I. Wavelet Preliminaries

Wavelets are localized oscillatory functions with finite energy and zero mean, widely used for analyzing non-stationary signals such as speech, biomedical signals, and audio streams. Unlike Fourier-based representations, wavelets provide joint time–frequency localization, enabling the analysis of transient, short-duration, and scale-dependent signal characteristics. Owing to these properties, wavelet transforms have proven effective in capturing subtle artifacts introduced by synthetic audio generation. Given a signal $x(t)$, its Continuous Wavelet Transform (CWT) with respect to a mother wavelet $\psi(t)$ is defined as

$$W_x(a,b) = \int_{-\infty}^{\infty} x(t)\, \psi_{a,b}^{*}(t)\, dt, \tag{1}$$

where

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}}\, \psi\left(\frac{t-b}{a}\right), \tag{2}$$

$a > 0$ denotes the scale parameter controlling dilation, b denotes the translation parameter, and $(\cdot)^*$ represents complex conjugation. For real-valued wavelets, the conjugation has no effect. The different types of wavelets are as discussed below.

A. Derivative of Gaussian (DoG) Wavelet

The Derivative of Gaussian (DoG) wavelet is obtained by differentiating a Gaussian function. The first-order DoG wavelet is defined as

$$\psi_{\text{DoG}}(u) = u\, e^{-\frac{u^2}{2}}, \tag{3}$$

where

$$u = \frac{t-b}{a}. \tag{4}$$

This wavelet is real-valued, antisymmetric, and inherently zero-mean. It provides excellent temporal localization and is particularly effective in highlighting rapid transitions, discontinuities, and edge-like structures in signals.

B. Mexican Hat Wavelet

The Mexican Hat wavelet, also known as the Ricker wavelet, corresponds to the second derivative of a Gaussian function. It is defined as

$$\psi_{\text{MH}}(u) = (1-u^2)\, e^{-\frac{u^2}{2}}, \tag{5}$$

with $u = (t-b)/a$.

This wavelet is real-valued and symmetric, featuring a central positive lobe with surrounding negative lobes. It effectively suppresses slow-varying components and emphasizes peaks, blobs, and localized spectral energy variations.

C. Morlet-Inspired Wavelet

The classical Morlet wavelet is a complex-valued modulated Gaussian. In practical signal analysis and learning-based systems, a real-valued Morlet-inspired formulation is often employed, defined as

$$\psi_{\text{Morlet}}(u) = e^{-\frac{u^2}{2}} \cos(\omega u), \tag{6}$$

where ω denotes the center angular frequency.

This formulation can be interpreted as the real part of a complex Gabor atom. While it does not strictly satisfy the admissibility correction of the classical Morlet wavelet, it retains strong time–frequency localization and is effective for analyzing oscillatory components and harmonic structures in speech signals.

D. Bump Wavelet

The bump wavelet is a smooth, compactly supported function characterized by infinite differentiability. It is defined as

$$\psi_{\text{Bump}}(u) = \begin{cases} \exp\left(-\frac{1}{1-u^2}\right), & |u| < 1, \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

Due to its strict locality and absence of side lobes, the bump wavelet provides excellent frequency localization while minimizing spectral leakage. These properties make it well suited for capturing subtle, localized distortions in audio signals.

E. Morse-Inspired Wavelet

The generalized Morse wavelet is originally defined in the frequency domain and is characterized by two parameters controlling symmetry and decay. A commonly used time-domain parametric envelope inspired by the Morse family is expressed as

$$\psi_{\text{Morse}}(u) = |u|^{\beta} e^{-|u|^{\gamma}}, \tag{8}$$

where $\beta > 0$ controls the order of the wavelet and $\gamma > 0$ determines its decay behavior.

This formulation provides a flexible approximation capable of modeling a wide range of wavelet shapes. Its parameterized structure enables adaptation to varying signal characteristics, particularly for analyzing modulated and non-stationary components in speech.

F. Scale-Adaptive Feature Equalization (SAFE)

To ensure numerical stability and consistent feature magnitudes, a scale-adaptive normalization strategy is applied to all wavelet responses. Specifically, each wavelet activation is normalized by its standard deviation along the feature dimension:

$$\hat{\psi}(u) = \frac{\psi(u)}{\sqrt{\text{Var}[\psi(u)] + \epsilon}}, \tag{9}$$

where ϵ is a small constant for numerical stability.

This operation, referred to as Scale-Adaptive Feature Equalization (SAFE), enforces unit-variance wavelet responses and prevents uncontrolled growth or collapse of feature amplitudes caused by variations in the learnable scale parameter. Since wavelet functions are inherently sensitive to dilation, unnormalized activations may lead to gradient explosion, feature dominance, or training instability. The proposed SAFE mechanism ensures balanced energy across different wavelet scales, improves convergence behavior, and enhances robustness to initialization and optimization dynamics.