Speech Features for Schizophrenia

Subjects: Engineering, Biomedical

Contributor: Felipe Lage Teixeira, Miguel R. Costa, José Pio Abreu, Manuel Reis, Salviano Pinto Soares, João Paulo Teixeira

It is known that speech and language provide unique and essential information about human thought. Speech in subjects with schizophrenia is perceived as a negative symptom because it is mainly reflected in a lack of emotion (blunted affect) and poor speech (alogia). Other speech symptoms in schizophrenia include slow speech, reduced pitch variability, more pauses, and less synchronization in syllable variability. Speech production in patients with schizophrenia is usually stimulated via clinical interviews, free speech activities, image description or reading. Free speech can be compromised in people with a diagnosis of schizophrenia. Therefore, techniques such as asking patients to report activities or plans for the future, tasks done the previous day, and dreams can be implemented. Narrative of Emotions Tasks can also be used during medical consultation.

Keywords: schizophrenia ; speech ; EEG ; features

1. State of the Art (Speech)

Considering what has been reviewed thus far, the works that use speech to diagnose schizophrenia are limited, which may be due to the difficulty in obtaining authorization to collect a speech dataset. Typically, the authors record the dataset used. However, within this limitation, most of the works, as in the case of [1][2], are based on speech in a natural context. Speech patterns in schizophrenia have also been analyzed in some works but to a lesser extent, as in the case of [2].

According to this study, the features used to identify schizophrenia are divided into four categories: prosodic, spectral, temporal, and statistical features. However, in addition to these characteristics, a quantitative measure, namely, the number of words (verbosity), was used.

Gosztolya et al. [3] used only temporal features obtained from spontaneous speech, such as articulation rate, speech tempo, duration of utterance, number of pauses, duration of pauses, pause duration rate, pause frequency, and average pause duration. The authors achieved 70–80% accuracy in classifying subjects with schizophrenia diagnosis [3].

Other authors used two categories of features. Kliper et al. [4] used temporal and prosodic features to identify schizophrenia, depression and control. The parameters used include spoken ratio, utterance duration, gap duration, pitch range, the standard deviation of pitch, power standard deviation, mean waveform correlation, mean jitter, and mean shimmer. These parameters allowed the classification of control vs. schizophrenia with an accuracy of 76.19%, control vs. depression with an accuracy of 87.5%, and schizophrenia vs. depression with an accuracy of 71.43%. For multiclass classification they achieved 69.77%.

Martínez-Sánchez et al. [5] and Rapcan et al. [6] showed that patients with schizophrenia tend to have slow speech, reduced pitch variability, and a more significant number of pauses. Rapcan et al. [6] investigated the fundamental frequency (F0) and the relative variation of vocal pitch and, using temporal and prosodic features, attempted to study the total speech duration but did not find a statistical significance and argued that the lack of academic qualifications of the subjects under analysis compromised the results.

Compton et al. [7] also used two categories of features but, in their case, used prosody and spectral categories. They studied schizophrenic patients and healthy subjects with and without aprosody. They concluded that patients with aprosody present lower F0, F2, and intensity/loudness values.

The severity of negative symptoms in the first outbreak of schizophrenia is correlated with the second-order formant F2. This conclusion was obtained after the study with fundamental frequency F0 and the first and second-order formants F1 and F2 [8].

He et al. [9] also detected negative symptoms using the parameters: symmetric spectral difference level (SSDL), quantification error and vector angle (QEVA), and standard dynamic volume value (SDVV), thus discriminating subjects with and without a diagnosis of schizophrenia with an accuracy of 98.2% (with decision trees).

Other authors used three categories of speech features. To identify cognitive and thought disorders. Voleti et al. [10] tried to find acoustic features of speech. These disorders include various neurological impairments (e.g., dementia) and psychiatric conditions (e.g., schizophrenia). Prosodic articulation temporal and vocal quality features were used. Temporal features include the duration of voiced segments and the duration of silent segments. The prosodic features covered loudness, periodicity measures, and F0. The spectral or articulation features comprise formant frequencies (F1, F2, and F3) and MFCCs. They also used jitter, shimmer, and harmonic-to-noise ratio (HNR) features.

Parola et al. [11] analyzed three categories of parameters, qualitative indices, quantitative analysis, and multivariate machine learning (ML) tools. Using ML, the results are more promising. For schizophrenia and healthy identification, free speech-based studies provide higher differences between groups.

Some authors used features of all four categories. Agurto et al. [12] could predict psychosis with 90% accuracy using prosodic, spectral, temporal, and statistical measures. The feature set was formed for spectral characterization by MFCCs, spectral slope, interquartile range (IQR), maximum energy, and frequency. For vowel characterization, they used F1, F2, and F3 (frequencies and their corresponding bandwidth). For voice quality, they used jitter (local absolute value and ppq5), shimmer (local absolute and apq5), autocorrelation, harmonic-to-noise ratio (HNR), and noise-to-harmonic ratio (NHR). For the rhythm changes, pauses (threshold of −25 dB and minimum duration of 100 ms) and voiced parts were considered. For each category mentioned above, the authors calculated the median, IQR, pct5, pct95, skewness, kurtosis, total interventions, speech rate, articulation rate, and speech/non-speech ratio (and corresponding percentages). In addition, they calculated speech rate/velocity of speech and articulation rate. These equations are indicators of cerebral activity, and, using them, it is possible to obtain a cerebral processing rate.

Tahir et al. [13] state that using a Multi-Layer Perceptron (MLP) neural network classifier allows an assessment of negative symptoms using the features of speaking rate, frequency, and volume entropy. The author also experimented with other types of features such as prosodic features (F0), spectral features including first, second, and third order formants (F1, F2, F3), MFCCs, amplitude (minimum, maximum, and mean volume), conversational/temporal features including duration of speech, speaking turns, interruptions, interjections, and statistical features such as entropy.

Similar to the aim of the previous study, Low et al. [14] concluded that features such as peak slope, linear predictive coefficients, and mean pause duration are directly correlated with schizophrenia. Quasi-open Quotient—QOQ, F1 range, articulation rate, pause rate, speech rate, time talking, and mean intensity are negatively correlated with schizophrenia. Moreover, the parameters including total number of pauses, mean speech duration, intensity variability, and F0 variability, among others, despite being used in many studies, do not show any correlation with schizophrenia.

Other authors used the semantic level of features. Mota et al. [15] evaluated the structural characteristics of each interview conducted so that each interview was converted into a graph in which each word is represented by a node and the temporal sequence between two words is represented by an edge (margin). The same procedure was performed every 30 consecutive words to analyze the verbosity. After this procedure, the authors evaluated the number of edges (margins) and the node connection. For semantic analysis, the median semantic distance between two sentences was calculated using latent semantic analysis (LSA). The authors stated that schizophrenia speech produces fewer linked words and less semantic coherence via the structural and semantic features.

On the other hand, for the prediction of psychotic outbreaks (in young people at high clinical risk—CHR), Bedi et al. [16] evaluated semantic and syntactic features. They detected two features in semantic coherence: the minimum semantic distance for first-order coherence (e.g., the minimum coherence or maximum discontinuity between two sentences) and the average semantic distance for first-order coherence (e.g., the average coherence between sentences). With the studied features, the authors could predict the development of psychosis with 100% accuracy.

The formal linguistic aspects of auditory verbal hallucinations (AVHs) indicate that speaking in the first person is less common in hallucinated speech. Sentences have no grammatical connectivity; speech has no connection and, usually, it is personalized. Thus, although there are individual variations, there is a linguistic profile of typical speech in people with verbal auditory hallucinations [17].

Some works combine speech acoustic features with text features. Xu et al. [18] transcribed the interviews (with software help), so it was possible to use speech and text parameters. The verbal speech parameters were LIWC, diction, Latent

Dirichlet Allocation, and Doc2vec features. The non-verbal speech parameters were composed of conversational, OpenSmile, and DisVoice elements, thus distinguishing diagnosed and undiagnosed subjects with an accuracy of 76.2% [18][19].

One work's authors [6] suggest that the lack of academic qualifications can compromise studies in this context. To increase performance, techniques could be applied as suggested in [20], in which speech is transcribed and text parameters are used simultaneously with the speech parameters.

Speech analysis was also combined with other parameters. In [1], the algorithm's performance increased when body movements were implemented as input parameters. For example, [1] applied low-level descriptors (LLD) and body movements to detect negative symptoms. The LLD set is composed of intensity, loudness, MFCC (12), pitch (F0), probability of voicing, F0 envelope, 8 LSF (Line Spectral Frequencies), and Zero-Crossing Rate. Using an SVM classifier with the LLD alone, the authors obtained an accuracy of 79.49%. If these features were combined with body movements, the accuracy improved to 86.36%.

Feature selection procedures were also implemented. To make a selection of the most promising parameters in the identification of schizophrenia via speech, Espinola et al. [21] used the Particle Swarm Optimization (PSO) method. Out of a set of 33 features, zero-crossing rate, Hjorth parameter complexity, average amplitude changes, mean absolute value, kurtosis, third and fourth moments, maximum amplitude, peak frequency, power spectrum ratio, mean, and total power (12 out of 33) were selected. With SVM, the authors reached an accuracy of 91.79% in classifying subjects with and without a diagnosis of schizophrenia.

Argolo et al. [19] concluded that structured interviews or task descriptions are the most commonly used for automated speech evaluation in these studies, similarly to studies based on free speech.

One of the most used machine learning tools is SVM, which has an accuracy rate between 70% and 91.79%. Using MLP, one author [13] obtained an accuracy of 81.3%. Utilizing Linear Discriminant Analysis (LDA), the authors of [6] achieved 79.4% accuracy. Using Signal Processing Algorithms, the authors of [5] achieved 93.8% accuracy in the discrimination between patients and controls. With decision trees, another author [9] obtained 98.2%. Lastly, the best accuracy achieved was obtained in the work of [16], which had an accuracy of approximately 100%, but for the prediction of psychotic outbreaks.

Although the set of previously analyzed features can indicate typical characteristics of schizophrenia, they do not identify schizophrenia exclusively. Other mental disorders or an anatomic deformation in the vocal tract can compromise these features. Therefore, the combination of several features is required for a schizophrenia diagnosis.

A summary of the features most used in the literature is presented in **Table 1**. The most frequently used speech parameters are divided into four main categories. The prosodic category features mostly used are F0, Intensity/Loudness/Amplitude, Jitter, and Shimmer. In the spectral category, the features more frequently used are frequency formants F1, F2, and F3 and MFCCs. The temporal features mostly used are utterance duration, the duration of pauses, and the number of pauses. For quantitative measures, some authors, such as [12][15][22], suggest that the number of pauses can be promising. Finally, the statistical features mostly used are the number of words and verbosity.

Table 1. Speech features used to identify schizophrenia.

Category of Feature	Feature	Work
Prosodic Characteristics	F0/Pitch	[1][5][6][7][8][10][13][20][23][24]
	Intensity/Loudness/Amplitude	[1][6][7][10][12][13][20][21][24]
	Jitter Shimmer	[4][10][12]
	HNR	[5]
	NHR	[12]
	Quantization Error and Vector Angle (QEVA); Standard Dynamic Volume Value (SDVV)	[9]
	Articulation rate	[18][25]
	Peak slope	[14]
Spectral Characteristics	MFCCs	[1][10][12][13][20]
	F1 F2	[7][8][10][12][13][20][24]
	F3	[10][12][13][20][21][24]
	Line Spectral Frequencies (LSF);	[23]
	Linear Predictive Coefficients (LPC)	[14]
	Symmetric Spectral Difference Level (SSDL)	[9]
	Zero-crossing rate	[4][26]
Temporal Characteristics	Duration of pauses	[3][5][6][10][11][14][27]
	Utterance duration	[3][4][6][11][13][20][28][29][30]
	Number of pauses	[3][6][12]
	Gap duration	[3][21]
	The proportion of silence	[10][11]
	Total recording time	[6]
	Voiced/unvoiced percentages; voiced/unvoiced ratio; velocity of Speech	[12]
	Quasi-open Quotient (QOQ)	[14]
	Number of words; verbosity (use of excessive words)	[15][27][29]
	Speaking turns, interruptions, and interjections	[2][10]
	Probability of voicing	[23]
	IQR (interquartile range) of MFCCs and F0 variation	[12]
Statistical Measures	Skewness and kurtosis (of log Mel freq. band); mean value (of waveform Correlation, jitter, and shimmer), slope sign changes	[4][27]
	Third, fourth, and fifth moments; Hjorth parameter activity; mobility and complexity; waveform length	[24]
	Minimum semantic distance for first-order coherence; mean semantic distance for first-order coherence	[16]
	Pitch range; standard deviation of pitch; power standard deviation; mean waveform correlation	[4]

Table 2 shows the parameters used by several authors organized according to the categories to which they belong. Not all of the authors mentioned in **Table 2** attempted to identify schizophrenia via speech; therefore, the “accuracy” column contains a short description. In the case of these authors, no accuracy was reported, and the table presents only

theoretical conclusions. In work [16], the authors achieved 100% accuracy but in classifying psychotic outbreaks in young people at CHR. Therefore, this work is excluded from this accuracy comparison.

Table 2. Accuracy of the speech features in the classification.

Number of Used Categories	Categories	Ref.	Accuracy (%)
1	Prosodic	[25]	To evaluate the relative contributions of motor and cognitive symptoms on speech output in persons with schizophrenia
		[26]	Language and thought disorder in multilingual schizophrenia
		[28]	Understanding constricted affect in schizotypal via computerized prosodic analysis,
	Temporal	[3]	80
		[11]	They identified weak untypicalities in pitch variability related to flat affect and stronger untypicalities in proportion of spoken time, speech rate, and pauses related to alogia and flat affect.
	Statistical	[30]	93.8% (emotion detection)
		[15]	They characterized the relationship between structural and semantic features, which explained 54% of negative symptoms variance.
		[2]	93
		[16]	100 (psychotic outbreaks in young people at CHR).
		[27]	87.56
2	Prosodic and Spectral	[7]	The authors used such methods to understand the underpinnings of aprosody.
		[1]	79.49
		[8]	F2 was statistically significantly correlated with the severity of negative symptoms.
		[9]	98.2
	Temporal and Statistical	[29]	85
	Prosodic and Temporal	[5]	93.8
		[6]	79.4
	Acoustic and Text Features	[18] [19]	76.2
3	Prosodic, Spectral, and Temporal	[13]	81.3
		[20]	90.5
		[21]	91.79
	Prosodic, Spectral, and Statistical	[23]	82
		[24]	The association between disorganized speech and adjunctive use of mood stabilizers could perhaps be understood in the context of a relationship with impulsiveness/aggressiveness or in terms of deconstructing the Kraepelinian dualism.
4	Prosodic, Temporal, and Statistical	[4]	87.5
	Prosodic, Spectral, Temporal, and Statistical	[14]	The authors provide an online database with their search results and synthesize how acoustic features appear in each disorder.
		[10]	90
		[12]	90

The most common approach is to use a combined set of parameters (two or more categories). With two categories, the best result obtained was using prosodic and spectral parameters, as in the work of [9] (98% accuracy). Using three categories, the best result was obtained with prosodic, spectral, and temporal features (92% accuracy in [21]). Using the four categories, the maximum accuracy of 90% was achieved in two works.

The use of temporal features alone does not present a discriminant power that can be considered for the identification of schizophrenia, and similarly to other authors, it will be an advantage to combine at least two categories of parameters. The more promising category are the prosodic and spectral features.

The prosodic features F0 and its derived ones, such as QEVA, SDVV, and the spectral SSDL (derived from the spectrogram), have the best performance in schizophrenia classification.

2. Speech Features Description

This section describes the speech features mentioned previously.

The fundamental frequency (or pitch) measures the frequency of vibration of the vocal folds; consequently, its inverse is the fundamental or glottal period. There are several methods for estimating the fundamental frequency. The most robust is estimating the first peak of the normalized autocorrelation of the signal [31].

The intensity (loudness or amplitude) is defined as the acoustic intensity in decibels relative to a reference value and is perceived as loudness [14].

Jitter measures deviations in frequency between consecutive glottal periods, and this commonly used method is based on the DSYPA algorithm (dynamic programming project phase slope algorithm). This algorithm estimates the opening and closing instants of the glottis (glottal closure instant) [31]. Jitter can be measured in four different ways, but the most used ways are relative jitter (jitter) and absolute jitter (jitta). Relative jitter is the mean absolute difference between the consecutive glottal periods divided by the mean period and is expressed as a percentage. The absolute jitter is the variation of the glottal period between cycles (the mean absolute difference between consecutive periods) [32].

The shimmer is related to the magnitude variation along the glottal periods, which can be measured in four different ways. Relative Shimmer (Shim) and Absolute Shimmer (ShdB) are the most used. Relative Shimmer is defined as the mean absolute difference between the magnitudes of consecutive periods divided by the mean magnitude and is expressed as a percentage. The Absolute Shimmer is expressed as the peak-to-peak magnitude variation in decibels [32].

The remaining determinations forms of jitter and shimmer are not used because in a statistical study carried out by [33] they did not show statistically significant differences between jitter and relative shimmer correspondingly.

The Harmonic-to-Noise Ratio (HNR) measures the ratio between harmonic and noise components, quantifying the relationship between the periodic component (harmonic part) and aperiodic components (noise). HNR can be measured by the ratio between the amplitude of the first peak of the normalized autocorrelation, considering that this is the energy of the harmonic component of the signal, and its difference to one, that is the noise energy. This feature can be obtained with Equation (1), where H is the harmonic component given by the energy of the signal's first peak of the normalized autocorrelation. The final value of HNR is the average along all segments [32].

$$\mathrm{HNR(dB)} = 10^{*}\log_{10}\frac{\mathrm{H}}{1-\mathrm{H}} \quad (1)$$

The Noise-to-Harmonic Ratio NHR can be calculated by Equation (2). To determine the autocorrelation, it is necessary to multiply the normalized autocorrelation of a segment of a speech signal by the normalized autocorrelation of a window (ex. Hanning window). Then, the first peak of the segment signal is the autocorrelation.

$$\mathrm{NHR} = 1 - \mathrm{Autocorrelation} \quad (2)$$

The Quantization Error and Vector Angle (QEVA) contain two indicators, the mean value of the cumulative error and the mean value of the vector angle. Both indicators are calculated based on the fundamental frequency curve and fit the

fundamental frequency curve. The QEVA permit evaluates the stability and similarity of the successive fundamental frequencies of the speech signals [9].

The Standard Dynamic Volume Value (SDVV) considers the monotonous speed and intensity of speech. Considering the speaking behavior of schizophrenic people, it is related to flat affect in schizophrenic patients. The calculation is divided into three steps. The first step is the intensity calculation based on voice segments (Equation (3)), where Mws represents the intensity of speech, M is the number of voice segments, ω denotes the voice segment, L is the length of one voice segment, i denotes the index of speech content from a speaker, j represents the index of voice segments in the speech content, and r is adopted to regularize the amplitudes of voice segments.

$$\mathrm{M_{ws}} = \left(\frac{1}{\mathrm{ML}} \sum_{\mathrm{i=1}}^{\mathrm{M}} \sum_{\mathrm{j=1}}^{\mathrm{L}} \omega(\mathrm{i,j}) \right)^{\mathrm{r}} \tag{3}$$

The next step consists of determining the normalized exponent variance calculation using Equation (4), where Vs represents the exponent variance in a sentence; s(n) denotes the normalized sentence; $\overline{\mathrm{s(n)}}$

is the mean value of all the data points in the sentence, including those in the word intervals; S_l is the length of the whole sentence; and t is also adopted as r.

$$\mathrm{V_s} = \left(\frac{\sum \left(\mathrm{s(n)} - \overline{\mathrm{s(n)}} \right)^2}{\mathrm{S_l}} \right)^{\mathrm{t}} \tag{4}$$

The last step consists of the standard dynamic volume value calculation using Equation (5). It aims to represent the intensity variations in speech signals more objectively.

$$\mathit{S}\mathrm{DVV} = \frac{\mathrm{S_l^t}}{(\mathrm{ML})^{\mathrm{r}}} \frac{\sum_{\mathrm{i=1}}^{\mathrm{M}} \sum_{\mathrm{j=1}}^{\mathrm{L}} \mathrm{w(i,j)^r}}{\sum \left(\mathrm{s(n)} - \overline{\mathrm{s(n)}} \right)^{2\mathrm{t}}} \tag{5}$$

The Velocity of Speech and Articulation Rate (Equations (6) and (7)) correspond to the ratio between the number of syllables and the total time recorded with and without the duration of pauses.

$$\mathrm{Velocity\ \ of\ \ Speech} = \frac{\mathrm{Number\ \ of\ \ Syllables}}{\mathrm{Total\ \ Time\ \ Recording}}, \tag{6}$$

$$\mathrm{Articulation\ \ Rate} = \frac{\mathrm{Number\ \ of\ \ Syllables}}{\mathrm{Total\ \ Time\ \ Recording\ \ (after\ \ pause\ \ remove)}}, \tag{7}$$

The peak slope corresponds to the slope of the regression line that fits lof10 of the maxima of each frame [14].

The Mel Frequency Cepstral Coefficients (MFCC) are used to obtain an approximation of the perception of the human auditory system to the frequencies of sound. They are calculated via the frequency spectrum of small windows of the speech signal, which is obtained by the Fast Fourier Transform (FFT). Subsequently, the frequency spectrum is subjected to a bank of triangular filters, equally spaced in the Mel frequency scale, via the discrete cosine transform applied to the output of the filters. Between 13 and 20 coefficients are usually determined. Finally, energy and delta (variations along the sequence of MFCCs speech segments) are calculated [32].

The frequency formants F1, F2, and F3 correspond to the first, second, and third peaks in the spectrum resulting from a human vocal tract resonance.

The linear predictive coding (LPC) coefficients are the best method to predict the values of the next time point of the audio signal using the values from the previous *n* time points, which is used to reconstruct filter properties [14].

Symmetric Spectral Difference Level (SSDL) reflects the distribution of frequency components in the speech spectrum. It is calculated using Equation (8) [9], where N is the number of words in one emotional text; n is the word index; m denotes a factor for adjusting the symmetric amplitude difference; and a is the exponential factor, which constrains the distribution range of SSDL values.

$$\mathrm{SSDL} = \frac{1}{\mathrm{N}.10^{\mathrm{a}}} \sum_{\mathrm{n}=1}^{\mathrm{N}} \frac{\sum_{\mathrm{i}=1}^{\frac{\mathrm{fs}}{\mathrm{d}}-1} \left|\mathrm{Sn}\left(\mathrm{f}\left(\frac{\mathrm{fs}}{\mathrm{d}} - \mathrm{i}\right)\right) - \mathrm{Sn}\left(\mathrm{f}\left(\frac{\mathrm{fs}}{\mathrm{d}} + \mathrm{i}\right)\right)\right|^{\mathrm{m}} . \mathrm{fn}\left(\frac{\mathrm{fs}}{\mathrm{d}} - \mathrm{i}\right)}{\mathrm{Cn}}, \tag{8}$$

Cn is the inverse of En (Equation (9)):

$$\mathrm{Cn} = \frac{1}{\mathrm{En}} \qquad \mathrm{En} = \int_0^{\mathrm{fs}/\mathrm{d}} \mathrm{Sn}(\mathrm{fn})\, \mathrm{dfn} \tag{9}$$

The Zero-Crossing rate (ZCR) is the rate at which the signal changes from positive to negative and back, which is defined in Equation (10), and sgn x(k) in Equation (11).

$$\mathrm{ZCR} = \frac{1}{2\mathbf{N}} \sum_{\mathrm{k}=1}^{\mathrm{N}} |\mathrm{sgn}\ \mathrm{x}(\mathrm{k}) - \mathrm{sgn}[\mathrm{x}(\mathrm{k}-1)]| \tag{10}$$

$$\mathrm{sgn}[\mathrm{x}(\mathrm{k})] = \begin{cases} 1, \ \mathrm{x}(\mathrm{k}) \geq 0 \\ -1, \mathrm{x}(\mathrm{k}) < 0 \end{cases} \tag{11}$$

The utterance duration corresponds to the time taken to say the utterance, and the number of pauses corresponds to the number of silences in the speech without counting the silence of occlusions in the stop consonants. The duration of pauses corresponds to the time duration of these silences. The gap duration is any segment of recording with no subjects' speech [4]. The proportion of silence (in percentage) is the relationship between the duration time of all silence segments (without the occlusion of stop consonant) and the total duration of the speech. The total recording time is the total duration of the conversation.

Voiced and unvoiced percentages correspond to the relationship between speech and silence in total time recorded in the discourse. Quasi-open Quotient (QoQ) is the ratio of the vocal folds' opening time [14]. The number of words and verbosity correspond to the number of words in the discourse. Speaking turns correspond to the number of changes between the speakers in the discourse. The interruption is when someone speaks and is interrupted. The interjection corresponds to a sound that contains no information (e.g., "hmmm").

The probability of voicing is the probability that speech is present and generally returns a row vector with the same length of speech signal. This value can be obtained with a function such as "voiceActivityDetector" in Matlab Software.

The Interquartile range (IQR) is the difference between the upper and lower quartile in an order data set. The skewness is a measure of the lack of symmetry; the data are symmetrical if it looks the same to the left and right of the center point. The kurtosis is a measure of the relative peakedness of a distribution. The slope sign changes are a statistical feature defined as the number of times the slope of the signal waveform changes sign within an analysis window. The Hjorth feature is divided in three parameters: activity, mobility, and complexity. The activity gives a measure of the squared standard deviation of the amplitude of the signal x(t) (Equation (12)), the mobility represents the mean frequency or the proportion of the standard deviation of the power spectrum (Equation (13)), and the complexity indicates how the shape of a signal is like a pure sine wave and gives an estimation of the bandwidth of the signal (Equation (14)) [34].

$$\mathrm{activity} = \mathrm{var}(\mathrm{x}(\mathrm{t})) \tag{12}$$

$$\mathrm{mobility} = \sqrt{\frac{\mathrm{activity}(\mathrm{x}'(\mathrm{t}))}{\mathrm{activity}\ (\mathrm{x}(\mathrm{t}))}} \tag{13}$$

$$\text{complexity} = \frac{\text{mobility } (x'(t))}{\text{mobility}(x(t))} \quad (14)$$

The minimum and mean semantic distance for first-order coherence are measured as an index of “disorder” in the text [16].

3. Emotion Detection in Speech

It is not easy to understand human emotions quantitatively, but understanding them is fundamental to human social interactions. The best way to analyze them is by assessing facial expressions or speech [35].

The emotional state is vital for ensuring a good lifestyle and can be influenced by social relations, physical conditions, or health status. Various sources of information such as facial expression, brain signals (EEG), and speech can be used to identify a person’s emotion [35].

There are six basic emotions, including anger, happiness/joy, disgust, surprise, fear, and sadness, and a neutral emotional state. The other emotions are derived from these [22].

Since anhedonia (the inability to feel pleasure or satisfaction), hallucinations, and delirium are symptoms of schizophrenia, the last two of which can be accompanied by strong emotions, these symptoms can lead to a decrease in motivation and a limitation of social life. Hallucinations and delusions can also lead to an increase in anxiety and stress levels.

Emotions are convoluted psychological states composed of several components, such as personal experience and physiological, behavioral, and communicative reactions [36]. Studies with schizophrenic people show that they suffer difficulties in emotional recognition [37].

An emotional state is a feature in patients with schizophrenia [14]. **Figure 1** represents the most common emotions in schizophrenia. If possible, finding an emotional state based on speech features may be a further advantage for applications in the future context of this work.

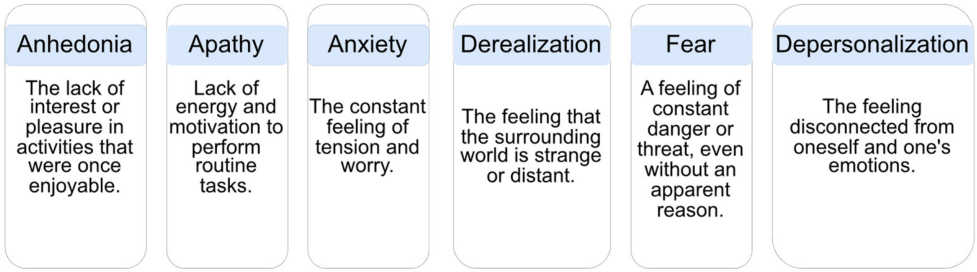

Figure 1. Emotions in schizophrenia.

Modulations in pitch [38] often control the emotional state. Most of the relevant developed work is based on using prosodic analysis to recognize emotional features.

Emotion classification is one of the most challenging tasks in speech signal processing [39]. In the work developed in [30], the authors show that acoustic and prosodic information can be combined and integrated with a speech recognition system using suprasegmental states. The same authors state that prosodic information is essential for the reliable detection of a speaker’s emotional state.

Speech emotion recognition (SER) parameters can be divided into acoustic and non-acoustic. Within acoustic, they can be grouped into different categories: prosody, spectral, wavelet, nonlinear, speech quality, and deep learning-based (encoder). The prosody features, mainly derived from F0, discriminate well between high and low arousal emotions (sad and happy). Spectral features extract the energy content of different frequency bands; the most used in emotion recognition are MFCC, Linear Predictive Cepstral Coefficients (LPCC), and Perceptual Linear Prediction (PLP) coefficients. The wavelet-based features provide better temporal resolution for the high-frequency components and better frequency resolution for the low-frequency components. Voice quality features measure the attributes related to the vocal cords (e.g., jitter, shimmer, instantaneous pitch, phase, energy, autocorrelation, harmonic-to-noise ratio (HNR), normalized noise energy (NNE), and glottal noise excitation (GNE)). Nonlinear features capture the complexity of speech signals on different emotions. The most popular are correlation dimension (CD), largest Lyapunov exponent (LLE), Hurst exponent (HE), and Lempel–Ziv complexity. The deep-learning-based features are directly given to a machine learning tool, such as

a convolutional neural network (CNN) or a long–short-term memory network (LSTM). The encoder layer of the deep-learning architecture model contains the abstract features of input speech. Non-linguistic features include non-verbal activities, such as laughter or crying, that can be detected using an automatic speech recognition system [40].

Paralinguistic features include attitudinal, intentional, and stylistic information [41]. They are essential for understanding and interpreting the pronunciation and identification of an emotional state [5]. Word choice likely indicates a speaker's emotional state [30].

For the detection of an emotional state, the MFCCs [42][43][44], zero crossing rate, energy, the entropy of energy, spectral centroid, spectral spread, spectral entropy, spectral flux, spectral roll-off, chroma vector, and chroma deviation [45] were used in previous studies.

Yadav et al. [46] presented a method to detect moments in the emotional state using Zero Time Windowing (ZTW) based on spectral energy. This method sums up the three spectral peaks at each instant of the sample Hilbert envelope of Numerator Group Delay (HNGD).

References

1. Chakraborty, D.; Yang, Z.; Tahir, Y.; Maszczyk, T.; Dauwels, J.; Thalmann, N.; Zheng, J.; Maniam, Y.; Amirah, N.; Tan, B.-L.; et al. Prediction of Negative Symptoms of Schizophrenia from Emotion Related Low-Level Speech Signals. In Proceedings of the ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing—Proceedings, Calgary, AB, Canada, 15–20 April 2018; IEEE: Piscataway Township, NJ, USA, 2018; pp. 6024–6028.
2. Tahir, Y.; Chakraborty, D.; Dauwels, J.; Thalmann, N.; Thalmann, D.; Lee, J. Non-verbal speech analysis of interviews with schizophrenic patients. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 20–25 March 2016; IEEE: Piscataway Township, NJ, USA, 2016; pp. 5810–5814.
3. Gosztolya, G.; Bagi, A.; Szalóki, S.; Szendi, I.; Hoffmann, I. Identifying schizophrenia based on temporal parameters in spontaneous speech. Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH 2018, 2018-Septe, 3408–3412.
4. Kliper, R.; Portuguese, S.; Weinshall, D. Prosodic Analysis of Speech and the Underlying Mental State; Springer: Berlin, Germany, 2016; Volume 604, ISBN 9783319322698.
5. Martínez-Sánchez, F.; Muela-Martínez, J.A.; Cortés-Soto, P.; Meilán, J.J.O.G.; Ferrándiz, J.A.N.V.; Caparrós, A.E.; Valverde, I.M.P. Can the Acoustic Analysis of Expressive Prosody Discriminate Schizophrenia? Span. J. Psychol. 2015, 18, E86.
6. Rapcan, V.; D'Arcy, S.; Yeap, S.; Afzal, N.; Thakore, J.; Reilly, R.B. Acoustic and temporal analysis of speech: A potential biomarker for schizophrenia. Med. Eng. Phys. 2010, 32, 1074–1079.
7. Compton, M.T.; Lunden, A.; Cleary, S.D.; Pauselli, L.; Alolayan, Y.; Halpern, B.; Broussard, B.; Crisafio, A.; Capulong, L.; Balducci, P.M.; et al. The aprosody of schizophrenia: Computationally derived acoustic phonetic underpinnings of monotone speech. Schizophr. Res. 2018, 197, 392–399.
8. Covington, M.A.; Lunden, S.L.A.; Cristofaro, S.L.; Wan, C.R.; Bailey, C.T.; Broussard, B.; Fogarty, R.; Johnson, S.; Zhang, S.; Compton, M.T. Phonetic measures of reduced tongue movement correlate with negative symptom severity in hospitalized patients with first-episode schizophrenia-spectrum disorders. Schizophr. Res. 2012, 142, 93–95.
9. He, F.; Fu, J.; He, L.; Li, Y.; Xiong, X. Automatic Detection of Negative Symptoms in Schizophrenia via Acoustically Measured Features Associated with Affective Flattening. IEEE Trans. Autom. Sci. Eng. 2021, 18, 586–602.
10. Voleti, R.; Liss, J.M.; Berisha, V. A Review of Automated Speech and Language Features for Assessment of Cognitive and Thought Disorders. IEEE J. Sel. Top. Signal Process. 2020, 14, 282–298.
11. Parola, A.; Simonsen, A.; Bliksted, V.; Fusaroli, R. Voice patterns in schizophrenia: A systematic review and Bayesian meta-analysis. Schizophr. Res. 2020, 216, 24–40.
12. Agurto, C.; Pietrowicz, M.; Norel, R.; Eyigoz, E.K.; Stanislawski, E.; Cecchi, G.; Corcoran, C. Analyzing acoustic and prosodic fluctuations in free speech to predict psychosis onset in high-risk youths. In Proceedings of the No. 42nd Annual International Conference of the IEEE-Engineering-in-Medicine-and-Biology-Society (EMBC), Montreal, QC, Canada, 20–24 July 2020; IEEE: New York, NY, USA, 2020; pp. 5575–5579.
13. Tahir, Y.; Yang, Z.; Chakraborty, D.; Thalmann, N.; Thalmann, D.; Maniam, Y.; Rashid, N.A.B.A.; Tan, B.-L.; Keong, J.L.C.; Dauwels, J.; et al. Non-verbal speech cues as objective measures for negative symptoms in patients with schizophrenia. PLoS ONE 2019, 14, e0214314.

14. Low, D.M.; Bentley, K.H.; Ghosh, S.S. Automated assessment of psychiatric disorders using speech: A systematic revie w. Laryngoscope Investig. Otolaryngol. 2020, 5, 96–116.

15. Mota, N.B.; Carrillo, F.; Slezak, D.F.; Copelli, M.; Ribeiro, S. Characterization of the relationship between semantic and structural language features in psychiatric diagnosis. In Conference Record—Asilomar Conference on Signals, System s and Computers, Proceedings of the No. 50th Asilomar Conference on Signals, Systems and Computers (ASILOMAR SSC), Pacific Grove, CA, USA, 29 October 2017; Matthews, M.B., Ed.; IEEE: Natal, Brazil, 2017; pp. 836–838.

16. Bedi, G.; Carrillo, F.; Cecchi, G.A.; Slezak, D.F.; Sigman, M.; Mota, N.B.; Ribeiro, S.; Javitt, D.C.; Copelli, M.; Corcoran, C.M. Automated analysis of free speech predicts psychosis onset in high-risk youths. NPJ Schizophr. 2015, 1, 15030.

17. Tovar, A.; Fuentes-Claramonte, P.; Soler-Vidal, J.; Ramiro-Sousa, N.; Rodriguez-Martinez, A.; Sarri-Closa, C.; Sarró, S.; Larrubia, J.; Andrés-Bergareche, H.; Miguel-Cesma, M.C.; et al. The linguistic signature of hallucinated voice talk in schizophrenia. Schizophr. Res. 2019, 206, 111–117.

18. Xu, S.H.; Yang, Z.X.; Chakraborty, D.; Tahir, Y.; Maszczyk, T.; Chua, V.Y.H.; Dauwels, J.; Thalmann, D.; Magnenat, N.; Tan, T.B.L.; et al. Automatic Verbal Analysis of Interviews with Schizophrenic Patients. In International Conference on D igital Signal Processing, DSP, Proceedings of the no. 23rd IEEE International Conference on Digital Signal Processing (DSP), Shanghai, China, 19–21 November 2018; Institute of Electrical and Electronics Engineers Inc.: Singapore, 201 9.

19. Argolo, F.; Magnavita, G.; Mota, N.B.B.; Ziebold, C.; Mabunda, D.; Pan, P.M.M.; Zugman, A.; Gadelha, A.; Corcoran, C.; Bressan, R.A.A. Lowering costs for large-scale screening in psychosis: A systematic review and meta-analysis of pe rformance and value of information for speech-based psychiatric evaluation. Brazilian J. Psychiatry 2020, 42, 673–686.

20. Xu, S.H.; Yang, Z.X.; Chakraborty, D.; Chua, Y.H.V.; Dauwels, J.; Thalmann, D.; Thalmann, N.M.M.; Tan, B.-L.L.; Chee Keong, J.L.; Chua, Y.H.V.; et al. Automated Verbal and Non-verbal Speech Analysis of Interviews of Individuals with Sc hizophrenia and Depression. In Proceedings of the Annual International Conference of the IEEE Engineering in Medici ne and Biology Society, EMBS, Singapore, 23–27 July 2019; pp. 225–228.

21. Espinola, C.W.; Gomes, J.C.; Pereira, J.M.S.; dos Santos, W.P. Vocal acoustic analysis and machine learning for the id entification of schizophrenia. Res. Biomed. Eng. 2021, 37, 33–46.

22. Bandela, S.R.; Kumar, T.K. Speech emotion recognition using unsupervised feature selection algorithms. Radioenginee ring 2020, 29, 353–364.

23. Chakraborty, D.; Xu, S.H.; Yang, Z.X.; Chua, Y.H.V.; Tahir, Y.; Dauwels, J.; Thalmann, N.M.; Tan, B.-L.L.; Lee, J. Predicti on of negative symptoms of schizophrenia from objective linguistic, acoustic and non-verbal conversational cues. In Pr oceedings—2018 International Conference on Cyberworlds, CW 2018, Proceedings of the No. 17th International Confe rence on Cyberworlds (CW), Kyoto, Japan, 3–5 October 2018; Sourin, A., Sourina, O., Rosenberger, C., Erdt, M., Eds.; Institute of Electrical and Electronics Engineers Inc.: Singapore, 2018; pp. 280–283.

24. Park, Y.C.; Lee, M.S.; Si, T.M.; Chiu, H.F.K.; Kanba, S.; Chong, M.Y.; Tripathi, A.; Udomratn, P.; Chee, K.Y.; Tanra, A.J.; et al. Psychotropic drug-prescribing correlates of disorganized speech in Asians with schizophrenia: The REAP-AP stu dy. Saudi Pharm. J. 2019, 27, 246–253.

25. Cannizzaro, M.S.; Cohen, H.; Rappard, F.; Snyder, P.J. Bradyphrenia and bradykinesia both contribute to altered speec h in schizophrenia: A quantitative acoustic study. Cogn. Behav. Neurol. 2005, 18, 206–210.

26. Bhatia, T.K. Language and thought disorder in multilingual schizophrenia. World Engl. 2019, 38, 18–29.

27. Espinola, C.W.; Gomes, J.C.; Mônica, J.; Pereira, S.; Pinheiro, W.; Santos, D. Detection of major depressive disorder u sing vocal acoustic analysis and machine learning-an exploratory study. Res. Biomed. Eng. 2020, 37, 53–64.

28. Cohen, A.S.; Lee Hong, S. Understanding Constricted affect in schizotypy through computerized prosodic analysis. J. P ers. Disord. 2011, 25, 478–491.

29. Mota, N.B.; Copelli, M.; Ribeiro, S. Thought disorder measured as random speech structure classifies negative sympto ms and schizophrenia diagnosis 6 months in advance. NPJ Schizophr. 2017, 3, 18.

30. Polzin, T.S.; Waibel, A.H. Detecting Emotions in Speech. Proc. Coop. Multimodal Commun. 1998. Available online: http s://www.ri.cmu.edu/pub_files/pub1/polzin_thomas_1998_1/polzin_thomas_1998_1.pdf (accessed on 18 September 20 22).

31. Cordeiro, H.T. Reconhecimento de Patologias da Voz Usando Técnicas de Processamento da Fala. 2016. Available on line: https://run.unl.pt/bitstream/10362/19915/1/Cordeiro_2016.pdf (accessed on 18 September 2022).

32. Fernandes, J.; Silva, L.; Teixeira, F.; Guedes, V.; Santos, J.; Teixeira, J.P. Parameters for Vocal Acoustic Analysis—Cur ed Database. Procedia Comput. Sci. 2019, 164, 654–661.

33. Teixeira, J.P.; Fernandes, P.O. Acoustic Analysis of Vocal Dysphonia. Procedia Comput. Sci. 2015, 64, 466–473.

34. Galvão, F.; Alarcão, S.M.; Fonseca, M.J. Predicting exact valence and arousal values from EEG. Sensors 2021, 21, 3414.
35. Teixeira, F.L.; Teixeira, J.P.; Soares, S.F.P.; Abreu, J.L.P. F0, LPC, and MFCC Analysis for Emotion Recognition Based on Speech. In Optimization, Learning Algorithms and Applications; Springer International Publishing: Cham, Switzerland, 2022; pp. 389–404.
36. Akcay, M.B.; Oguz, K.; Akçay, M.B.; Oğuz, K. Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers—ScienceDirect. Speech Commun. 2020, 116, 56–76.
37. Souto, M.T.S. Reconhecimento Emocional de Faces em Pessoas Com Esquizofrenia: Proposta de um Programa Com Recurso à Realidade Virtual; Universidade do Porto: Porto, Portugal, 2013.
38. Cohen, A.S.; Kim, Y.; Najolia, G.M. Psychiatric symptom versus neurocognitive correlates of diminished expressivity in schizophrenia and mood disorders. Schizophr. Res. 2013, 146, 249–253.
39. Davletcharova, A.; Sugathan, S.; Abraham, B.; James, A.P. Detection and Analysis of Emotion from Speech Signals. Procedia Comput. Sci. 2015, 58, 91–96.
40. Fahad, M.S.; Ranjan, A.; Yadav, J.; Deepak, A. A survey of speech emotion recognition in natural environment. Digit. Signal Process. A Rev. J. 2021, 110, 102951.
41. Teixeira, J.P.; Fernandes, J.; Teixeira, F.; Fernandes, P.O. Acoustic analysis of chronic laryngitis statistical analysis of sustained speech parameters. In BIOSIGNALS 2018—11th International Conference on Bio-Inspired Systems and Signal Processing, Proceedings; Part of 11th International Joint Conference on Biomedical Engineering Systems and Technologies, BIOSTEC, Funchal, Portugal, 19–21 January 2018; SciTePress: Setúbal, Portugal, 2018; Volume 4.
42. Ververidis, D.; Kotropoulos, C. Emotional speech recognition: Resources, features, and methods. Speech Commun. 2006, 48, 1162–1181.
43. Pribil, J.; Pribilova, A.; Matousek, J. Comparison of formant features of male and female emotional speech in czech and slovak. Elektron. Elektrotechnika 2013, 19, 83–88.
44. Nunes, A.; Coimbra, R.L.; Teixeira, A. Voice quality of European Portuguese emotional speech. Lect. Notes Comput. Sci. (Incl. Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinform.) 2010, 6001 LNAI, 142–151.
45. Papakostas, M.; Siantikos, G.; Giannakopoulos, T.; Spyrou, E.; Sgouropoulos, D. Recognizing Emotional States Using Speech Information. Adv. Exp. Med. Biol. 2017, 989, 155–164.
46. Yadav, J.; Fahad, M.S.; Rao, K.S. Epoch detection from emotional speech signal using zero time windowing. Speech Commun. 2018, 96, 142–149.