HAL open science

Benchmarking flexible adaptive time-frequency transforms for underdetermined audio source separation

Andrew Nesbit, Emmanuel Vincent, Mark D. Plumbley

To cite this version:

Andrew Nesbit, Emmanuel Vincent, Mark D. Plumbley. Benchmarking flexible adaptive time-frequency transforms for underdetermined audio source separation. 2009 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), Apr 2009, Taipei, Taiwan. pp.37–40. inria-00544160

HAL Id: inria-00544160

https://inria.hal.science/inria-00544160

Submitted on 7 Dec 2010

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

BENCHMARKING FLEXIBLE ADAPTIVE TIME-FREQUENCY TRANSFORMS FOR UNDERDETERMINED AUDIO SOURCE SEPARATION

Andrew Nesbit[1,], Emmanuel Vincent[2] and Mark D. Plumbley[1]*

[1]Electronic Engineering & Computer Science, Queen Mary, University of London,
Mile End Road, London, E1 4NS, United Kingdom

[2]METISS Group, IRISA-INRIA,
Campus de Beaulieu, 35042 Rennes Cedex, France

ABSTRACT

We have implemented several fast and flexible adaptive lapped orthogonal transform (LOT) schemes for underdetermined audio source separation. This is generally addressed by time-frequency masking, requiring the sources to be disjoint in the time-frequency domain.

We have already shown that disjointness can be increased via adaptive dyadic LOTs. By taking inspiration from the windowing schemes used in many audio coding frameworks, we improve on earlier results in two ways. Firstly, we consider non-dyadic LOTs which match the time-varying signal structures better. Secondly, we allow for a greater range of overlapping window profiles to decrease window boundary artifacts. This new scheme is benchmarked through oracle evaluations, and is shown to decrease computation time by over an order of magnitude compared to using very general schemes, whilst maintaining high separation performance and flexible signal adaptivity. As the results demonstrate, this work may find practical applications in high fidelity audio source separation.

***Index Terms*—** Time-frequency analysis, Discrete cosine transforms, Source separation, Benchmark, Evaluation

1. INTRODUCTION

Our goal is to tackle the problem of *audio source separation* where the mixtures are *underdetermined* and *instantaneous*. In particular, we aim to estimate $J > 2$ simultanously active sources when the number of mixture channels is two, according to the following model:

$$x(n) = As(n) \tag{1}$$

where $x(n) = (x_1(n), x_2(n))$ and $s(n) = (s_1(n), \ldots, s_J(n))$ are the mixture and source vectors respectively, $A = (a_{i,j})$ is a $2 \times J$ matrix with real-valued entries $a_{i,j}$, and the discrete-time index ranges as $0 \leq n < N$. In most practical scenarios only $x(n)$ is observed, and little or no information is known about $s(n)$ or A (the *blind* case). In benchmarking contexts for algorithm evaluation we assume that both $s(n)$ and A are known (the *oracle* case; see Section 3).

We approach the problem of estimating $s(n)$ in the underdetermined case using the principle of *time-frequency masking* [1, 2]. This depends on the assumption that, after transforming $x(n)$ by an appropriate linear, invertible time-frequency (TF) transform (for example, a short-time Fourier transform, STFT), we have at most two sources active at each TF index m. Denote by J'_m the estimated number of active (non-zero) source coefficients at the TF index m. We assume that $J'_m = 2$ as it has been shown that this gives better performance than the simpler $J'_m = 1$ (*binary masking*) case [3, 4]. (Efficient source estimation under the general $J'_m \leq J$ assumption is currently an open research problem.) Let $X(m) = (X_1(m), X_2(m))$ be the transform of $x(n)$, and $S(m) = (S_1(m), \ldots, S_J(m))$ be the transform of $s(n)$.

Now we denote by $\mathcal{J}_m = \{j : S_j(m) \neq 0\}$ the set of all J'_m sources contributing to $X(m)$, and call it the *local activity pattern* at m. Equation (1) then reduces to a determined system at each m:

$$X(m) = A_{\mathcal{J}_m} S_{\mathcal{J}_m}(m),$$

where $A_{\mathcal{J}_m}$ is the $2 \times J'_m$ submatrix of A formed by taking columns A_j, and $S_{\mathcal{J}_m}(m)$ is the subvector of $S(m)$ formed by taking elements $S_j(m)$, whenever $j \in \mathcal{J}_m$. Once $\mathcal{J}_m$ has been estimated for each m we estimate the sources in the TF domain according to the following

$$\begin{cases} \widehat{S}_j(m) = 0 & \text{if } j \notin \mathcal{J}_m, \\ \widehat{S}_{\mathcal{J}_m}(m) = A^{-1}_{\mathcal{J}_m} X(m) & \text{otherwise,} \end{cases} \tag{2}$$

where $A^{-1}_{\mathcal{J}_m}$ is the inverse of $A_{\mathcal{J}_m}$ [2]. Finally, we invert $\widehat{S}(m)$ to obtain the estimated source vector in the time domain $\hat{s}(n)$.

It has been shown that using *lapped orthogonal transforms* (LOTs) which adapt to the time-varying signal structures in the TF domain has the potential to yield sparser representations and superior performance compared to the commonly used STFT [3–7]. Inspired by MPEG audio coding, our scheme expands on previous work [7] by allowing the LOTs to adapt to the signal flexibly and by decreasing artifacts at window boundaries in order to improve separation performance, whilst drastically decreasing computation time. This work is useful in high fidelity applications, e.g., sampling musical sources in creative or compositional contexts, where the highest possible separation quality is often more important than real-time computation.

The rest of this article is structured as follows. In Section 2 we describe the adaptive LOT framework for fast and flexible TF transforms. Section 3 outlines oracle becnhmarking techniques for algorithm evaluation. We present experimental results for mixtures of three audio sources in Section 4 and discuss these results in Section 5. Finally, in Section 6, we conclude by outlining ideas for further work.

*Supported by the EPSRC (Grant EP/E045235/1)

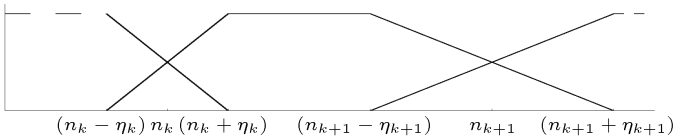

Fig. 1. Schematic representation of window β_k^λ.

2. ADAPTIVE SIGNAL EXPANSIONS

Adapting a LOT to the mixture channels $x_i(n)$ entails forming an appropriate partition of their domain $[0, N-1]$, that is, a finite set of ordered pairs

$$\lambda = \{(n_k, \eta_k)\}$$

such that

$$0 = n_0 < n_1 < \cdots < n_k < \cdots < n_{K-1} = N - 1\,,$$

where K is the number of partition points. This segments the domain of $x_i(n)$ into adjacent intervals $\mathcal{I}_k = [n_k, n_{k+1}-1]$ which should be relatively long over durations which require good frequency resolution, and relatively short over the durations requiring good time resolution. This is achieved by windowing $x_i(n)$ with windows $\beta_k^\lambda(n)$, each of which is supported in $[n_k - \eta_k, n_{k+1} + \eta_{k+1} - 1]$, thus partly overlapping with its immediately adjacent windows β_{k-1}^λ and β_{k+1}^λ by η_k and η_{k+1} points respectively (see Fig. 1). These *bell parameters* η_k are thus subject to the constraint

$$n_{k+1} - n_k \geq \eta_{k+1} + \eta_k\,. \tag{3}$$

Previous work [3–6] imposed the constraint that $\eta_k = \eta$ is constant across all $k = 1, \ldots, K-2$, whereas we now relax that constraint by allowing the η_k to vary across k. Note that $\eta_0 = \eta_{K-1} = 0$ and appropriate border modifications need to be made for this special case [8]. For every partition λ we form its associated windows according to the following function:

$$\beta_k^\lambda(n) = \begin{cases} r\left(\frac{n-(n_k-\frac{1}{2})}{\eta_k}\right) & \text{if } n_k - \eta_k \leq n < n_k + \eta_k, \\ 1 & \text{if } n_k + \eta_k \leq n < n_{k+1} - \eta_{k+1}, \\ r\left(\frac{(n_{k+1}-\frac{1}{2})-n}{\eta_{k+1}}\right) & \text{if } n_{k+1} - \eta_{k+1} \leq n < n_{k+1} + \eta_{k+1}, \\ 0 & \text{otherwise,} \end{cases}$$

where the *bell function* $r(t)$ satisfies $r^2(t) + r^2(-t) = 1$ for $-1 \leq t \leq 1$, $r(t) = 0$ for $t < -1$ and $r(t) = 1$ for $t > 1$, where t is real-valued and satisfies various regularity properties [8]. The bell parameters η_k and η_{k+1} determine how quickly the window monotonically rises on its left side and monotonically falls on its right side. Although there are many possible bell functions which satisfy these constraints, in practice we use a sine bell [8]. The local cosine basis associated with the interval $\mathcal{I}_k$ is then given by modulating $\beta_k^\lambda(n)$ by functions from a cosine-IV basis as follows:

$$\mathcal{B}_k^\lambda = \left\{ \beta_k^\lambda \sqrt{\frac{2}{n_{k+1} - n_k}} \cos\left[\pi\left(m + \frac{1}{2}\right)\frac{n - (n_k - \frac{1}{2})}{n_{k+1} - n_k}\right] \right\}$$

where $m \in [0, n_{k+1} - n_k - 1]$ is the discrete frequency index. This defines the basis B^λ for the orthogonal LOT, adapted to the partition λ, for the space of signals of length N:

$$B^\lambda = \bigcup_{k=0}^{K-1} \mathcal{B}_k^\lambda.$$

This basis is only one of many possibilities. Since our aim is to find the 'best' basis, we will consider all admissible partitions $\lambda \in \Lambda$ subject to some relatively lenient constraints, each of which determines a different orthonormal basis. Thus we obtain a *library* of possible cosine packet bases for this space of signals of length N:

$$\mathcal{L} = \bigcup_{\lambda \in \Lambda} B^\lambda.$$

2.1. Computing the Best Basis

To find that $B^\lambda \in \mathcal{L}$ which gives the 'best' representation of some signal $y(n)$, we aim to minimise its *cost* $C(y)$, which should be inversely related to our separation performance criterion. We ensure that $C(y)$ is an additive function, for example, the ℓ^1 norm [8] in a blind context (where $y(n) = x_i(n)$), or an oracle benchmarking criterion (see Section 3). Past work involves restricting the set of admissible partitions to be dyadic so that each n_k is proportional to a power of 2 for all m, and also so that all bell parameters $\eta_k = \eta$, for $k = 1, \ldots, K-1$, are the same [3,5,6]. This means that we can use computationally efficient, dynamic programming algorithms [8] to determine the best orthogonal basis with minimum cost $C(y)$. On the other hand, this has two major restrictions. Firstly, constraining $\mathcal{L}$ to admit only dyadic partitions means that the time-varying signal structures may not correspond well to any partition. Secondly, if η is small, which is often the case with previous dyadic partitioning schemes, then window artifacts will be more likely to occur in the estimated sources [5,9].

2.2. Fast and Flexible Partitioning Schemes

To overcome these problems, we have developed several alternative partitioning schemes and associated algorithms, also based on dynamic programming [10, 11]. In previous work [7] we described a *flexible segmentation* (FS) scheme which admits all possible partitions λ with some 'resolution' L, so that if the signal length N is an integral multiple of L, then each partition point can be written as $n_k = cL$ for $c \geq 0$, and where η_k is subject only to the condition (3). Provided that both L and N are powers of two, any library $\mathcal{L}$ admitted by FS is a superset of the library admitted by the less flexible dyadic partitioning scheme.

Although FS gives excellent separation results in the oracle context, its library $\mathcal{L}$ is very large due to a combinatorial explosion between the range of allowed interval lengths, interval onsets and bell parameters. Therefore, its computation time is very high. (For example, see Table 1.) As we wish to maintain flexible partitioning on the domain of the signal, yet decrease the time required for estimation of $s(n)$, we are motivated by the corresponding ideas from the MPEG-4 AAC audio coding framework [12] and introduce the following partitioning schemes which call *MPEG-like*:

Long-Short (LS) We restrict the range of allowable partitions to admit intervals $\mathcal{I}_k$ of only two lengths, that is, a *long interval* of length L_L and a *short interval* of length $L_S = L$, where L_L is an integral multiple of L_S, and we admit only bell parameters such that $2\eta_k \in \{L_L, L_S\}$. Apart from this restriction of interval lengths and bell parameters, there are no additional constraints, and LS is otherwise the same as FS.

Window Shapes (WS) This is equivalent to LS with the additional constraint that if $\mathcal{I}_k$ is long, then at most one of η_k and η_{k+1} is short. In other words, the four different window shapes admitted (compared to five in LS) correspond to a long window ($2\eta_k = 2\eta_{k+1} = L_L$), a short window ($2\eta_k = 2\eta_{k+1} = L_S$), a long-short *transition window* ($2\eta_k = L_L, 2\eta_{k+1} = L_S$), and a short-long ($2\eta_k = L_S, 2\eta_{k+1} = L_L$) transition window in the MPEG-4 framework.

Onset Times (OT) This is equivalent to LS with the additional constraint if any interval $\mathcal{I}_k$ is long, then n_k must satisfy $n_k = cL_L$ for some $c = 0, \ldots, \frac{N}{L_L} - 1$. Although this is equivalent to dyadic partitioning if L_L is equal to a power of two, it still admits variable bell parameters η_k.

WS/OT This scheme imposes both the WS and OT constraints simultaneously.

WS/OT/Successive Transitions (WS/OT/ST) This scheme imposes the WS/OT constraints in addition to disallowing adjacent transition windows, i.e., a transition window must be adjacent to a long window and a short window. This implements the windowing scheme used by MPEG-4, apart from the choice of the bell function $r(t)$.

Even though the sizes of the libraries become significantly smaller as we impose more constraints, we expect that the MPEG-like partitioning schemes are nevertheless sufficiently flexible that benefits gained in computation time will outweigh any decrease in separation performance.

3. ORACLE ESTIMATION

Oracle estimation allows us to judge the difficulty of estimating the sources $s(n)$ from a given mixture $x(n)$, and to gain insight into the upper performance bounds of our class of separation algorithms [4]. As it depends on knowing the reference source signals $s(n)$ and the mixing matrix A it is intended to be used for algorithm evaluation rather than for practical (semi-)blind separation applications. The aim is to determine those $\mathcal{J}_m$ and $B^\lambda \in \mathcal{L}$ which give the best possible separation performance for every time-frequency index m, by minimising the following oracle performance criterion (cost function):

$$C(\hat{s}) = \sum_{n=0}^{N-1} \sum_{j=1}^{J} \left(\hat{s}_j(n) - s_j(n)\right)^2 . \tag{4}$$

Minimising (4) is equivalent to maximising the *signal to distortion ratio* (SDR), given by

$$\text{SDR [dB]} = 10 \log_{10} \frac{\sum_{n=0}^{N-1} \sum_{j=1}^{J} \left(s_j(n)\right)^2}{\sum_{n=0}^{N-1} \sum_{j=1}^{J} \left(\hat{s}_j(n) - s_j(n)\right)^2},$$

as has already been shown [4]. Moreover, minimising $C(\hat{s})$ is equivalent to minimising at each m independently, by computing oracle local activity patterns:

$$\widehat{\mathcal{J}}_m^{\text{ora}} = \underset{\mathcal{J}_m \in \mathcal{P}_m}{\arg\min} \sum_{j=1}^{J} \left(\widehat{S}_j(m) - S_j(m)\right)^2 ,$$

where $\widehat{S}_j(m)$ on the right hand side is given by (2), and $\mathcal{P}_m$ is the set of all possible activity patterns subject to $J'_m = 2$. (As J'_m is small an exhaustive search over all $\mathcal{J}_m \in \mathcal{P}_m$ is computationally feasible.) The best orthogonal *oracle* basis is computed by determining $\widehat{\mathcal{J}}_m^{\text{ora}}$ over all $0 \leq m < N$, for each $B^\lambda \in \mathcal{L}$, and selecting that B^λ corresponding to the activity patterns giving the highest SDR using dynamic programming techniques [10, 11].

Scheme	L_L	L_S	Av. SDR [dB]	Av. Time [s]
FS	-	-	26.3	13183.0
LS	2^{10}	2^4	24.9	256.3
WS	2^{10}	2^4	24.8	185.6
OT	2^{10}	2^9	23.8	44.9
WS/OT	2^{10}	2^9	23.7	40.9
WS/OT/ST	2^{10}	2^9	23.7	42.1
FB	2^9	2^9	22.7	25.4

Table 1. Results of oracle estimation. Where applicable, the values of L_L and L_S which correspond to the average best SDR are given.

4. EXPERIMENTS AND RESULTS

We applied our methods to twenty mixtures in total, each of which had $J = 3$ sources. Ten of the mixtures were music, ten were speech. To simulate a *panned mono* mixing process, we used the following instantantaneous mixing matrix:

$$A = \begin{pmatrix} 0.2125 & 0.9487 & 0.6430 \\ 0.9772 & 0.3162 & 0.7658 \end{pmatrix}.$$

The pitched sources in the musical mixtures were harmonically related so that overlapping harmonics between different sources were expected. To ease computation time, we downsampled from 44.1 kHz to 22.05 kHz, keeping a resolution of 16 bits per sample. Each source and mixture channel signal was of length 2^{18} samples (approximately 11.9 s).

For each mixture, we performed oracle evaluations of $s(n)$ for each of the LS, WS, OT, WS/OT and WS/OT/ST partitioning schemes (see Section 2.2), with $L_L = 2^c$, where $c \in [8, \ldots, 11]$, and $L_S = 2^c$, where $n \in [4, \ldots, 9]$. We exclude all long-short combinations with $L_L \leq L_S$. This means that short intervals range from 0.73 ms to 23 ms in length, and the long intervals' range is between 12 ms and 93 ms.

A selection of the results is given by Table 1, where each entry is the average over twenty different mixtures corresponding to a particular transform scheme with given block lengths. We compare the MPEG-like schemes to FS and the baseline *fixed basis* (FB) transform (where $L_L = L_S$ and $2\eta_k = L_L$ for all k). Table 2 gives a more extensive presentation for each of the MPEG-like transforms.

5. DISCUSSION

From Table 1 we can see that as the partitioning schemes get more and more restrictive, performance naturally decreases due to their respective libraries becoming smaller. The question is, to what extent is this offset by improvements in computation time? Simply going from FS to LS (with no additional constraints apart from having long and short windows and their corresponding bell parameters) decreases average computation time from 3.7 h to 4.3 min, an approximately 98% improvement in computation time in exchange for a 1.4 dB decrease in average SDR. Blind and oracle performances evolve similarly across changes in interval lengths for adaptive dyadic and fixed basis partition schemes [6], so we expect that proportionally similar changes in computation time and performance

Scheme	L_L	L_S and Av. SDR [dB]					
		2^4	2^5	2^6	2^7	2^8	2^9
LS	2^8	22.8	22.6	22.4	22.0	-	-
	2^9	24.5	24.4	24.2	23.9	23.7	-
	2^{10}	**24.9**	24.7	24.5	24.4	24.4	24.2
	2^{11}	24.0	23.9	23.8	23.8	24.0	24.0
WS	2^8	22.7	22.5	22.3	22.0	-	-
	2^9	24.5	24.3	24.1	23.9	23.6	-
	2^{10}	**24.8**	24.6	24.4	24.3	24.3	24.1
	2^{11}	23.7	23.6	23.6	23.6	23.7	23.9
OT	2^8	21.4	21.5	21.5	21.6	-	-
	2^9	23.0	23.0	23.1	23.1	23.3	-
	2^{10}	23.2	23.2	23.3	23.4	23.6	**23.8**
	2^{11}	22.3	22.3	22.4	22.7	23.2	23.6
WS/OT	2^8	21.4	21.4	21.5	21.5	-	-
	2^9	23.0	23.0	23.0	23.1	23.2	-
	2^{10}	23.2	23.2	23.2	23.4	23.6	**23.7**
	2^{11}	22.2	22.2	22.3	22.6	23.0	23.5
WS/OT/ST	2^8	21.3	21.3	21.4	21.4	-	-
	2^9	22.9	22.9	22.9	23.0	23.2	-
	2^{10}	23.1	23.1	23.1	23.3	23.5	**23.7**
	2^{11}	22.0	22.1	22.2	22.4	22.9	23.4

Table 2. Full table of results for MPEG-like transforms. (Note that results for the FS transform do not appear in this table.)

would be experienced in the blind context with the considered new partitioning schemes. We emphasize that blind estimation typically yields around 12–13 dB for these mixtures [6] and that SDR improvements in the demonstrated range of 1–2 dB dB (in blind and oracle contexts) are significant in high fidelity applications.

As the WS scheme is identical to the LS scheme, apart from the relatively minor additional constraint of admitting only four different window shapes rather than five, it is not surprising that its performance is only 0.1 dB lower, yet it performs faster than the LS scheme by almost 28%. Indeed, this is also an acceptable tradeoff because the 'missing' window is a long window which tapers off with short bell parameters, and such relatively sharp tapers have to been shown to be undesirable, resulting in window border artifacts [5, 9].

The OT, WS/OT and WS/OT/ST schemes further improve computation time, however as they are dyadic partitioning schemes, the ranges of results indicate that most of the difference in performance, compared to the FS, LS and WS schemes, is due to dyadic vs non-dyadic partitioning, rather than variation in bell parameters η_k.

Table 2 displays a couple of interesting trends. For the LS and WS schemes, i.e., those which do not impose dyadic partitioning, as L_S decreases for any fixed L_L, the average SDR tends to increase by several tenths of a dB. However, for the OT, WS/OT and WS/OT/ST schemes, the reverse is the case; as L_S increases for any fixed L_L, the average SDR slightly increases. This further supports our claim that obviating the dyadic partitioning constraint is the primary factor in improving performance.

6. CONCLUSIONS AND FURTHER WORK

We have presented several different schemes for partitioning an audio signal in an adaptive LOT framework and suggest that the LS and WS schemes offer a good tradeoff between performance and computation time. The results highlight the performance advantages of non-dyadic partitioning of the signal domain.

Further work involves investigating methods for finding the oracle basis for each source individually, rather than for for the mixture as a whole. We expect that this would further improve performance. This would increase the computational complexity of estimating the sources, and one possible approach to tackle this would be to apply efficient, suboptimal, heuristic search algorithms to find orthogonal bases using in a flexible segmentation scheme. Also, we wish to perform listening tests to determine subjective, perceptual differences between the different partitioning schemes.

7. ACKNOWLEDGEMENTS

The authors would like to thank Rémi Gribonval for many helpful and insightful discussions. This research was performed at IRISA-INRIA, during which time Andrew Nesbit was graciously hosted as a visiting researcher.

8. REFERENCES

[1] Ö. Yılmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. Signal Process.*, vol. 52, no. 7, pp. 1830–1847, July 2004.

[2] R. Gribonval, "Piecewise linear source separation," in *Proceedings of the SPIE (Wavelets X)*, vol. 5207, pp. 297–310. WA, U.S.A., Nov. 2003.

[3] A. Nesbit and M. D. Plumbley, "Oracle estimation of adaptive cosine packet transforms for underdetermined audio source separation," in *Proc. ICASSP 2008*, Las Vegas, NV, U.S.A., Mar.–Apr. 2008, pp. 41–44.

[4] E. Vincent, R. Gribonval, and M. D. Plumbley, "Oracle estimators for the benchmarking of source separation algorithms," *Signal Process.*, vol. 87, no. 8, pp. 1933–1950, Aug. 2007.

[5] A. Nesbit, M. D. Plumbley, and M. E. Davies, "Audio source separation with a signal-adaptive local cosine transform," *Signal Process.*, vol. 87, no. 8, pp. 1848–1858, Aug. 2007.

[6] E. Vincent and R. Gribonval, "Blind criterion and oracle bound for instantaneous audio source separation using adaptive time-frequency representations," in *Proc. WASPAA2007*, New Paltz, NY, U.S.A., Oct. 2007, pp. 110–113.

[7] A. Nesbit, M. D. Plumbley, and E. Vincent, "Oracle evaluation of flexible adaptive transforms for underdetermined audio source separation," in *Proc. ICArn 2008*, Liverpool, U.K., Sept. 2008, pp. 17–20.

[8] S. Mallat, *A Wavelet Tour of Signal Processing*, Academic Press, second edition, 1999.

[9] X. Fang and E. Séré, "Adapted multiple folding local trigonometric transforms and wavelet packets," *Appl. Comput. Harmon. Anal.*, vol. 1, no. 2, pp. 169–179, 1994.

[10] Z. Xiong, K. Ramchandran, C. Herley, and M. T. Orchard, "Flexible tree-structured signal expansions using time-varying wavelet packets," *IEEE Trans. Signal Process.*, vol. 45, no. 2, pp. 333–345, Feb. 1997.

[11] Y. Huang, I. Pollak, C. A. Bouman, and M. N. Do, "Best basis search in lapped dictionaries," *IEEE Trans. Signal Process.*, vol. 54, no. 2, pp. 651–664, Feb. 2006.

[12] ISO, *Information technology—Coding of audio-visual objects—Part 3: Audio (ISO/IEC 14496-3:2005)*, ISO, 2005.