Speech Perceptual Hash Algorithm Based on Dual Features

Ning Zhang

Abstract— **A speech aware hash authentication algorithm with dual perceptual features is proposed. The method is composed of five sub methods. Each sub method is mapped into a hash bit. First, the line spectrum frequency coefficients (LSFs) and dual tree complex wavelet transform (DT-CWT) coefficients of each audio frame are extracted, and the norm, coefficient of variation and probability density estimates of the LSFs are calculated respectively. Then, the mean difference of the coefficients of the DT-CWT is calculated, and the DT-CWT are converted into the values of the matrix after singular value decomposition, Finally, each calculation result is mapped to a hash value. Simulation results show that the method is robust and discriminative to common voice content retention operations, and can detect whether the audio has been tampered with.**

Index Terms— **Perceptual Hash, Voice Content Authentication, DT-CWT, LSFs**

I. Introduction

With the rapid development of information technology, audio is gradually integrated into people's life. At the same time, it is also facing many serious security problems. Audio in the network may be subjected to a variety of malicious attacks or non-malicious operations. Therefore, we need to authenticate it to ensure that the received audio information is not damaged. At present, perceptual hash technology is an important technical means to ensure the security of multimedia information. Perceptual Hashing technology originates from digital watermarking technology. It draws lessons from the concepts and theories in related fields such as traditional cryptography hashing and multimedia authentication, and supports media perceptual content authentication through short summary and summary based matching of multimedia perceptual information [1]. The perceptual hash algorithm can uniquely map multimedia data with the same perceptual content into a digital summary. It is robust to content preserving operations and vulnerable to malicious attacks. As a perceptual hash function, it should meet the basic properties of discrimination, robustness, unidirectionality, randomness and so on [1]. The technical difficulty of audio aware hash is to balance the robustness of content operation and the vulnerability of malicious attacks, but this is also a problem we must overcome. The main stages of perceptual hash are feature extraction and Quantization Compression. Audio perceptual hash feature extraction methods mainly include linear prediction coefficient, mel frequency cepstrum coefficient, wavelet coefficient and so on. In [2], the author proposes an algorithm based on the line prediction coefficients (LPC). Firstly, the LPC matrix is constructed, and then the hash sequence is obtained by non-negative matrix decomposition. In [3], the author proposes the LSFs based on compressed domain, which takes the feature coefficients in the speech MELP coding process as the perceptual features to generate the perceptual hash sequence. In [4], the author proposes an algorithm based on linear prediction (LP) analysis, which uses the short-term average energy and LPC of each frame to construct a linear prediction matrix, then divides the matrix into blocks, performs singular value decomposition on each block, and finally constructs a hash sequence. The above LPC related algorithms have good robustness, but they are generally in computational efficiency or discrimination. In [5], the author proposes an algorithm based on correlation coefficient of the Mel frequency cepstrum (MFCC), which calculates the correlation between MFCC and pseudo-random, and then maps it to hash sequence. In [6], the author proposes an algorithm based on the fusion of MFCC and linear prediction cepstrum coefficients(LPCC). Firstly, the MFCC and LPCC are fused to form a matrix, the matrix is partitioned and decomposed into two-dimensional non-negative matrix, and then the hash sequence is constructed. In [7], the author proposes an algorithm based on MFCC, which uses MFCC to construct a matrix, uses non-negative matrix decomposition, and finally obtains the hash sequence. The above MFCC related algorithms have good robustness and discrimination, but the authentication efficiency is general or insensitive to malicious tampering. In [8], the author proposes an algorithm based on discrete wavelet transform (DWT), which performs DWT transform on audio signal to extract low-frequency coefficients, then reduces the dimension of measurement matrix, and finally constructs hash sequence. In [9], the author proposes a perceptual hash algorithm based on DWT. Firstly, the algorithm extracts the DWT coefficients of audio signal, constructs a matrix, decomposes it into non negative matrix, and finally obtains the hash sequence. In [10], the author proposes an algorithm based on radon and DCT in wavelet domain. The algorithm first extracts the wavelet coefficients, constructs them into a matrix, performs Radon transform on the matrix, performs DCT transform, and finally generates hash sequence. The above DWT related perceptual hash algorithm has good robustness, but its discrimination or authentication efficiency is not excellently good. From the above discussion, we can see that an excellent perceptual hash algorithm should have good robustness. In addition, in order to improve the performance of the algorithm, we should balance the discrimination, the vulnerability of malicious

Manuscript received December 07, 2022

Ning Zhang, School of Software, Tiangong University, Tianjin, China.

tampering and computational efficiency as much as possible. Therefore, we propose a perceptual hash algorithm based on LSFs and DT-CWT. The algorithm combines the advantages of the LSFs and DT-CWT, and has good robustness to content preserving operation, but it is vulnerable to malicious tampering.

II. PROPOSED METHOD

A. Construction of perceptual hash based on LSFs

Linear prediction analysis is one of the most effective methods to analyze speech signals [11], and linear prediction coefficients are also widely used in speech synthesis, audio coding and so on. As a deduction of LP analysis, LSFs has good quantization and interpolation characteristics, and LSFs can well correspond to the position and bandwidth of the formant of the frequency envelope. LSFs converted from LP coefficients satisfy the sorting attribute from 0 to π, as shown below:

$$0<\theta_1<w_1<\ldots<\theta_{p/2}<w_{p/2}<\Pi \qquad (1)$$

where p indicates the LP order, φi, 1 ≤ i ≤ p, are the LSFs. In this paper, p = 10. We perform 10 order LP analysis on each frame to obtain 10 LSFs. LSFs have good vulnerability and robustness. Whether they are subjected to ordinary attacks or malicious attacks, the amplitude transformation of LSFs is limited. Considering this characteristic, we calculate the second norm of the coefficient, the coefficient of variation and the probability density to quantify the compression. These three features can reflect the overall situation of LSFs coefficients from different angles. The second norm represents the distance from a group of data to the zero point, which can also be well reflected when there are small changes in this group of data. We know that the coefficient of variation is the ratio of standard deviation to average value, which can reflect the dispersion of data. Since the LSFs coefficient is relatively stable, its probability density is calculated. We compress LSFs coefficients from these three angles to construct hash values. Before calculating the norm, we first enlarge the LSFs coefficient, which is helpful to highlight the difference between the coefficients.

$$L_s(i)=\log(L(i)+1)*10 \qquad (2)$$

$$L_c(n)=L_s(i)L_s(j) \qquad (3)$$

Where is n=1, 2,…,100；i= 1,2,…,10；j= 1, 2,…,10

Then, each coefficient of LS is multiplied by all coefficients of LSFs to construct redundant data. A total of P 2 coefficients are obtained, which are represented by $L_c(n)$. Then the second norm NL of $L_c(n)$ is calculated.

$$NL=\sqrt{\sum_{i=1}^{n}Lc(i)^2} \qquad (4)$$

Next, we calculate the coefficient of variation. During the calculation, LSFs(i) is divided into two groups, and the coefficient of variation of each group is calculated respectively, so that we can construct the hash value according to the relationship between the two groups.

$$Cof_a=\sqrt{\frac{\sum_{i=1}^{n}(La(i)-\mu_a)^2}{n_a}}/\mu_a \qquad (5)$$

$$Cof_b=\sqrt{\frac{\sum_{i=1}^{n}(Lb(i)-\mu_b)^2}{n_b}}/\mu_b \qquad (6)$$

Then calculate the Difference of $\mathrm{Cof_a}$ and $\mathrm{Cof_b}$

$$Cof=Cof_a-Cof_b \qquad (7)$$

Finally, the probability density f(t) of LSFs(i) is calculated, and the number N greater than the set probability threshold δ is counted .

[f(t), mt] = ksdensity(LSF s(i)), 1 ≤ t ≤ 100 (8)

After obtaining the norm, coefficient of variation and probability density, we use different mapping methods to construct the hash value. We counted the number greater than NL in $L_c(n)$, represented by M, and compared M with the threshold T1.

$$h_1=\begin{cases}1,\mathrm{M}>T_1\\0,否则\end{cases} \qquad (9)$$

$$h_2=\begin{cases}1,Cof>T_2\\0,否则\end{cases} \qquad (10)$$

$$h_3=\begin{cases}1,N>T_3\\0,否则\end{cases} \qquad (11)$$

B. Construction of perceptual hash based on DT-CWT

Wavelet decomposition is a multi-scale, multi-resolution time frequency signal analysis tool, so many people like to use this method when extracting features. Due to the good robustness of low-frequency components of discrete wavelet, it is also widely popular in the field of Perceptual Hashing. In [8], [9], [10], DWT coefficients are extracted and hash sequences are constructed. The results also show good robustness. Because the robustness is too strong, the discrimination is not so ideal. Therefore, we use DT-CWT [12], which not only retains the advantages of DWT, but also has good translation invariance, direction selectivity and limited data redundancy.

Considering that the robustness and vulnerability of the second and third groups of coefficients are relatively balanced, and the average value can represent the overall data of a group of data and is relatively stable. Therefore, the method uses these two groups of coefficients and calculates their mean value. The fifth bit hash value is calculated below: The coefficients C (m, n), 1 ≤ m ≤ 5, 1 ≤ n ≤ 2 of speech signal are obtained after 4-stage DT-CWT transformation. We take the second group of coefficients C (2, n) and the third group of coefficients C (3, n), calculate the amplitudes of these two groups of coefficients, denoted by SM(k) and TM(p), k = 80, p = 40.

$$Sc(k) = C(2,1)^2 + C(2,2)^2 \tag{12}$$

$$Tc(m) = C(3,1)^2 + C(3,2)^2 \tag{13}$$

Then the average values of Sc and Tc, ASM and ATM can be calculated.

ASM = mean(SM), ATM = mean(TM) (14)

Let D represent the difference between ATM and ASM, and finally compare D with the preset threshold T4.

D = ASM – ATM (15)

Then:

$$h_4 = \begin{cases} 1, D > T_4 \\ 0, 否则 \end{cases} \tag{16}$$

The fifth bit hash value is calculated below: In order to have better robustness, the method uses five groups of lower coefficients and calculates its singular values.

Firstly, we calculate the amplitude of the fifth coefficient, denoted by Fc(n), n = 20, transform FM into matrixA5×4.

$$Fc(k) = C(5,1)^2 + C(5,2)^2 \tag{17}$$

$$A = \begin{pmatrix} Fc(1) & \dots & Fc(4) \\ \vdots & \ddots & \vdots \\ Fc(16) & \cdots & Fc(20) \end{pmatrix} \tag{18}$$

Singular value decomposition (SVD) is used to compress the matrix A, and finally the perceptual features are obtained.

S = SVD(A) (19)
MS = max(S) (20)

Then:

$$h_5 = \begin{cases} 1, V > T_5 \\ 0, 否则 \end{cases} \tag{21}$$

Finally, we concatenate all the bit values obtained by the five sub-methods to obtain the hash value of this frame.

$$H_i = h_1 \sqcup h_2 \_ h_3 \sqcup h_4 \sqcup h_5 \tag{22}$$

All other frames are operated according to the above method, and we will get H1, H2, · · ·, Hl. l is the number of frames.

C. *Perceptual Hash Authentication*

In the authentication process, we use frame by frame authentication. Because each frame is mapped to 5 bits, when the number of wrong bits in a frame is greater than the set intra threshold, it is determined that the frame signal has been tampered, otherwise it is determined that it has not been tampered. We detect whether the judgment is correct by calculating the bit error rate. The calculation formula is as follows:

$$BER(S1,\ S2) = \sum_{n=1}^{N} H_1(n) \oplus H_2(n) / L_n \tag{23}$$

L represents the length of the frame and H represents the hash sequence. In order to evaluate the performance of the algorithm, the false accept rate (FAR) and false rejection rate(FRR) are defined, which are calculated as follows:

$$FAR(\tau) = \int_{-\infty}^{\tau} f(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\Pi\sigma}} \int_{-\infty}^{\tau} e^{\frac{-(x-\mu)^2}{2\sigma^2}} dx \tag{24}$$

$$FRR = 1 - \frac{1}{\sqrt{2\Pi\sigma}} \int_{-\infty}^{\tau} e^{\frac{-(x-\mu)^2}{2\sigma^2}} dx \tag{25}$$

III. EVALUATIONS

A series of experiments have been done to detect the robustness, anti-collision and tamper detection and location ability of the algorithm. All twelve speech stimuli (male/female Japanese sentences) in the ATR speech database(B set), which were clipped into 8.1-sec durations, sampled at 16 kHz, and quantized with 16 bits were used as the database. In the experiment, the total length of the audio clip is 129600, which is divided into frames in a non-overlapping manner. The frame length is 20ms, and the number of frames is 405.The LP order for speech analysis/synthesis was adopt as 10 to balance inaudibility and robustness. Then, the performance of the algorithm will be measured from several aspects.

A. Robustness experiment and analysis

Robustness indicates whether the audio signal processed by the content retention operation is the same as the perceptual hash value extracted from the original audio signal. In order to detect the robustness of the algorithm, we did 22 kinds of ordinary attacks and malicious attacks on 12 audio as shown in Table 1, and obtained a total of 264 audio samples. We extract the hash value of each audio sample and compare it with the original audio hash value to calculate the average bit error rate.

Robustness against general processing: The proposed method were first evaluated with four processing: (a) re-sampling at 24 kHz and (b) at 12 kHz, and (c) re-quantization with 24 bits and (d) 8 bits. When the threshold is set to 3, we can see from table 1 that the overall robustness of the echo operation algorithm is excellent. We then evaluated the proposed method with common speech processing, such as (a) a single100-ms echo addition of−6 db, speech analysis/synthesis by (b) short-time Fourier transform (STFT), (c) gammatone filterbank (GTFB), (d) Gaussian noise addition with an overall average SNR of 36db, (e) Scale2, (f) White Gaussian noise, (g) flipping, (h) jitter, (i) flanger, and (j) repetition. In addition, when the threshold is set to 2, we also test the algorithm. From the results, we can see that the robustness of the algorithm decreases as a whole. Among them, the 8k Requantize causes the most increase in bit error rate, and the bit error rate of echo attack is high. The specific average bit error rate is shown in Table 1.

Table 1. BER of algorithms with different thresholds

Operation means	BER(τ = 3)	BER(τ = 2)
Resampling	0	0.12
Resampling	0	0.98
Requantize	0	0.83
Requantize	0.74	11.18
Noise	1.21	3.6
Echoe	12.22	24.86
Speech_Flanger	1.15	7.92
Speech_Flipping	0.020	3.07
Speech_GTFB	0.12	1.36
Speech_Jitter	0	0.021
Speech_Repetition	4.59	14.75
Speech_Scale2	0	0.45
Speech_SIFT	0	0

In addition, we also implemented a separate LSFs correlation algorithm and DT-CWT correlation algorithm to calculate their bit error rate. The end of the experiment is shown in table 2.

Table 2. LSF and DT-CWT of algorithms with different thresholds

Operation means	BER(LSF)	BER(DT-CWT)
Resampling	0.12	0
Resampling	0.97	0
Requantize	0.89	0
Requantize	10.26	0
Noise	1.02	0
Echoe	0.02	0
Speech_Flanger	4.67	0.45
Speech_Flipping	2.9	0
Speech_GTFB	2.74	0
Speech_Jitter	16.48	4.74
Speech_Repetition	0	0.45
Speech_Scale2	0	0
Speech_SIFT	0.12	0

We test the LSF related algorithm, where the threshold is set to 2. It can be seen from table 2 that the LSF correlation algorithm is not good for the Requantize of 8K, and echo and is robust to the rest of the operations. As can be seen from table 2, the robustness of DT-CWT related methods is stronger than that of LSF. This may be because it consists of two sub methods and the threshold is just set to 2.

Robustness against speech codecs: Speech codecs are usually applied to speech for transmission. Therefore, robustness against speech codecs is very important to guarantee the effectiveness of perceptual hash. We chose three typical speech codecs of G.711 (pulse code modulation (PCM)), G.726 (adaptive differential PCM (ADPCM)), and G.729 (Code-excited linear prediction (CELP)), and G.7231 to evaluate the robustness.

Table 3. Coding error rate with different thresholds

Operation means	BER(τ = 3)	BER(τ = 2)
G.711	0	1.17
G.726	0.04	2.53
G.729	1	9.18
G.7231	1	9.15

As can be seen from table 3, when the threshold is set to 3, the algorithm is very robust to four coding attacks. However, when the and threshold is set to 2, the robustness of the algorithm also decreases as a whole, and the robustness to G.729 and G.7231 decreases the most.

Table 4 below are the experimental results of LSF part and DT-CWT part respectively, we can see that the overall robustness of DT-CWT part is better, while LSF part is more sensitive to G.729 and G.7231 coding.

Table 4. Coding error rate of LSF and DT-CWT

Operation means	BER(τ = 3)	BER(τ = 2)
G.711	1.15	0
G.726	2.51	0
G.729	6.63	0.45
G.7231	7.05	0.37

B. Vulnerability experiment and analysis

In the vulnerability test part, we conducted the following five attacks with high intensity:(a) highpass, (b) lowpass, (c) white noise, (d) Reverb 1s, and (e) concat. All four attacks attack frames between 135 and 270 frames of audio. When the threshold is set to 3, it can be seen that the error rate of the algorithm for splicing attack is the lowest, 48.7%, while the error rate for white noise is the highest, 89.9%. When the threshold is set to 2, although the robustness decreases, the vulnerability increases.

Table 5. Vulnerability of algorithms with different thresholds

Operation means	BER(τ = 3)	BER(τ = 2)
Highpass	85.54	47.80
Lowpass	66.36	38.97
Reverb 1s	71.63	48.35
WhiteNoise	89.89	67.10
Concat	48.71	24.75

Operation means	BER(LSFs)	BER(DT-CWT)
Highpass	86.52	66.36
Lowpass	46.39	99.02
Reverb 1s	65.38	88.54
WhiteNoise	72.79	99.94
Concat	49.69	67.22

From the test results of LSFs and DT-CWT respectively, the vulnerability of DT-CWT algorithm is relatively poor, especially for highpass and white noise attacks, while LSFs algorithm is also relatively poor for these two attacks, but the overall effect is better than DT-CWT.

C. Distinguishing experiment and analysis

Discrimination can measure the excellence of the algorithm, so we do a discrimination test in this part. Previously, we have done 12 content retention operations for 12 audio. The audio operated by maintaining the content and its corresponding original audio are divided into a group, so there are 12 groups with 13 audio in each group. So a total of 11154 hash distances can be obtained. We calculate FAR and FRR respectively according to formulas (23) and (24) and draw FAR-FRR diagram.

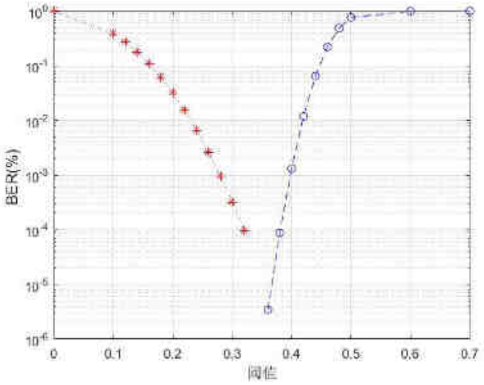

Fig. 1. FAR-FRR

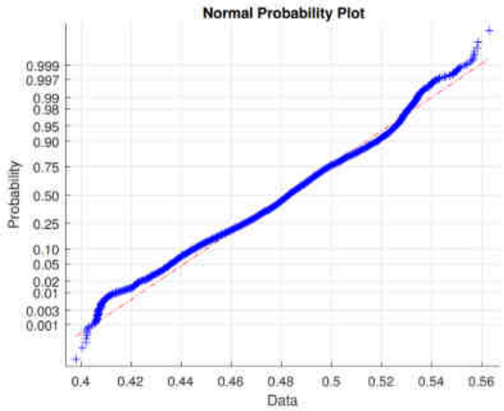

Fig. 2. BER

As shown in Figure 1, the BER of the FRR curve is below the threshold value of 0.1, indicating that the algorithm has good robustness, and the FAR-FRR curve has no intersection in the figure, indicating that the discrimination of the algorithm is also very excellent. In addition, we can also see that content authentication can be performed accurately when the threshold is less than 0.5 The BER of audio perceived hash values of different contents basically follows the normal distribution. In order to further measure the discrimination of the algorithm, we draw the distribution diagram of BER value according to the previous 11154 BER data, and Figure 2 shows the normal probability of BER. According to the central limit theorem of Oliver Laplace, theoretically, BER approximately obeys the normal distribution with parameter $(\mu0 = p, p\,\sigma = p(1-p)/N)$,p is the probability of occurrence of 1, p = 0.5, and N is the length of perceptual hash sequence. Therefore, the theoretical value of bit error rate $\mu0 = 0.5$ and the theoretical standard deviation $\sigma = 0.0111$. The mean and standard deviation obtained in the experiment are 0.4818 and 0.0267 respectively. It can be seen that the standard deviation obtained in the experiment is close to the theoretical value.

IV. CONCLUSION

A. This paper proposes an audio content algorithm based on perceptual hash which can realize tamper location. The algorithm uses DT-CWT and LSFs coefficients as perceptual features to construct hash values, maps different features into different eigenvalues through different sub algorithms, and uses Hamming distance to detect the location of audio content. The algorithm relies on the balance between the robustness and vulnerability of the two coefficients to achieve good results. It can also be seen from the experiment that the algorithm can effectively resist conventional operations, and can also detect and locate some malicious attacks. However, because the algorithm consists of five sub methods, the efficiency of the algorithm is poor, which is what we need to improve in the next step.

REFERENCES

[1] CHEN Zhen DU Ling, “Survey on image tamper detection withperceptual hashing,” 2019.

[2] Ning Chen and Wanggen Wan, “Robust speech hash function,” ETRI Journal, vol. 32, no. 2, pp. 345–347, 2010.

[3] Yuhua Jiao, Qiong Li, and Xiamu Niu, “Compressed domain perceptual hashing for MELP coded speech,” in 4th International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP 2008), Harbin, China, 15- 17 August 2008, Proceedings, Jeng-Shyang Pan, Xiamu Niu, Hsiang-Cheh Huang, and Lakhmi C. Jain, Eds. 2008, pp. 410– 413, IEEE Computer Society.

[4] Y. B. Huang, Q. Y. Zhang, and Z. T. Yuan, “Perceptual speech hashing authentication algorithm based on linear prediction analysis,” Telkomnika Indonesian Journal of Electrical Engineering, vol. 12, no. 4, 2014.E. H. Miller, “A note on reflector arrays (Periodical style—Accepted for publication),” IEEE Trans. Antennas Propagate., to be published.

[5] Jinfeng Li, Tao Wu, and Hongxia Wang, “Perceptual hashing based on correlation coefficient of mfcc for speech authentication,” Journal of Beijing University of Posts and Telecommunications, vol. 000, no. 002, pp. 89–93, 2015.

[6] Y. Huang, Q. Zhang, Z. Yuan, and Z. Yang, “The hash algorithm of speech perception based on the integration of adaptive mfcc and lpcc,” Journal of Huazhong University of Science and Technology, 2015.Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interfaces(Translation Journals style),” *IEEE Transl. J. Magn.Jpn.*, vol. 2, Aug. 1987, pp. 740–741 [*Dig. 9th Annu. Conf. Magnetics* Japan, 1982, p. 301].

[7] Chen, N., Xiao, H.-D., Wan, and W., “Audio hash function based on non-negative matrix factorisation of mel-frequency cepstral coefficients,” Information Security, IET, vol. 5, no. 1, pp. 19–25, 2011.

[8] Qiu-yu Zhang, Si-Bin Qiao, Yi-Bo Huang, and Tao Zhang, “A high-performance speech perceptual hashing authentication algorithm based on discrete wavelet transform and measurement matrix,” Multim. Tools Appl., vol. 77, no. 16, pp. 21653–21669, 2018.

[9] Ning Chen, Wanggen Wan, and He-D. Xiao, “Robust audio hashing based on discrete-wavelet-transform and non-negative matrix factorisation,” IET Commun., vol. 4, no. 14, pp. 1722– 1731, 2010.

[10] Jinfeng Li and Tao Wu, “Perceptual audio hashing using RT and DCT in wavelet domain,” in 11th International Conference on Computational Intelligence and Security, CIS 2015, Shenzhen, China, December 19-20, 2015. 2015, pp. 363–366, IEEE Computer Society.

[11] Peter Vary and Rainer Martin, Digital Speech Transmission: Enhancement, Coding and Error Concealment, Digital Speech Transmission: Enhancement, Coding and Error Concealment, 2006.

[12] Ivan W. Selesnick, Richard G. Baraniuk, and Nick G. Kingsbury, “The dual-tree complex wavelet transform,” IEEE Signal Process. Mag., vol. 22, no. 6, pp. 123–151, 2005.