Etude pour l'amélioration de la parole codée par transformation en paquets de framelette serrée

Souhir Bousselmi Kais Ouni
Unité de Recherche Traitement du Signal, Traitement de l'Image et Reconnaissance de Formes
Ecole Nationale d'Ingénieurs de Tunis (ENIT), BP-37, Le Belvédère, 1002 Tunis, Tunisie
souhir.bousselmi@laposte.net, kais.ouni@enit.rnu.tn

RÉSUMÉ

Dans ce papier nous proposons d'étudier les performances d'une nouvelle représentation temps-frence dite la transformation en paquets de framelette serrée dans le codage de la parole. Nous avons effectué, pour cela, une étude comparative avec la transformation en paquets d'ondelette. L'évaluation des performances a été effectué en utilisant différents critères objectifs : le gain de codage, l'erreur quadratique moyenne à racine normalisée, le rapport signal sur bruit de crête, le rapport signal sur bruit segmental, le rapport signal sur bruit segmental à fréquence pondérée et le PESQ. Les résultats obtenus montrent que le codage de la parole basé sur la transformation en paquets de framelette fournit une qualité supérieure à celui basé sur la transformation en paquets d'ondelette.

ABSTRACT

Study for improving the coded speech by tight framelet packet transform

In this paper we propose to study the performance of a new time-frequency representation called the tight framelet packets transform in speech coding. For this, we performed a comparative study with the wavelet packets transform. Performance evaluation was done using various objective criteria : the coding gain, the normalized root mean square error, the peak signal to noise ratio, the segmental signal to noise ratio, the frequency weighted segmental signal to noise ratio and PESQ. The obtained results show that the speech coding by framelet packets transform provides a higher quality than that using wavelet packet transform.

MOTS-CLÉS : Codage de la parole, frame d'ondelette, transformation en paquets de framelette, transformation en paquets d'ondelette.

KEYWORDS: Speech coding, wavelet frame, framelet packets transform , wavelet packets transform.

1 Introduction

Le codage par transformée des signaux de la parole est une application du traitement de la parole en pleine expansion. La particularité essentielle de ce type de codage est la modification de la représentation temporelle du signal d'entrée par une représentation temps-fréquence. Ce changement d'espace de représentation réduit la redondance due à la corrélation du signal ce qui rend la quantification plus efficace que la quantification directe des échantillons du signal (Dia, 1993)(Mariani, 2002). La transformation en paquets d'ondelette est une représentation

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 1: JEP, pages 473–480,
Grenoble, 4 au 8 juin 2012.

temps-fréquence qui a été utilisée dans l'élaboration des codeurs de la parole et audio (Kastantin, 1996)(Sinha et Tewfik, 1993). Toutefois, cette transformation présente des inconvénients qui limitent ses performances dans le codage de la parole. En effet, les ondelettes orthogonales à support compact ne sont pas symétriques et ne possèdent pas un déphasage linéaire, à l'exception de l'ondelette triviale de Haar. De plus, vu l'échantillonnage critique les ondelettes orthogonales ne sont pas invariantes par translation (Abdelnour et Selesnick, 2005)(Selesnick, 2001). De ce fait, il est intéressant d'intégrer les frames d'ondelettes possédant des propriétés souhaitables en codage de la parole. La symétrie des frames d'ondelettes permet d'améliorer le traitement aux bords des blocs du signal. La linéarité de phase permet d'éliminer les distorsions fréquentielles. La régularité "smoothness" conduit à une représentation plus compacte du signal. La redondance des frames d'ondelettes engendre un plan temps-fréquence dense ce qui entraîne une invariance par translation approximative. Autres ces propriétés, les frames d'ondelettes assurent une reconstruction parfaite et robuste des signaux et une forte résistance aux bruits de quantification (Goyal et Vetterli, 1998). La méthode la plus exploitée pour construire une frame d'ondelette consiste à utiliser un banc de filtres sur-échantillonné à trois bandes composé d'un filtre passe-bas et deux filtres passe-haut (Selesnick et Sendur, 2000)(Selesnick, 2004). La représentation temps-fréquence issu des frames d'ondelettes appelée la transformation en frame d'ondelette ou la transformation en framelette est obtenue par des itérations successives du banc de filtre sur-échantillonné sur les sorties du filtre passe-bas. La différence essentielle entre la transformation en ondelette et la transformation en framelette est que, dans le cas de la transformation en framelette, chaque étape de décomposition est constituée de deux filtres passe-haut. La transformation en framelette a permis d'obtenir une meilleur reconstruction des signaux de la parole comparé à la transformation en ondelette classique (Bousselmi et Ouni, 2010). Cependant, la transformation en framelette présente l'inconvénient de ne pas avoir un découpage en sous-bandes tenant compte du modèle de l'oreille humaine, ce qui limite ces performances en codage de la parole. Pour remédier à cet inconvénient, une généralisation qui consiste en plus à décomposer les bandes passe-hauts est construite. Elle est baptisée la transformation en paquets de framelette (Lu et Fan, 2011)(SUQI, 2009). L'objectif de ce papier est d'étudier les performances de cette nouvelle transformation dans le codage de la parole. Nous avons effectué, pour cela, une étude comparative avec la transformation en paquets d'ondelette. L'évaluation des performances a été faite en utilisant différents critères objectifs : le gain de codage, l'erreur quadratique moyenne à racine normalisée, le rapport signal sur bruit de crête, le rapport signal sur bruit segmental, le rapport signal sur bruit segmental à fréquence pondérée et le PESQ. Ce papier est organisé comme suit : dans la deuxième section nous introduisons les concepts de base des frames d'ondelettes. Dans la troisième section nous présentons la transformation en paquets de framelette et l'arbre de décomposition adopté dans notre étude. Dans la quatrième section nous étudions les performances de la transformation en paquets de framelette dans le codage des signaux de la parole.

2 Frame d'ondelette

La famille de fonctions ψ_{mn} (m,n ∈ Z) est une frame, s'il existe deux nombres positives A et B tel que pour tout f ∈ $L^2(R)$ on a (Daubechies, 1992) :

$$A\left\|f\right\|^2 \leq \sum_{m,n}\left|\left\langle f,\psi_{mn}\right\rangle\right|^2 \leq B\left\|f\right\|^2 \quad (1)$$

Les nombres A et B sont appelés les bornes de frame. Le plus grand nombre $A \succ 0$ et le plus petit nombre $B \succ 0$ satisfaisant l'inégalité 1 sont appelés les bornes de frame optimale. Si $A = B$ ont dit que la frame est ajustée ou serrée. La frame $\left\{\left(\psi^i_{mn}\right)_{m,n\in Z}\right\}_{i=1}^{N}$ où $\psi^i_{mn}(t) = 2^{m/2}\psi_i(2^m t - n)$ est appelée frame d'ondelette. Les fonctions ψ^i_{mn} sont appelées les " framelettes"(Petukhov, 2003). Dans le but de construire une frame d'ondelette dyadique, on se base sur l'analyse multirésolution et le principe d'extension unitaire, vu l'existence des algorithmes d'implémentation rapide (Daubechies et Shen, 2003)(Benedetto et Li, 1998). Le type de frame d'ondelette utilisé dans ce travail possède une seule fonction d'échelle $\phi(t)$ et deux framelettes $\psi_1(t)$ et $\psi_2(t)$. D'après la structure multirésolution, la fonction d'échelle et les framelettes sont définies par les relations suivantes (Selesnick, 2004) :

$$\phi(t) = \sqrt{2}\sum_n h_0(n)\phi(2^j t - n) \tag{2}$$

$$\psi_i(t) = \sqrt{2}\sum_n h_i(n)\phi(2t - n) \quad i = 1, 2 \tag{3}$$

où $h_i(n)$, $n \in Z$ sont les filtres à réponse impulsionnelle finie et à support compact, associé au banc de filtres sur-échantillonnés à trois bandes, présenté dans la figure 1. Le filtre h_0 est un filtre passe bas et les deux filtres h_1 et h_2 sont des filtres passe-haut. Chaque bande du banc de filtre est décimée par 2. Du fait que la frame est serrée, les filtres de synthèse sont donnés par l'inverse des filtres d'analyse. La conception d'une frame d'ondelette symétrique serrée à deux générateurs

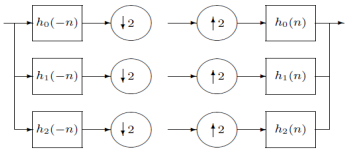

FIGURE 1 – Banc de filtres sur-échantillonné à trois sous-bandes

consiste à déterminer les filtres h_0, h_1 et h_2 vérifiant les conditions de reconstruction parfaite et de symétrie (Selesnick, 2004). Les conditions de reconstruction parfaite dans le cas du banc de filtre de la figure 1 sont données par les deux équations ci-dessous, où $H_i(z) = \sum_n h_i z^{-n}$ $i = 0, 1, 2$:

$$H_0(z)H_0(1/z) + H_1(z)H_1(1/z) + H_2(z)H_2(1/z) = 2 \tag{4}$$

$$H_0(-z)H_0(1/z) + H_1(-z)H1(1/z) + H_2(-z)H_2(1/z) = 0 \tag{5}$$

Généralement, dans le cas d'une fonction d'échelle symétrique et à support compact, il est impossible d'avoir une frame d'ondelette serrée uniquement à deux ondelettes symétriques ou antisymétriques. Cependant Petukhov fournit une condition sur $h_0(n)$ pour que ceci soit possible (Petukhov, 2003).

3 Transformation en paquets de framelette serrée

La transformation en paquets de framelette est une généralisation de la transformation en framelette. Elle est construite à partir d'un traitement répété dans la bande passe-bas ainsi que dans les deux bandes passe-haut du banc de filtre sur-échantillonné de la figure 1 (Lu et Fan, 2011). Ceci nous permet d'avoir une bonne localisation temps-fréquence et un découpage en fréquence en accord avec les bandes critiques de l'oreille humaine. Nous présentons dans la figure 2 l'arbre de décomposition correspondant à une analyse en paquets de framelette au niveau 3. Dans cette figure les indices 0, 1 et 2 associés à chaque feuille/noeud de l'arbre correspondent respectivement aux filtres d'analyses $\{h_0, h_1, h_2\}$ décimés par 2. Cet arbre est dite ternaire, vu qu'elle est basé sur deux fonctions d'ondelettes (framelettes). La représentation d'un signal de

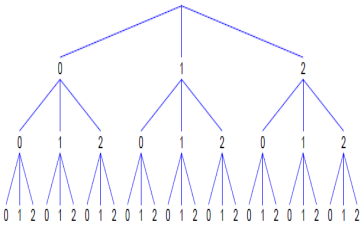

FIGURE 2 – Arbre de décomposition complète correspondant à la transformation en paquets de framelette

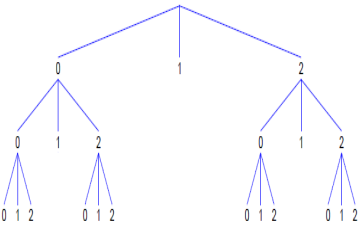

FIGURE 3 – Arbre de décomposition tronqué correspondant à la transformation en paquets de framelette

parole basé sur l'arbre complet ternaire de paquets de framelette présenté dans la figure 2 ne permet pas une amélioration de la parole codée. De ce fait, il est important de convertir cette décomposition en une décomposition binaire en réduisant le nombre de feuilles (l'arbre est donc tronqué). Ainsi les noeuds auxiliaire associés au filtre de bande passante étroite h_1 seront considérés comme des noeuds terminals qui ne seront pas considérés dans l'étape d'analyse. Les noeuds associés aux filtres h_0 et h_2 sont appelés les noeuds dyadiques. La condition principale pour qu'un arbre en paquets de framelette soit admissible est que chaque noeud dyadique possède 0 ou 3 enfants, et chaque noeud auxiliaire possède 0 enfants. Pour un niveau de décomposition donnée, chaque noeud dyadique représente une étape d'analyse qui divise chaque sous-signal en deux bandes de fréquence séparées, où la largeur de chaque bande de fréquence est égale à la moitié de la largeur de la bande du niveau précédant (Parker, 2005).

4 Evaluation et résultats

L'objectif principal de ce papier est d'étudier les performances d'une nouvelle représentation temps-fréquence, la transformation en paquets de framelette (TPF) dans le codage de la parole. Nous avons considéré pour cela le codeur dont le schéma de principe est présenté dans la figure 4. Nous sommes particuliment intéressés de l'étape de décomposition temps-fréquence où nous

comparons cette nouvelle transformation avec la transformation en paquets d'ondelette (TPO). Pour obtenir le signal codé $\hat{x}$ nous segmentons le signal original en des trames de 256 échantillons avec un chevauchement de 16 échantillons. Après allocation fixe des bits les coefficients de la transformation en paquets de framelette sont quantifiés avec un quantificateur scalaire uniforme. Au décodeur, nous effectuons la quantification et la transformation inverse des coefficients. Finalement les trames adjacentes sont ajoutées. Dans cette analyse, les signaux de parole sont

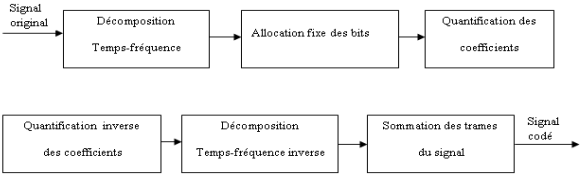

FIGURE 4 – Schéma synoptique du codeur/décodeur

synthétisés en considérant différents pourcentages de coefficients les plus énergétiques et la qualité de la parole codée est évaluée en utilisant différentes mesures de performance : le gain de codage, l'erreur quadratique moyenne à racine normalisée NRMSE, le rapport signal sur bruit de crête PSNR, le rapport signal sur bruit segmental SNRseg, le rapport signal sur bruit segmental à fréquence pondérée fwSNRseg et l'evaluation perceptive de la qualité de la parole PESQ. Le pourcentage des coefficients les plus énergétiques est un autre critère objectif qui vise à minimiser l'erreur de reconstruction. Pour valider notre approche, nous avons utilisé des signaux de parole issus de la base TIMIT et échantillonnés à 8 kHz. Le calcul des coefficients de la transformation en paquets de framelette est effectué en utilisant les filtres conçus par Selesnick (Selesnick, 2004) et en se basant sur l'arbre d'analyse à trois niveau de décomposition schématisé dans la figure 3. Les coefficients de la transformation en ondelettes sont obtenus en utilisant l'ondelette de Daubechies d'ordre 4 (4 moments nuls) et trois niveaux de décomposition. En vue d'évaluer avec précision les valeurs du SNRseg, du fwSNRseg et du PESQ nous proposons de calculer des valeurs moyennes sur 20 signaux codés issus du corpus TIMIT. Les valeurs du gain de codage, du PSNR et du NRMSE sont des valeurs moyennes sur 2182 trames obtenues par segmentation de 20 phrases issues du même corpus TIMIT. La figure 5 montre les valeurs du NRMSE dans le cas de la transformation en paquets de framelette TPF et de la transformation en paquets d'ondelette TPO en utilisant différents pourcentages de coefficients les plus énergétiques dans la synthèse. Nous remarquons que pour les différents pourcentages de coefficients une erreur minimale est obtenue en utilisant la transformation en paquets de framelette TPF. La figure 6 montre les valeurs du PSNR pour différentes pourcentages de coefficients dans le cas de la transformation en paquets de framelette TPF et de la transformation en paquets d'ondelette TPO. Il est à noter que la transformation en paquets de framelette fournit les meilleurs résultats. Nous remarquons la même chose pour le rapport signal sur bruit segmental et le rapport signal sur bruit segmental à fréquence pondérée dont les valeurs pour les différentes pourcentages sont présentées respectivement dans les figures 7 et 8. En effet pour 70% de coefficients retenus, les valeurs du PSNR, du SNRseg et du fwSNRseg dans le cas de la transformation en paquets de framelette sont respectivement 69.50 dB, 6.55 dB and 8.36 dB, tandis que dans le cas de la transformation en paquets d'ondelette, ils sont respectivement de 64.47 dB, 1.04 dB et 1 dB. Nous présentons dans la figure 9 les valeurs du

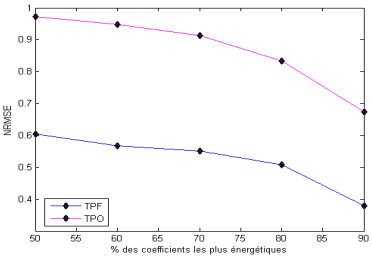

FIGURE 5 – variation du NRMSE avec le % de coeficients pour la TPF et la TPO

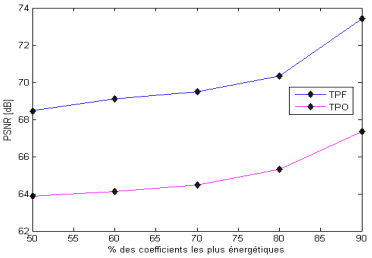

FIGURE 6 – variation du PSNR avec le % de coeficients pour la TPF et la TPO

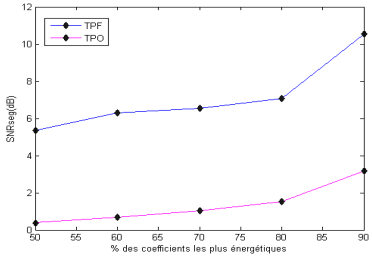

FIGURE 7 – variation du SNRseg avec le % de coefficients pour la TPF et la TPO

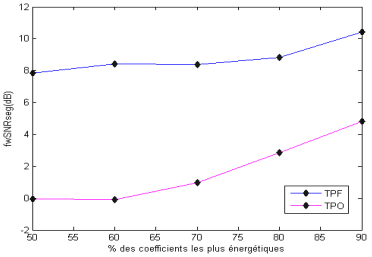

FIGURE 8 – variation du fwSNRseg avec le % de coefficients pour la TPF et la TPO

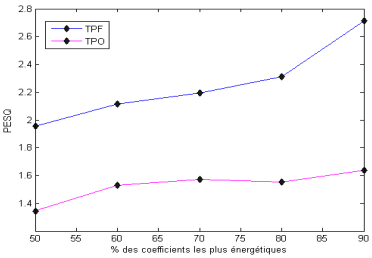

FIGURE 9 – variation du PESQ avec le % de coefficients pour la TPF et la TPO

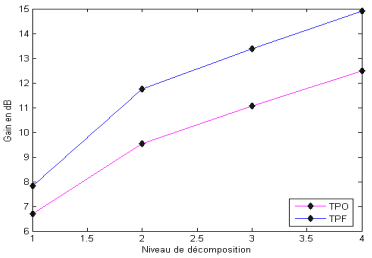

FIGURE 10 – Variation du gain de codage en fonction du niveau de décomposition pour la TPF et la TPO

PESQ pour les deux transformations. Nous remarquons que les valeurs du PESQ obtenues à partir de la transformation en paquets de framelette sont supérieures à ceux obtenues par la transformation en paquets d'ondelette. En effet, dans le cas où 90% des coefficients sont retenus dans la synthèse le PESQ est de 2.71 dans le cas de la TPF et de 1.63 pour la TPO. Les résultats objectifs montrent que les performances de la transformation en paquets de framelette sont supérieures à ceux de la transformation en paquets d'ondelette. Par ailleurs, les tests informels montrent que les signaux de parole synthétisés sont non distinguables dans le cas de la TPF avec 50% des coefficients retenus. Cependant, pour le même pourcentage les signaux de parole synthétisés en utilisant la TPO sont plus perçus. Ceci est du aux propriétés captivantes des frames d'ondelettes : la reconstruction parfaite est stable, la meilleure localisation temps-fréquence, la régularité et la symétrie. Dans une deuxiéme expérience nous avons étudié l'effet du niveau de décomposition sur la qualité de la parole codée. Pour cela nous avons calculé le gain de codage pour les niveaux de 1 à 4 dans le cas de la transformation en paquets de framelette et de la transformation en paquets d'ondelelle. Le gain de codage ou compacité d'énergie est un critère objectif utilisé pour comparer les performances entre différentes transformations. Il est donné par le rapport entre la moyenne arithmétique et la moyenne géométrique des variances des coefficients dans les sous-bandes. La courbe du gain en fonction du niveau de décomposition est schématisée dans la figure 10. Nous constatons que pour les différents niveaux de décomposition, la transformation en paquets de framelette possède le gain le plus élevé. Ce résultat est confirmé par les tests informels. Une autre étude comparative des performances de la transformation en

Débit en kbits/s	PESQ		fwSNRseg		SNRseg	
	TPF	TPO	TPF	TPO	TPF	TPO
24	2.6569	2.1738	14.1172	9.8585	12.8564	7.1485
32	2.9636	2.6061	16.4927	11.9969	19.1417	10.7230

TABLE 1 – Valeurs du PESQ, du fwSNRseg et du SNRseg dans le cas de la TPF et la TPO pour les débits de 24 kbits/s et de 32 kbits/s

paquets de framelette et la transformation en paquets d'ondelette consiste à calculer le PESQ, le fwSNRseg et le SNRseg pour différents débits. Dans le tableau 1, nous présentons les valeurs de ces mesures pour les deux transformations et ceci pour les débits de 24 kbits/s et de 32 kbits/s. Nous remarquons que pour un même débit la transformation en paquets de framelette fournit les meilleurs résultats.

5 Conclusion

Dans ce papier, nous avons étudié les performances d'une nouvelle représentation temps-fréquence basée sur des frames de paquets d'ondelettes dans le codage des signaux de la parole. Nous avons mené une étude comparative avec la transformation usuelle en paquets d'ondelette. Plusieurs critères de mesure ont été utilisés : le gain de codage, l'erreur quadratique moyenne à racine normalisée, le rapport signal sur bruit de crête, le rapport signal sur bruit segmental, le rapport signal sur bruit segmental à fréquence pondérée et le PESQ. Les résultats obtenus montrent l'importance de la transformation en paquets de framelette serrée dans la suppression des distorsions et l'amélioration des signaux codés. Comme perspective à ce travail, nous proposons de concevoir un codeur de parole de haute qualité basé sur la transformation en

paquets de framelette serrée, dans lequel nous envisageons une allocation adaptative des bits et une quantification vectorielle optimale.

Références

ABDELNOUR, A. F. et SELESNICK, I. W. (2005). Symmetric nearly shift-invariant tight frame. *In IEEE Transactions on Signal Processing*, volume 53, pages 231–239.

BENEDETTO, J. et LI, S. (1998). The theory of multiresolution analysis frames and applications to filter banks. *In Applied and Computational Harmonic Analysis*, volume 5, pages 389–427.

BOUSSELMI, S. et OUNI, K. (2010). Speech signal reconstruction based on the symmetric tight wavelet frame decomposition. *In International Congress on Image and Signal Processing (CISP)*, volume 17, pages 3453–3456.

DAUBECHIES, I. Han, B. R. A. et SHEN, Z. (2003). Framelets : Mra-based constructions of wavelet frames. *In Applied and Computational Harmonic Analysis*, volume 14, pages 1–46.

DAUBECHIES, I. (1992). Ten lectures on wavelets. *In CBMS Conference Series in Applied Mathematics*, volume 61.

DIA, H. (1993). Codage par transformée de la parole à bande élargie (0-7khz). *In Thèse de Doctorat, Institut National Polytechnique de Grenoble*.

GOYAL, V.K Thao, N. et VETTERLI, M. (1998). Quantized overcomplete expansions in r^n : analysis synthesis, and algorithms. *In IEEE Transaction on Information Theory*, volume 44, pages 16–31.

KASTANTIN, R. (1996). Codage de la parole basé sur la transformation en ondelettes. *In Thèse de Doctorat, Institut National Polytechnique de Grenoble*.

LU, D. et FAN, Q. (2011). A class of tight framelet packets. *In Czechoslovak Mathematical Journal*, volume 61, pages 623–639.

MARIANI, J. (2002). Analyse, synthèse et codage de la parole. *In Edition Hermes*.

PARKER, S. (2005). Cfs : Time-frequency representations of acoustic signals based on redundant wavelet methodologies. *In Thesis, University of Wisconsin Madison*.

PETUKHOV, A. (2003). Symmetric framelets. *In Constructive Approximation*, volume 19, pages 309–328.

SELESNICK, I. W. (2001). Smooth wavelet tight frames with zero moments. *In Applied and Computational Harmonic Analysis*, volume 10, pages 163–181.

SELESNICK, I. W. (2004). Symmetric wavelet tight frames with two generators. *In Applied and Computational Harmonic Analysis*, volume 17, pages 211–225.

SELESNICK, I. W. et SENDUR, L. (2000). Iterated oversampled filter banks and wavelet frames. *In Wavelet Applications in Signal and Image Processing*.

SINHA, D. et TEWFIK, A. H. (1993). Low bit rate transparent audio compression using adapted wavelets. *In IEEE Transactions on Signal Processing*, volume 41, pages 3464–3479.

SUQI, P. (2009). Tight wavelet frame packet. *In Thesis, Departement of Mathematics, National University of Singapore*.