Envelope Modulation Spectrum (EMS)

The Envelope Modulation Spectrum (EMS) is a representation of the slow amplitude modulations in a signal and the distribution of energy in the amplitude fluctuations across designated frequencies, collapsed over time [1]. It has been shown to be a useful indicator of atypical rhythm patterns in pathological speech.

The original speech segment (see Figure 2), $x_0[n]$, is first filtered into 9 octave bands with center frequencies of approximately 30, 60, 120, 240, 480, 960, 1920, 3840, and 7680 Hz, using eight-order Butterworth filters (see Figure 3). Let $h_i[n]$ denote the filter associated with the ith octave. The octave band filtered signals $x_i[n]$ is then denoted by,

$$x_i[n] = h_i[n] * x[n], \quad i = 1, \ldots, 9 \tag{1}$$

The envelope for the ten signals (the original signal and 9 octave band signals- see Figure 4), denoted by $e_i[n]$, is extracted by:

$$e_i[n] = h_{LPF}[n] * |\mathcal{A}\{x_i[n]\}|, \quad i = 0, \ldots, 9 \tag{2}$$

where, $\mathcal{A}\{x(t)\} = x[n] + j\mathcal{H}\{x[n]\}$ is the analytic signal, $h_{LPF}[n]$ is the impulse response of a fourth-order, 30 Hz low-pass Butterworth filter, and $\mathcal{H}\{\cdot\}$ denotes the Hilbert transform.

Once the amplitude envelope of each signal is obtained, the mean is removed and the power spectrum for each of the bands (see Figure 5), $PowSpec_i$, is estimated by evaluating the DFT using the Goertzel algorithm [cite] at frequencies $0\ Hz < f \leq 10\ Hz$. From the power spectrum, six EMS metrics are computed for each of the 9 octave bands, and the full signal:

1. Peak frequency in the spectrum from 0-10 Hz
2. Peak amplitude / The mean amplitude in the spectrum from 0-10 Hz
3. Energy in the spectrum from 3-6 Hz / Energy in the spectrum from 0-10 Hz
4. Energy in spectrum from 0-4 Hz / Energy in the spectrum from 0-10 Hz
5. Energy in spectrum from 4-10 Hz / Energy in the spectrum from 0-10 Hz
6. Energy ratio between 0-4 Hz band and 4-10 Hz band

This results in a 60-dimensional feature vector denoted by $\mathbf{f}_{EMS}$.

Long-Term Average Spectrum (LTAS)

The Long-Term Average Spectrum (LTAS) [2] captures atypical average spectral information in the signal. Nasality, breathiness, and atypical loudness variation, common causes of intelligibility deficits in pathological speech, present themselves as atypical distributions of energy across the spectrum; LTAS attempts to measure these cues in each octave.

The original speech segment (see Figure 2), $x_0[n]$, is first filtered into 9 octave bands with center frequencies of approximately 30, 60, 120, 240, 480, 960, 1920, 3840, and 7680 Hz, using eight-order Butterworth filters (see Figure 3). Let $h_i[n]$ denote the filter associated with the ith octave. The octave band filtered signals $x_i[n]$ is then denoted by,

$$x_i[n] = h_i[n] * x[n], \quad i = 1, \ldots, 9 \tag{3}$$

Each of the ten band signals (the original full-band signal and 9 octave band signals), $x_i[n],\ i = 0, \ldots, 9$, is then framed using a 20 ms non-overlapping rectangular window the Root Mean Square (RMS) of each frame is taken and denoted $X_{RMSi}[k]$ for the kth frame of the ith band. Finally, the following features are extracted for each of the i bands:

1. The RMS value normalized by the full-band RMS value, $\frac{rms\{x_i[n]\}}{rms\{x_0[n]\}}$.*
2. The normalized mean frame RMS, $\frac{mean\{X_{RMSi}[k]\}}{rms\{x_0[n]\}}$.
3. The standard deviation of frame RMS, $std\{X_{RMSi}[k]\}$.
4. The frame standard deviation normalized by full-band RMS, $\frac{std\{X_{RMSi}[k]\}}{rms\{x_0[n]\}}$.
5. The frame standard deviation normalized by band RMS, $\frac{std\{X_{RMSi}[k]\}}{rms\{x_i[n]\}}$.
6. The skewness of frame RMS, $skew\{X_{RMSi}[k]\}$.
7. The kurtosis of frame RMS, $kurt\{X_{RMSi}[k]\}$.
8. The range of frame RMS, $range\{X_{RMSi}[k]\}$.
9. The normalized range of frame RMS, $\frac{range\{X_{RMSi}[k]\}}{rms\{x_0[n]\}}$.
10. Pairwise variability of RMS energy between ensuing frames, $\frac{mean\{abs(X_{RMSi}[k]-X_{RMSi}[k-1])\}}{rms\{x_0[n]\}}$.

This results in a 99-dimensional feature vector denoted by $\mathbf{f}_{LTAS}$.

*Note that for $i = 1$ this feature always has value equal to 1; thus this feature is removed from the set.

Rhythm Metrics

A Praat script is used to automatically extract a pitch track. The Praat script assesses voicing on a frame-by-frame basis by estimating periodicity using an autocorrelation-based method [3]. When the pitch track is undefined, we assume the speech to be consonantal. The duration of the vocalic and intervocalic segments are computed and a series of metric are then extracted:

ΔV Standard deviation of vocalic intervals.

ΔIV Standard deviation of intervocalic intervals.

ΔV-IV Standard deviation of vocalic + intervocalic intervals.

pctV Percent of utterance duration composed of vocalic intervals.

VarcoV Standard deviation of vocalic intervals divided by mean vocalic duration ($\times$100).

VarcoIV Standard deviation of intervocalic intervals divided by mean intervocalic duration ($\times$100).

VarcoV-IV Standard deviation of vocalic + intervocalic intervals divided by mean vocalic + intervocalic duration ($\times$100).

nPVI-V Normalized pairwise variability index for vocalic interval. Mean of the difference between successive vocalic intervals divided by their sum ($\times$100).

rPVI-IV Pairwise variability index for intervocalic interval. Mean of the difference between successive intervocalic intervals.

nPVI-V-IV Normalized pairwise variability index for vocalic + intervocalic interval. Mean of the difference between successive vocalic + intervocalic intervals divided by their sum ($\times$100).

rPVI-V-IV Normalized pairwise variability index for vocalic + intervocalic interval. Mean of the difference between successive vocalic + intervocalic intervals.

Articulation rate Number of vocalic + intervocalic intervals produced per second excluding pauses.

This results in a 12-dimensional feature vector denoted by $\mathbf{f}_{AutoRhythm}$.

The details of the algorithm used to partition a speech signal into voiced (i.e. vocalic, V) and unvoiced (i.e. intervocalic, IV) segments is described in detail in the pitch estimation algorithm in [4]. Here we provide an overview. The input speech signal is first split into frames. For each frame, we compute the autocorrelation of the Hanning-windowed speech signal and the autocorrelation of the Hanning window. The final estimate of the signal's short-term autocorrelation function is taken as the ratio of these two byproducts. A silence

and voicing detector based on signal energy and periodicity are used to detect voiced frames (those that remain are unvoiced). For each frame, a number of candidate pitch estimates are first computed from the final autocorrelation and a cost of selection is associated with each. A final, minimum-cost, estimate of the pitch is selected from these candidates. Here, we only make use of the voicing decision a byproduct of the algorithm. The algorithm depends on a number of different parameters. The parameter values used in our PRAAT implementation are shown in Table 1.

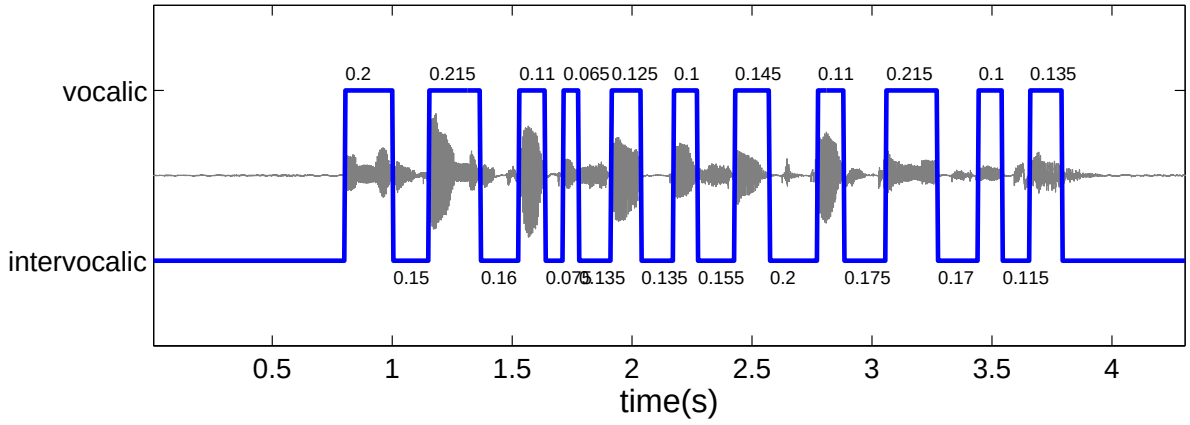

Figure 1: Example showing a speech segment(grey line) overlaid with a voicing indicator (blue line). The length of vocalic ($\sqcap$) and intervocalic ($\sqcup$) segments are labeled. The vocalic + intervocalic interval ($\sqcap$+$\sqcup$) can be obtained by adding the duration of vocalic and intervocalic pairs.

Table 1: Praat Pitch Extraction Settings

Parameter	Value
Time Step (sec)	0.005
Pitch Floor (Hz)	75*
Pitch Candidate	15*
Very Accurate	on
Silence Threshold	0.03*
Voicing Threshold	0.45*
Octave Cost	0.01*
Octave-Jump Cost	0.35*
Voiced-Unvoiced Cost	0.25
Pitch Ceiling (Hz)	600*

*Denotes default parameter value.

Phonation Features

A Praat script is used to automatically extract a voice report. The voice report consists of several metrics:

Median F0 (Hz) Median fundamental frequency.

Mean F0 (Hz) Mean fundamental frequency.

St.dev. F0 (Hz) Standard deviation of the fundamental frequency.

Minimum F0 (Hz) Minimum fundamental frequency.

Maximum F0 (Hz) Maximum fundamental frequency.

Number of pulses desc.

Number of periods desc.

Mean period (ms) desc.

St.dev. period (ms) desc.

Number of voice breaks - The number of distances between consecutive pulses that are longer than 1.25 divided by the pitch floor.

Fraction of locally unvoiced frames - This is the fraction of pitch frames that are analyzed as unvoiced

Degree of voice breaks - This is the total duration of the breaks between the voiced parts of the signal, divided by the total duration of the analyzed part of the signal (MDVP calls it DVB). Since silences at the beginning and the end of the signal are not considered breaks, you will probably not want to select these silences when measuring this parameter.

Jitter (local) The average absolute difference between consecutive periods, divided by the average period.

Jitter (local abs) (ms) The average absolute difference between consecutive periods, in seconds.

Jitter (rap) The Relative Average Perturbation, the average absolute difference between a period and the average of it and its two neighbors, divided by the average period.

Jitter (ppq5) The five-point Period Perturbation Quotient, the average absolute difference between a period and the average of it and its four closest neighbors, divided by the average period.

Shimmer (local) The average absolute difference between the amplitudes of consecutive periods, divided by the average amplitude.

Shimmer (local dB) The average absolute base-10 logarithm of the difference between the amplitudes of consecutive periods, multiplied by 20.

Shimmer (apq3) The three-point Amplitude Perturbation Quotient, the average absolute difference between the amplitude of a period and the average of the amplitudes of its neighbors, divided by the average amplitude.

Shimmer (apq5) The five-point Amplitude Perturbation Quotient, the average absolute difference between the amplitude of a period and the average of the amplitudes of it and its four closest neighbors, divided by the average amplitude.

Shimmer (apq11) The 11-point Amplitude Perturbation Quotient, the average absolute difference between the amplitude of a period and the average of the amplitudes of it and its ten closest neighbors, divided by the average amplitude. MDVP calls this parameter APQ, and gives 3.070% as a threshold for pathology.

Mean autocorrelation desc.

Mean HNR desc.

Mean HNR (dB) desc.

This results in a 24-dimensional feature vector denoted by $\mathbf{f}_{VoiceReport}$.

The results of the voice measurements depend on pitch extraction settings. Table 2 shows the pitch settings used when extracting voice report features. Note that jitter and shimmer are computed for the whole sound duration.

Table 2: Praat Pitch Extraction Settings

Parameter	Value
Time Step (sec)	0.005
Pitch Floor (Hz)	75*
Pitch Candidate	15*
Very Accurate	on
Silence Threshold	0.03*
Voicing Threshold	0.45*
Octave Cost	0.1
Octave-Jump Cost	0.6
Voiced-Unvoiced Cost	0.25
Pitch Ceiling (Hz)	600*

*Denotes default parameter value.

References

[1] J. Liss, S. LeGendre, and A. Lotto, "Discriminating dysarthria type from envelope modulation spectra." *Journal of Speech Language and Hearing Research*, vol. 53, no. 5, pp. 1246–55, 2010.

[2] E. Mendoza, N. Valencia, J. Muoz, and H. Trujillo, "Differences in voice quality between men and women: Use of the long-term average spectrum (ltas)," *Journal of Voice*, vol. 10, no. 1, pp. 59 – 66, 1996. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0892199796800191

[3] J. Liss, L. White, S. Mattys, K. Lansford, A. Lotto, S. Spitzer, and J. Caviness, "Quantifying Speech Rhythm Abnormalities in the Dysarthrias," *J Speech Lang Hear Res*, vol. 52, no. 5, pp. 1334–1352, 2009. [Online]. Available: http://jslhr.asha.org/cgi/content/abstract/52/5/1334

[4] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," in *IFA Proceedings 17*, 1993, pp. 97–110.

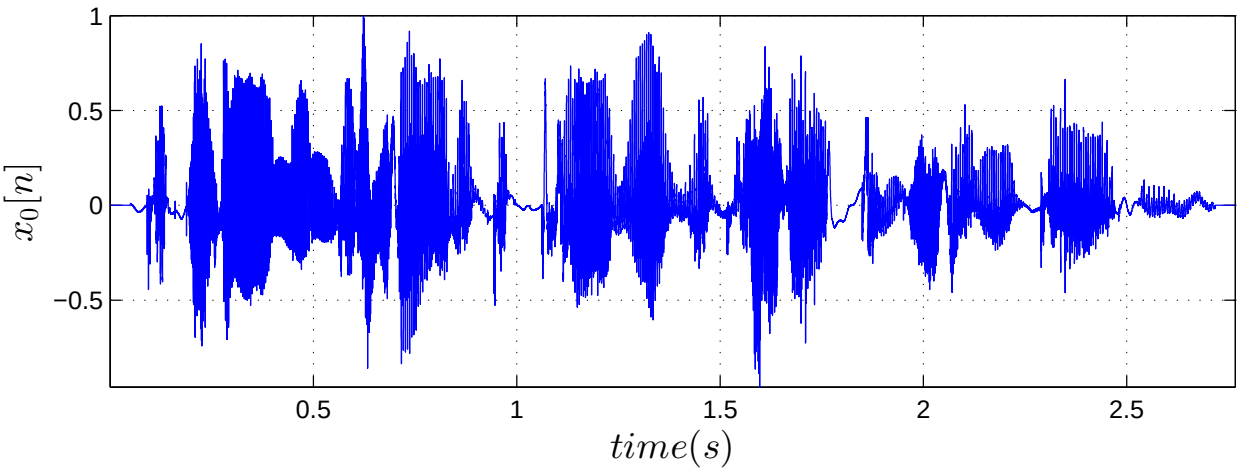

Figure 2: The original speech segment $x_0[n]$.

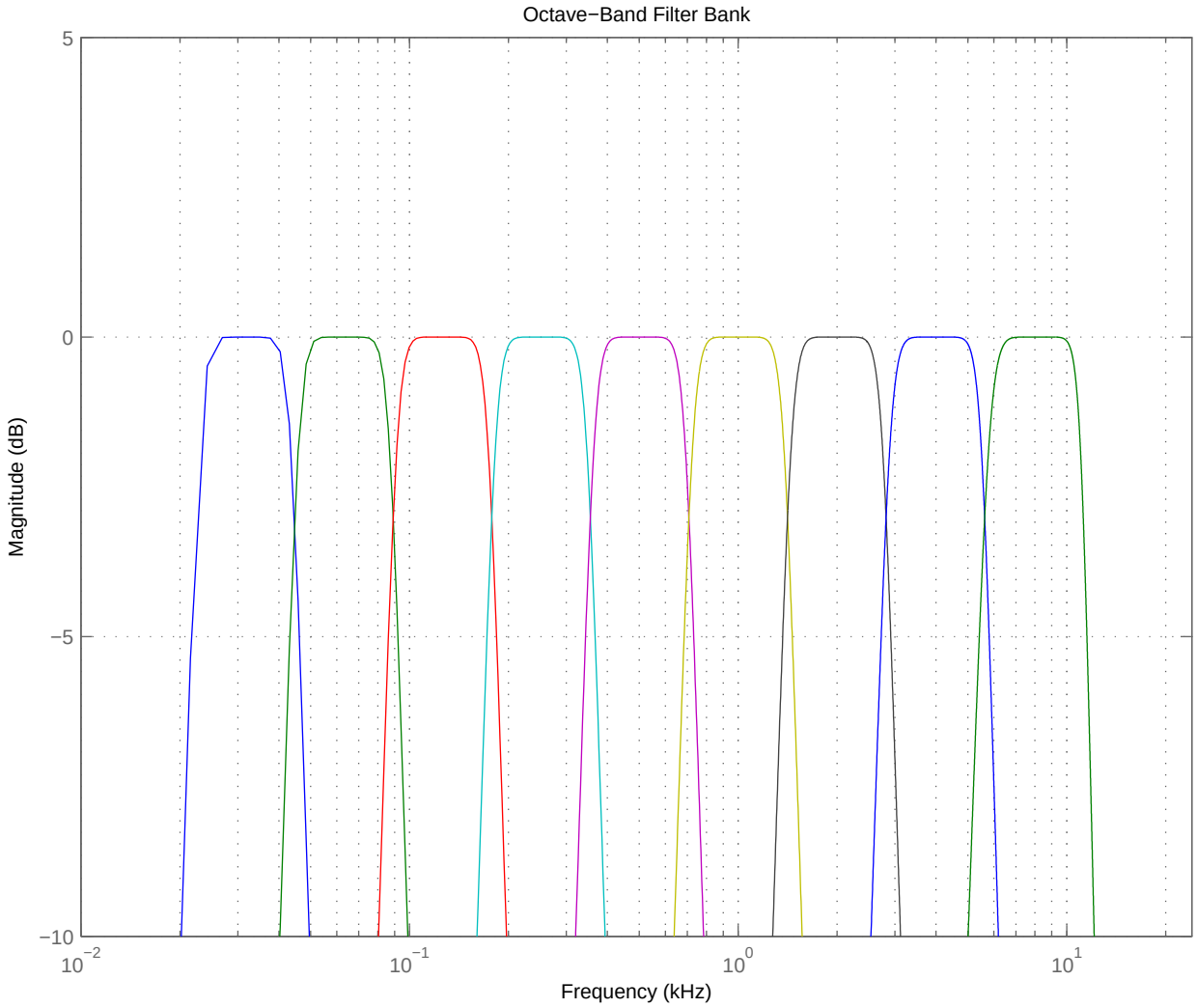

Figure 3: The octave filterbank.

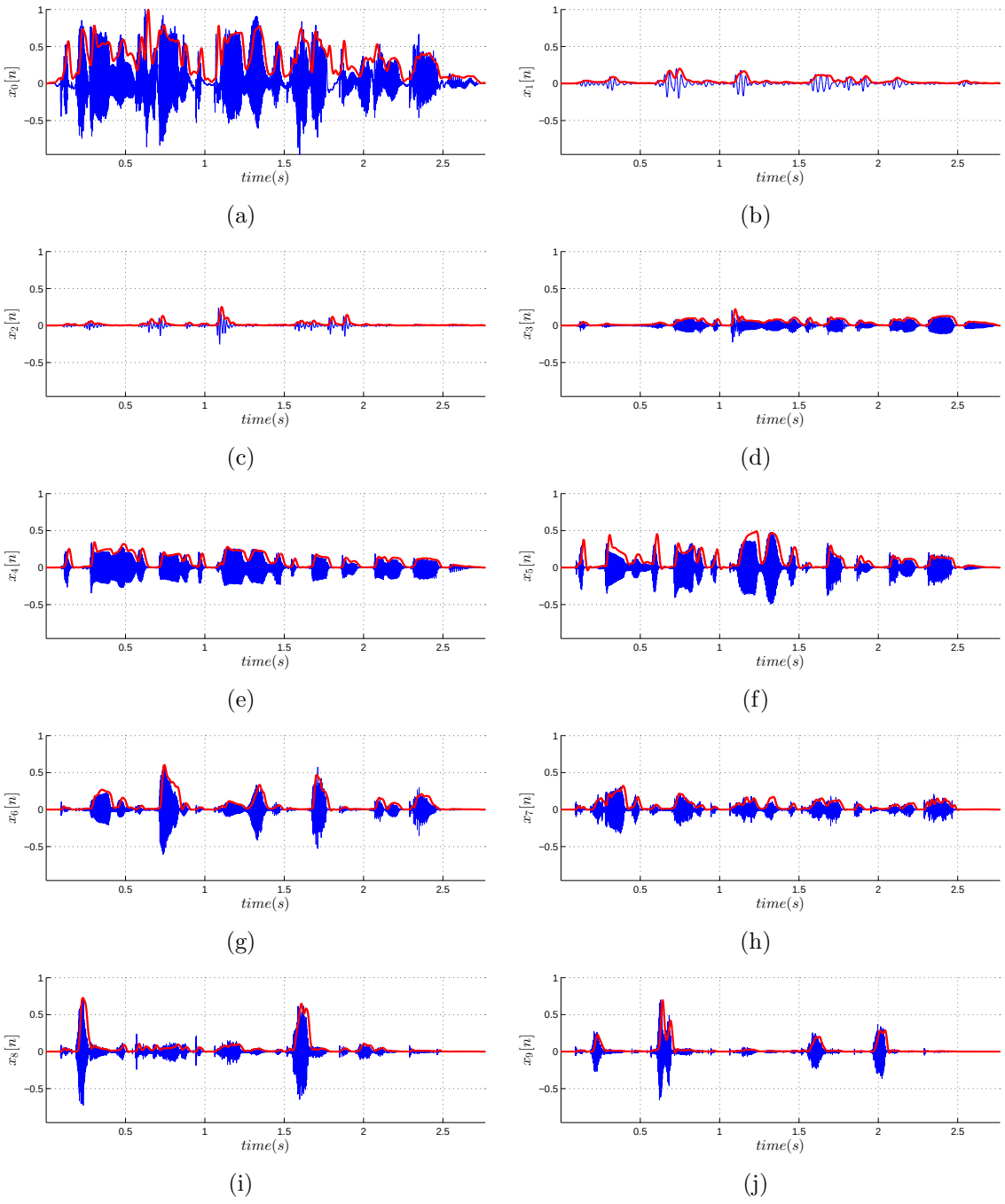

Figure 4: Bandpass filtered signals $x_i[n]$ shown in blue, and envelopes $e_i[n]$ shown in red.

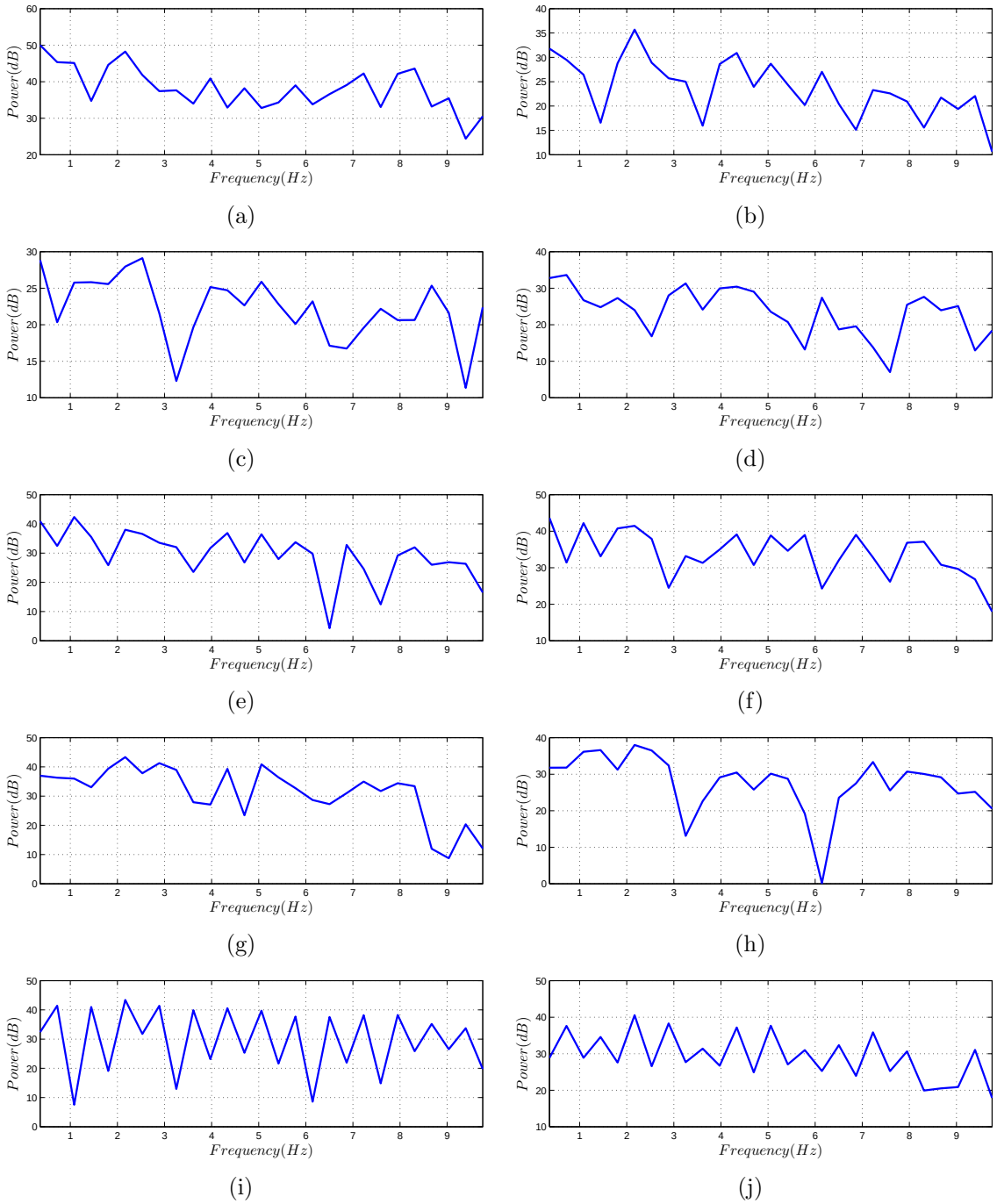

Figure 5: Estimated power spectrum $PowSpec_i$.