecutive F0 periods;
    - Loudness, estimate of perceived signal intensity from an auditory spectrum;
    - Harmonics-to-Noise Ratio (HNR), relation of energy in harmonic components to energy in noiselike components;
    - Spectral (balance) parameters;
    - Alpha Ratio, ratio of the summed energy from 50–1000 Hz and 1–5 kHz;
    - Hammarberg Index, ratio of the strongest energy peak in the 0–2 kHz region to the strongest peak in the 2–5 kHz region;

- Spectral Slope 0–500 Hz and 500–1500 Hz, linear regression slope of the logarithmic power spectrum within the two given bands;
- Formant 1, 2, and 3 relative energy, as well as the ratio of the energy of the spectral harmonic peak at the first, second, third formant's centre frequency to the energy of the spectral peak at F0;
- Harmonic difference H1–H2, ratio of energy of the first F0 harmonic (H1) to the energy of the second F0 harmonic (H2); and
- Harmonic difference H1–A3, ratio of energy of the first F0 harmonic (H1) to the energy of the highest harmonic in the third formant range (A3).

19. The system as claimed in claim 1 wherein, said bispectrum using third-order cumulants to analyze relation between frequency components in a signal correlative to said responses for examining nonlinear signals.

20. The system as claimed in claim 1 wherein, said feature extractor fuses higher-level feature embeddings using mid-level fusion.

21. The system as claimed in claim 1 wherein, said feature extractor comprising a dedicated linguistic feature extractor, with learnable weights per stimulus, correlative to linguistic tasks.

22. The system as claimed in claim 1 wherein, said feature extractor comprising a dedicated affective feature extractor, with learnable weights per stimulus, correlative to affective tasks.

23. The system as claimed in claim 1 wherein, said feature fusion module comprising an audio module configured with high-level feature extractors

and at least an autoencoder-based feature fusion to classify emotions from one or more responses.

24. The system as claimed in claim 1 wherein, said feature fusion module comprising an audio module configured with high-level feature extractors and at least an autoencoder-based feature fusion to classify emotions from one or more responses, in that, said audio modality using a local question-specific feature extractor to extract high-level features from a time-frequency domain relationship in said responses so as to output extracted high-level audio features.

25. The system as claimed in claim 1 wherein, said feature fusion module comprising a text module configured with high-level feature extractors and at least an autoencoder-based feature fusion to classify emotions from one or more responses.

26. The system as claimed in claim 1 wherein, said feature fusion module comprising a text module configured with high-level feature extractors and at least an autoencoder-based feature fusion to classify emotions from one or more responses, in that, said text modality using a Bidirectional Long Short-Term Memory network with an attention mechanism to simulate an intra-modal dynamic so as to output extracted high-level text features.

27. The system as claimed in claim 1 wherein, said feature fusion module comprising an audio module:

- using extracted features in acoustic feature embeddings from pre-trained models;
- comparing said extracted features on a spectral domain;

- determining vocal tract co-ordination features;
- determining recurrent quantification analysis features;
- determining Bigram count features and bigram duration features corerelative to speech landmarks; and
- fusing said features in an autoencoder.

28. The system as claimed in claim 1 wherein, said feature extractor comprising a speech landmark extractor configured to determine event markers associated with said responses, said determination of said event markers being correlative to location of acoustic events, from said response, in time, said determination including determination of timestamp boundaries that denote sharp changes in acoustic responses, independently of frames.

29. The system as claimed in claim 1 wherein, said feature extractor comprising a speech landmark extractor configured to determine event markers associated with said responses, each of said event markers having onset values and offset values, said event markers being selected from a group of landmarks consisting of glottis-based landmark (g), periodicity-based landmark (p), sonorant-based landmark (s), fricative-based landmark (f), voiced fricative-based landmark (v), and bursts-based landmark (v), each of said landmark being used to determine points in time, at which different abrupt articulatory events occur, correlative to rapid changes in power across multiple frequency ranges and multiple time scales.

30. The system as claimed in claim 1 wherein, said autoencoder being configured with a multi-modal multi-question input fusion architecture comprising:

- at least an encoder for mapping one or more said characteristics with a task type to a lower-dimensional representation, each task being multiplied by said characteristics with a learnable weight-encoded matrix based on task type, said weights being correlative to mental health assessment;
- at least a decoder for mapping said lower-dimensional representation to said one or more said characteristics in order to output mental health assessment;
- said autoencoder being trained to minimize reconstruction error between input tasks and decoder output by using a loss function.

31. Multi-modal methods for voice-based mental health assessment with emotion stimulation, said method comprising the steps of:
    - constructing tasks for capturing acoustic, linguistic, and affective characteristics of speech of a user;
    - receiving data said constructed tasks, comprising one or more stimuli, basis said constructed tasks, to be presented to a user in order to elicit a trigger of one or more types of user behaviour, said triggers being in the form on input responses;
    - presenting, to a user, the one of more stimuli, basis said constructed tasks, and, in response, receiving corresponding responses in one or more formats;
    - defining features, for each of said constructed task, said features being defined in terms of learnable heuristic weighted tasks;
    - extracting one or more defined features from said receive corresponding responses correlative to said constructed tasks, with learnable heuristic

weighted model considering at least one selected from a group consisting of ranked constructed tasks;

- fusing two or more defined features in order to obtain fused features;
- defining relationship/s, using said fused features, between:
    - an audio modality, of said feature fusion module, working in consonance with said responses in order to extract high-level features, in said responses, so as to output extracted high-level text features; and
    - a text modality, of said feature fusion module, working in consonance with said responses in order to extract high-level features, in said responses, so as to output extracted high-level audio features;

said step of defining relationship/s comprising a step of receiving extracted high-level text features and extracted high-audio text features, in parallel, from said audio modality and said text modality to output a shared representation feature data set for emotion classification correlative to said mental health assessment.

32. The method as claimed in claim 31 wherein, said constructed tasks comprising one or more questions, as stimulus, each question being assigned a question embedding with a 0-N vector such that a question-specific feature extractor is trained in relation to determination of embeddings' extraction, from said question, correlative to word-embedding, phone-embedding, and syllabl-level embedding, said extracted embeddings being and forced aligned for a mid-level feature fusion.

33. The method as claimed in claim 31 wherein, said step of defining relationship/s comprising a step of configuring a multi-modal multi-question input fusion architecture comprising:
    - mapping, via an encoder, one or more said characteristics with a task type to a lower-dimensional representation, each task being multiplied by said characteristics with a learnable weight-encoded matrix based on task type, said weights being correlative to mental health assessment;
    - mapping, via a decoder, said lower-dimensional representation to said one or more said characteristics in order to output mental health assessment;
    - training to minimize reconstruction error between input tasks and decoder output by using a loss function.

34. The method as claimed in claim 31 wherein, said step of extracting high-level audio features comprising a step of extracting high-level features from a time-frequency domain relationship in said responses so as to output extracted high-level audio features.

35. The method as claimed in claim 31 wherein, said step of extracting high-level text features comprising a step of using a Bidirectional Long Short-Term Memory network with an attention mechanism to simulate an intra-modal dynamic so as to output extracted high-level text features.