ISSN:0975-9646

Multi Scale Wavelet Fusion Based Vocal Singer Matching

Meenu, Monika Aggarwal
[1,2]*ECE Department BGIET, Punjab.*

***Abstract*— The proposed work deals with the practise of multi scale analysis of a vocal music signal. The wavelet domain can be introduced to classify an exclusive singer from a database containing different singers. As it is well known that the wavelet transform of a signal at low scale value corresponds to the approximate of the signal in the original form but more details of the signal can be achieved with the analysis obtained at higher scale values of the wavelet transform. Keeping in this property of the wavelet transforms, a novel method is proposed that fuses the information obtained from the different scale wavelet transforms of an audio signal, and then use this method to match different singers in a database with a particular singer. Results show that the proposed method achieves more accuracy as compared to direct analysis or even a single scale analysis of a signal.**

Keywords— **Wavelet Transform, Information Fusion**

I. INTRODUCTION

Singer matching is an interesting problem especially from a database which has none of the professional attributes of the track. Speaker recognition has already been done [1-3] but very less work is done in case of singing track recognition of a particular singer. Singer matching from a raw database of different singers is a tedious task and one may need a lot of information for it. Wavelet Transform is a handy transform for carrying out various audio and Image applications [4-6], [7-8]. The discrete wavelet transforms (DWT) and further Haar transform are quite efficient in performing audio processing tasks. The answer to the question that how wavelet transforms can be handy for audio analysis lies in its multi resolution scale analysis. Taking an example of image analysis by wavelets, one can think that Wavelets matches the way how our eyes recognise the world when they are faced to different distances. In the actual world, a forest can be seen from many different perspectives; they are, in fact, different scales of resolution. From the window of an airplane, for instance, the forest cover seems as a solid green roof. From the window of a car, the green roof gets renovated into individual trees, and if we leave the car and approach to the forest, we can gradually see details such as the trees branches and leaves. If we had a magnifying glass, we could see a dew drop on the tip of a leaf. As we get closer to even higher scales, we can discover details that we had not seen before. Similar is the case with human ears. When the observer is at a very far distance, then audio information collected by him is different as compared to when he takes note of the sound closely. The closeness option is also vulnerable sometimes (either in image or in audio), since, if one sees the picture from too near it may not be properly visible and also a complete view is not achieved and similarly a strong audio signal heard from very near position may not be clearly understood. So, the near and far approximations both need to be taken into account. Multi-resolution scale analysis using wavelet transforms is an appropriate solution to above type of problem. Information fusion always helps to bind the different opinions together and results in forming a single decision when different information available is collectively used as input to an algorithm. Various fusion methods for detection have earlier been proposed in literature for various applications [9-11], but among them the weighted fusion method seems to be simple and efficient measure to take in to account the information obtained from different wavelet scales. Section 2 enlightens the techniques used viz. Wavelet transform, its discrete version DWT and Haar transform along with the weighted mean fusion technique. Section 3 discusses the proposed methodology and Section 4 highlights the results obtained. Section 5 concludes the paper.

II. WAVELET TRANSFORM

Wavelet transform of a function is the improved version of Fourier transform. Fourier transform is a powerful tool for analysing the components of a stationary signal. But it has failed to analyse the non-stationary signal whereas wavelet transform allows the components of a non-stationary signal to be analysed. Wavelets are finite duration oscillatory functions with zero average value. The irregularity and good localization properties make them better basis for analysis of signals with discontinuities. Wavelets can be described by using two functions viz. the scaling function f (t), also known as 'father wavelet' and the wavelet function or 'mother wavelet'. 'Mother' wavelet ψ (t) undergoes translation (b) and scaling operations (a) to give self-similar wavelet families as follows:

$$\widetilde{W}(a,b) = \frac{1}{\sqrt{a}} \int f(t)\psi\left(\frac{t-b}{a}\right)dt \quad \text{--------- (1)}$$

2.1 Discrete Wavelet Transform

The Discrete Wavelet Transform (DWT) is a transformation that can be used to analyse the temporal and spectral properties of non-stationary signals like audio. Digital audio is becoming a major part of the average computer user experience. The increasing amounts of available audio data require the development of new techniques and algorithms for structuring this information. Although, there has been a lot of research on the problem of information extraction from speech signals, work on non-speech audio like music has only appeared recently. Practical implementation of wavelet transforms requires discretisation of its translation and scale parameters. This can be achieved by taking the logarithmic scale mapping as: a = a_0^l and b = kb_0a with l, k ∈ Z (Z is the set of integers), the discrete family wavelets can now be stated as:

$$\psi_{l,k}(t) = \frac{1}{\sqrt{a_0^{\,l}}}\psi\left(\frac{t-kb_0a_0^{\,l}}{a_0^{\,l}}\right) = \frac{1}{\sqrt{a_0^{\,l}}}\psi\left(a_0^{-l}t - kb_0\right), \quad \text{---(2)}$$

In order to have the feature of ortho- normality and compactness the coefficients can be taken as a0 = 2 and b0=1. If the above discretisation is on a dyadic grid then it is called standard DWT and can be expressed as:

$$\psi_{l,k}(t) = \frac{1}{\sqrt{2^l}}\psi\left(\frac{t-k2^l}{2^l}\right) = \frac{1}{\sqrt{2^l}}\psi\left(2^{-l}t - k\right) \quad \text{----- (3)}$$

2.2 Haar wavelet Transform

The Haar wavelet is the simplest possible wavelet. In mathematics, the Haar wavelet is a certain sequence of functions. It is renowned as the first known wavelet. This sequence was

proposed in 1909 by Alfred Haar. Haar used these functions to give an example of a countable orthonormal system for the space of square integrable functions on the real line. This wavelet operates on data by calculating the sums and differences of adjacent elements. The Haar Mother Wavelet function is defined as:

$$\psi_{Haar}(t) = \begin{cases} 1 & for\ 0 < t < 0.5 \\ -1 & for\ 0.5 < t < 1 \\ 0 & otherwise \end{cases}$$

------- (4)

Haar wavelet has finite no. of coefficients, hence it is very convenient to implement. Further, the orthogonality of Haar transform using a dyadic grid is an additional advantage for noiseless implementation. The Haar transform is performed in levels/scales. At each level, the Haar transform decomposes a discrete signal into two components with half of its length: an approximation (or trend) and a detail (or fluctuation) component. The first level of approximation $a^1 = (a1, a2, \dots, a_{n/2})$ is defined as:

$$a_m = \frac{X_{2m-1} + X_{2m}}{\sqrt{2}}$$

For, m= $1,2,\dots,n/2$, where X is the input signal. The multiplication of $\sqrt{2}$ ensures that the Haar transform preserves the energy of the signal. The values of a^1 represents the average of successive pairs of X value. The first level detail $d^1 = (d1, d2, \dots, d_{n/2})$ is defined as:

$$d_m = \frac{X_{2m-1} - X_{2m}}{\sqrt{2}}$$

For, m= $1,2,\dots,n/2$ .The values of d^1 represents the difference of successive pairs of X value.

2.3 Multi resolution analysis

Frequency analysis through conventional fixed window techniques such as the STFT are fixed window resolution operators in which the time duration of the analysis is inversely proportional to the bandwidth of the filters[12]. Likewise, high frequency localization results in poor time resolution as high time resolution results in poor frequency localization. In extracting features from a sampled speech waveform, it would be worthwhile to have a means to analyse the signal from a multi resolution perspective. Another motivation to pursue a multi resolution analysis of speech is that it somewhat models the cochlear mechanism of spectral decomposition during the initial stage of sound transduction in the ear, in which a time varying signal is spatially distributed in patterns along the basilar membrane. It has been shown that the nervous system processes spatially distributed patterns more efficiently than varying temporal signals [13]. Wavelets are based on mathematical constructs that deal with the linear expansion of a signal into contiguous frequency bands. Instead of analysing a signal with a single fixed window, as with short-time Fourier transform techniques, wavelets enable a signal analysis with multiple window durations that would allow for a coarse to fine multi resolution perspective of the signal [14]. The following figure descripts the role of different scales/resolutions for a simple sinusoid signal. The value of m depicts the different scales/levels of decomposition. The same signal is analysed at different scales and is providing different information at different levels.

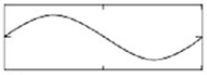

Fig. 1: Original Sinusoid Signal

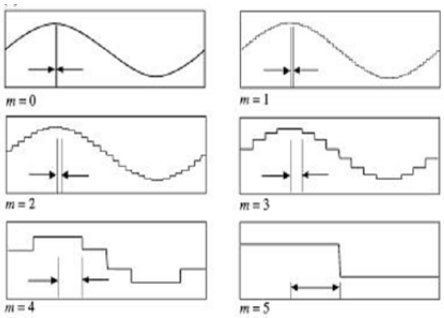

Fig. 2: Multi scale analysis of Fig 1

III. WAVELET FUSION

After obtaining the SAD information from the individual scales/resolutions, next step in the algorithm is to amalgamate the combined statistics, so that net result is obtained using some fusion strategy. The weighted mean fusion seems to be an appropriate technique to combine the information and resulting in a single conclusion. The technique is explained in the following equation:

$$f(x) = 1/n \sum_{i=1}^{n} i.x_i \qquad \text{--------- (5)}$$

Where, n stands for the total no. of scales used for wavelet and i=1, 2… n. x_i denotes the individual statistic SAD value obtained at a particular scale and f(x) resembles the final decision value after fusion.

IV. METHODOLOGY

The procedure followed for the proposed approach is portrayed in the following block diagram.

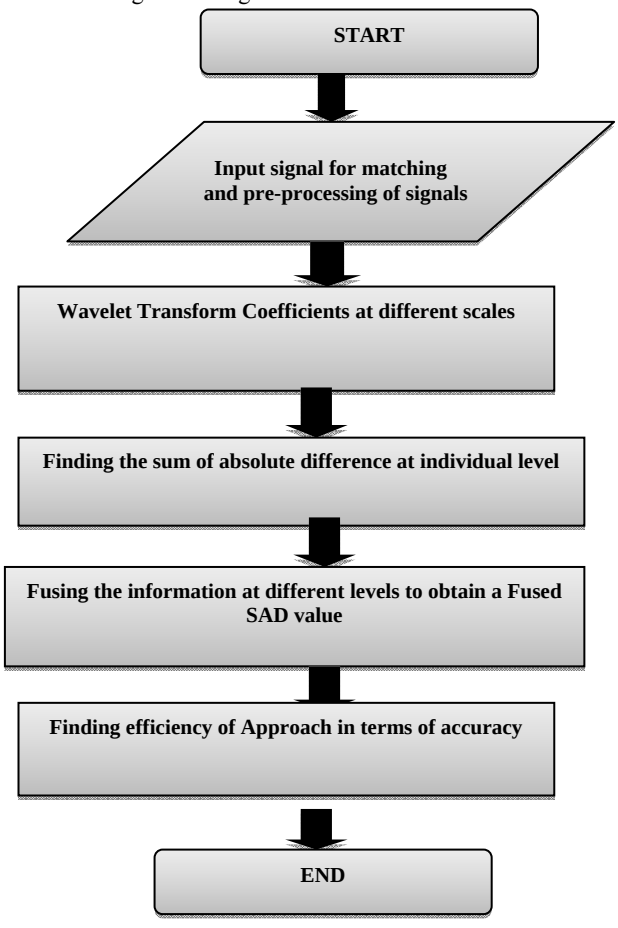

Fig. 3: Block Diagram illustrating proposed scheme

The details of the blocks are as follows:

Step 1: The first step before starting the matching procedure is to choose a particular singer track that is to be matched among an unknown database that contains mixed audio tracks of all the users/singers. A small pre-processing step is performed in which each track is normalised with the maximum intensity value in it. This helps in remedying the abnormalities in the different recorded tracks.

Step 2: Next thing is to find out the wavelet transform of all tracks individually. Haar transform for each signal is calculated at various scales in order to carry out the rest of the analysis. The procedure of finding the Haar transform has earlier been explained in the previous section.

Step 3: For finding the required parameter SAD, firstly the absolute value of each element of a track is obtained. Then the difference of absolute value of the input signal and all the unknown tracks in database is obtained and sum of all the elemental difference values for all the signals is obtained one by one respectively. The resultant that shows least SAD value is taken as the closest to the desired user vocal track. The above process is repeated with wavelet transform of vocal signal tracks, so that this time SAD values obtained through a high scale wavelet analysis. It is interesting to note that the direct analysis without applying the wavelet transform is analogous to the positive lowest scale value transform of the signal.

Step 4: After obtaining the wavelet domain features of the vocal tracks at different scales (1 to 17 in proposed work, because length of the sample taken approximately = 2 power 17) and finding the respective SAD values from these individual values, the net fused SAD value is decided by taking the weighted mean value of the individually obtained value. The technique of weighted mean and how the weights have been chosen is early explained in section 3 of this paper.

Step 5: After the information from various processing steps is acquired, final step is the analysis of data. For it, accuracy parameter is used. For a particular run of a unique singer vocal track with a small database, each time containing three different set of song samples, the result concluded on the basis of smallest SAD value. The individual accuracy of the result carried for 12 runs of experiment is found and a net accuracy is calculated for different type of matching.

V. RESULTS

The current problem was testified taking 4 samples each form 3 different users/singers hence making a case of 12 tracks in a database. And following are the cases taken for the testing of the validity of the algorithm. The feature or the parameter used for the testing of algorithm is chosen as Sum of Absolute Differences (SAD) that is calculated by firstly differencing the absolute value of the individual elements of a particular signal with another signal and then later summing up all the difference values obtained. The software used for performing the algorithm is MATLAB. The result section is further divided into three parts for showing various analysis schemes for a unchanged database of singers.

5.1 Direct analysis/Lowest Wavelet Scale Consequences

Table1 shows the performance when SAD is calculated without the application of any transform. This direct analysis is equivalent to lowest scale analysis of the wavelet transform, i.e., scale zero transform information. Her main emphasis is on analysing the approximate envelope of signal without considering the zooming details. As if the forest is seen from the window of aeroplane. Results show that the total accuracy for this analysis is equal to 66.67 %.

Table 1: Direct Analysis

TEST SAMPLE	TESTED DATABASE AND OBTAINED SAD VALUES			MOST SIMILAR SINGER
1a	**1b**	**2b**	**3c**	1b
	49.8196	55.57865	60.967	
	1c	**2b**	**2c**	2b
	61.05018	55.57865	61.22889	
	1d	**2b**	**3a**	3a
	51.31456	55.57865	33.08566	
	1d	**2c**	**3c**	1d
	51.3145	61.2288	60.967	
3c	**1c**	**1d**	**3b**	3b
	70.65986	64.31079	63.5081	
	1c	**2b**	**3a**	3a
	70.65986	65.78195	47.03106	
	2c	**1d**	**3b**	3b
	71.67904	64.31079	63.5081	
	3a	**1c**	**2c**	3a
	47.03106	70.65986	71.67904	
1c	**1b**	**2a**	**3b**	2a
	61.05481	59.1594	63.95788	
	1a	**2b**	**3c**	1a
	61.05018	67.1982	70.65986	
	2b	**1d**	**3c**	1d
	67.1982	61.32884	70.65986	
	3a	**1b**	**2c**	3a
	48.0841	61.05481	70.94665	
Accuracy Percentage				66.67

5.2 Highest Wavelet Scale Consequences

Table2 evaluates the SAD parameter when observations are achieved using high scale wavelet transform. Here another view of the signal, i.e., by taking the highest order transform of the signal (17th in this case because signal length is taken to be equivalent to 2^17 samples) is observed. This can be treated as the nearest view of a forest in which one is able to accumulate information when he is into the forest. In this case, highest level details are collected from the wavelet transform. Results here show that total average accuracy for this case is equal to 75 %.

Table 2: High Scale transform analysis

TEST SAMPLE	TESTED DATABASE AND OBTAINED SAD VALUES			MOST SIMILAR SINGER
1a	**1b**	**2b**	**3c**	1b
	20.95497	29.83038	27.81343	
	1c	**2b**	**2c**	1c
	27.66115	29.83038	32.83516	
	1d	**2b**	**3a**	3a
	24.22933	29.83038	14.93434	
	1d	**2c**	**3c**	1d
	24.22933	32.83516	27.81343	
3c	**1c**	**1d**	**3b**	3b
	31.27586	29.07722	27.19255	
	1c	**2b**	**3a**	3a
	31.27586	34.2744	20.48315	
	2c	**1d**	**3b**	3b
	36.30219	29.07722	27.19255	
	3a	**1c**	**2c**	3a
	20.48315	31.27586	36.30219	
1c	**1b**	**2a**	**3b**	2a
	25.35321	24.71562	27.42707	
	1a	**2b**	**3c**	1a
	27.66115	34.33696	31.27586	
	2b	**1d**	**3c**	1d
	34.33696	27.69766	31.27586	
	3a	**1b**	**2c**	3a
	20.25098	25.35321	36.09509	
Accuracy Percentage				75

Table 3: Analysis by Fusion

TEST SAMPLE	TESTED DATABASE AND OBTAINED SAD VALUES			MOST SIMILAR SINGER
1a	**1b**	**2b**	**3c**	1b
	170.8479	241.1037	225.5034	
	1c	**2b**	**2c**	1c
	224.3744	241.1037	265.0627	
	1d	**2b**	**3a**	3a
	196.2447	241.1037	121.399	
	1d	**2c**	**3c**	1d
	196.2447	241.1037	121.399	
3c	**1c**	**1d**	**3b**	3b
	253.7671	235.766	221.1421	
	1c	**2b**	**3a**	3a
	253.7671	276.8411	166.3973	
	2c	**1d**	**3b**	3b
	253.7671	276.8411	166.3973	
	3a	**1c**	**2c**	3a
	166.3973	253.7671	293.5101	
1c	**1b**	**2a**	**3b**	1b
	206.427	208.9804	223.0089	
	1a	**2b**	**3c**	1a
	224.3744	277.6727	253.7671	
	2b	**1d**	**3c**	1d
	277.6727	224.6264	253.7671	
	3a	**1b**	**2c**	3a
	164.845	206.427	291.7553	
Accuracy Percentage				83.33

5.3 Fusion Consequences

The final table3 highlights the results obtained by applying the proposed algorithm. Clearly one can see that total average accuracy in this case is obtained as 250/3 = 83.33%, which exhibits nearly 25% increase in the accuracy compared to the case of direct/low scale analysis and 11% increase in accuracy value compared to scenario of taking the high scale wavelet transform of the signal or choosing only the detail counterpart of the signal. The results clearly suggest that one view analysis of signal is not sufficient for accurate singer matching problem. Although, the high scale wavelet analysis provided certain increase in accuracy, but the maximum accuracy is obtained in the case of merging the information obtained at different scales and then inferring.

Here, the output from analysis is judged as correct if the most similar singer judged by the algorithm is true in reality and is termed as false if the similar singer comes out to be different. Hence accuracy is defined as no. of true outcomes divided by the total no. of cases.

VI. Conclusion And Future Scope

The proposed technique emphasis the use of wavelet transform and information fusion obtained from the multi-resolution nature of these transforms for vocal singer matching purposes. The method was tested for a database containing 12 songs from 3 different singers and the net increase in efficiency from direct analysis comes out to be 25% and multi-scale fusion further increased the accuracy of the task. The method if tested properly for music signal containing both vocal and instrumental audios can be a very handy tool in future for music matching since it does not use the pitch, MFCC parameters, etc. so provides a novel and simple approach of matching audio from a database having no prior information about it. Also, different fusion techniques can be applied in future to analyse and compare the effect of fusion strategies in accurate and precise matching of a singer.

REFERENCES

[1] M.I.Abdalla, H.S. Ali, "*Wavelet-Based Mel-Frequency Cepstral Coefficients for Speaker Identification using Hidden Markov Models*", Journal of telecommunications, pp-16-20, March 2010.

[2] K. R. S. Selva and N. S. Selva , "*Correction Robust speaker identification using vocal source information*", IEEE International Conference on Devices, Circuits and Systems (ICDCS), Coimbatore, pp-182-186, March 2012.

[3] Md.Sahidullah, S. Chakroborty and G. Saha, "*Improving Performance of Speaker Identification system using Complementary information fusion*", arXiv: 1105.2770v2 [cs.SD], 16 May 2011.

[4] M. K. Saini, S. Jain, "*Designing of speaker based wavelet filter*", IEEE International Conference on Signal Processing and Communication (ICSC), Noida, India, pp-261-266, December 2013.

[5] N. Baranwal and K. Datta , "*Peak Detection based Spread Spectrum Audio Watermarking using Discrete Wavelet Transform"*, International Journal of Computer Applications, pp-16-20,Vol.24 ,June 2011.

[6] R.Aggarwal, J.K.Singh, V.K.Gupta, "*Noise Reduction of Speech Signal using Wavelet Transform with Modified Universal Threshold*", International Journal of Computer Applications, pp-14-19, Vol.20, April 2011.

[7] A. Ellmauthaler, E.A.B. da Silva, C.L. Pagliari, M.M. Peraz, "*Multiscale Image Fusion Using the Undecimated Wavelet Transform With Non-Orthogonal Filter Banks*", IEEE Transactions on Image Processing, pp-1005-10017, December 2012.

[8] H. Zhu, B.Wu, P.Ren, "*AMedical Image Fusion Based on Wavelet Multi-Scale Decomposition*", Journal of Signal and Information Processing, pp-218-221, May 2013.

[9] H. Eldardiry, E. Bart, J. Liu, J. Hanley, B. Price, and O. Brdiczka, "*Multi-domain information fusion for insider threat detection,"* in Proc. IEEE Security and Privacy Workshops, San Francisco, California, USA. IEEE, pp. 45–51, May 2013.

[10] C. Garcia, T. D. Vu, O. Aycard and F. Tango, "*Fusion Framework for Moving-Object Classification"* IEEE International Conference on Information Fusion, Istanbul, pp. 1159–1166, July 2013.

[11] A.P. Grassy, V. Frolov and F. P. Leon, "*Information fusion to detect and classify pedestrians using invariant features",* Information Fusion, ELSEVIER, 2010.

[12] Gulick, W. L., Gescheider, G. A., and Frisna, R. D., Hearing: Physiological Acoustics, Neural Coding, and Psychoacoustics, Oxford University Press, 1989.

[13] Rioul, O. and Vetterli, M., Wavelets and signal processing, IEEE Signal Proc. Mag., 14, October 1991.

[14] Cody, M. A., The fast wavelet transform, Dr. Dobb's J., 16, April 1992