C'est en raison de ses bonnes propriétés de résolution temps-fréquence, de stabilité aux déformations, et de sparsité que le scalogramme $x_1(u_1, \lambda_1)$ est un outil de choix pour la classification de sons. Or, afin de pouvoir généraliser une classe de signaux à partir d'un nombre limité d'exemples, il est nécessaire de construire une représentation invariante par translation à l'échelle temporelle T des unités à classifier. Le scalogramme est une représentation *covariante* à la translation, bien qu'on puisse le rendre invariant en appliquant une moyenne glissante sur l'axe temporel. L'échelle T vaut typiquement 200 ms pour des notes de musique ou des syllabes ; pourtant, en raison de la nature transitoire des signaux de musique et de parole, les capacités de discrimination du scalogramme chutent dès que T dépasse 20 ms. Afin d'étendre cette limite, il s'agit de démoduler les oscillations du scalogramme $x_1(u_1, \lambda_1)$ dont la période est inférieure à T ; encore une fois, on requiert une bonne stabilité à la variabilité naturelle des sons à l'intérieur d'une classe.

Compte tenu de ces observations, [Andén et Mallat, 2011] ont introduit la transformée de scattering pour les sons, comme le « scalogramme du scalogramme ». Pour toute fréquence λ_1, il s'agit de transformer $x_1(u_1, \lambda_1)$ en ondelettes comme un signal unidimensionnel du temps u_1, et d'appliquer le module :

$$x_2(u_1, \lambda_1, \lambda_2) = |x_1(u_1, \lambda_1) * \psi_{\lambda_2}(u_1)|, \tag{1}$$

où $\psi_{\lambda_2}(u_1)$ est une ondelette de fréquence centrale λ_2 (en Hertz). La transformée de scattering capture explicitement les modulations d'amplitude du signal, ce qui a mené à des progrès considérables en classification de signaux. Cependant, elle n'est pas stable aux mouvement de hauteur au-delà de Q^{-1}, soit typiquement un seizième d'octave. Pour intégrer l'évolution de hauteur dans la transformée, on peut redéfinir x_2 comme

$$x_2(u_2, \lambda_1, \lambda_2) = |x_1 * \psi_{\lambda_2}|(u_2),$$

où u_2 est une variable générique construite à partir du temps u_1 et de la fréquence λ_1. Il faut remarquer que cette équation est une généralisation du modèle original du scattering 1 : en écrivant $u_2 = u_1$ et en factorisant l'action de $\psi_{\lambda_2}(u_2)$ sur toutes les fréquences λ_1, on parvient à $|x_1 * \psi_{\lambda_2}|(u_2)$.

Le modèle cortical de [Patil et al.] effectue des transformées sur les déplacements joints en temps et en log-fréquence : $u_2 = (u_1, \log_2 \lambda_1)$. La variabe de Fourier correspondante s'écrit $\lambda_2 = (a, b)$ où a est la fréquence temporelle (en Hertz) u_1 et b est la fréquence log-fréquentielle (en cycles par octave).

$$\psi_{\lambda_2}(u_1, \log_2 \lambda_1) = \psi_a(u_1) * \psi_b(\log_2 \lambda_1)$$

Nous proposons d'étendre ce modèle afin d'inclure la variabilité des déplacements sur les octaves. On pose donc $u_2 = (u_1, \log_2 \lambda_1, \lfloor \log_2 \lambda_1 \rfloor)$ la variable de déplacement sur la spirale. La fréquence correspondante s'écrit $\lambda_2 = (a, b, c)$ où a est la fréquence temporelle (en Hertz), b est la fréquence log-fréquentielle (en cycles par octave), c est la « fréquence selon les octaves » (en cycles par octave). L'expérience de Shepard nous apprend que, bien les fréquences b et

c soient exprimées dans la même unité, elles correspondent à des phénomènes différents : avec typiquement $|b^{-1}| < 1$ octave et $|c^{-1}| > 1$ octave, b quantifie un déplacement angulaire sur la spirale tandis que c quantifie un déplacement radial. On propose d'appeler « ondelettes de Shepard » les produits séparables d'ondelettes sur la spirale $u_2 = (u_1, \log_2 \lambda_1, \lfloor \log_2 \lambda_1 \rfloor)$:

$$\psi_{\lambda_2}(u_1, \log_2 \lambda_1, \lfloor \log_2 \lambda_1 \rfloor) = \psi_a(u_1) * \psi_b(\log_2 \lambda_1) * \psi_c(\lfloor \log_2 \lambda_1 \rfloor).$$

Sur la figure 2, on a représenté deux ondelettes de Shepard ψ_{λ_2} pour différentes valeurs de fréquences a, b et c.