Exploring a new method for food likability rating based on DT-CWT theory

Ya'nan Guo, Jing Han, Zixing Zhang, Björn Schuller, Yide Ma

Angaben zur Veröffentlichung / Publication details:

Guo, Ya'nan, Jing Han, Zixing Zhang, Björn Schuller, and Yide Ma. 2018. "Exploring a new method for food likability rating based on DT-CWT theory." In *Proceedings of the 2018 on International Conference on Multimodal Interaction - ICMI '18, Boulder, CO, USA, October 16 - 20, 2018*, edited by Sidney K. D'Mello, Panayiotis (Panos) Georgiou, and Stefan Scherer, 569–73. New York, NY: ACM Press. https://doi.org/10.1145/3242969.3243684.

Exploring A New Method for Food Likability Rating Based on DT-CWT Theory

Ya'nan Guo*
ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany
yanan.guo@student.uni-augsburg.de

Jing Han
ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany

Zixing Zhang
GLAM – Group on Language, Audio & Music, Imperial College London, UK

Björn Schuller†
ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany

Yide Ma
School of Information Sci. Eng, Lanzhou University, China

ABSTRACT

In this paper, we mainly investigate subjects' food likability based on audio-related features as a contribution to EAT – the ICMI 2018 Eating Analysis and Tracking challenge. Specifically, we conduct 4-level Double Tree Complex Wavelet Transform decomposition of an audio signal, and obtain five sub-audio signals with frequencies ranging from low to high. For each sub-audio signal, not only 'traditional' functional-based features but also deep learning-based features via pretrained CNNs based on SliCQ-nonstationary Gabor transform and a cochleagram map, are calculated. Besides, the original audio signals based Bag-of-Audio-Words features extracted by the openXBOW toolkit are used to enhance the model as well. Finally, the early fusion of all these three kinds of features can lead to promising results, yielding the highest UAR of 79.2 % by means of a leave-one-speaker-out cross-validation, which holds a 12.7 % absolute gain compared with the baseline of 66.5 % UAR.

KEYWORDS

Likability Sub-Challenges, Double Tree Complex Wavelet Transformation, Pre-Trained CNNs

ACM Reference Format:
Ya'nan Guo, Jing Han, Zixing Zhang, Björn Schuller, and Yide Ma. 2018. Exploring A New Method for Food Likability Rating Based on DT-CWT Theory. In *2018 International Conference on Multimodal Interaction (ICMI '18), October 16–20, 2018, Boulder, CO, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3242969.3243684

*The author is further affiliated with School of Information Sci. Eng, Lanzhou University, China
†The author is further affiliated with GLAM – Group on Language, Audio & Music, Imperial College London, UK.

ACM ISBN 978-1-4503-5692-3/18/10…$15.00
https://doi.org/10.1145/3242969.3243684

1 INTRODUCTION

EAT – The ICMI 2018 Eating Analysis and Tracking challenge [7] is targeted at eating condition analysis. i. e., recognising whether people are eating while speaking, and if yes, what kind of food, the likability level, and chewing and speaking difficulty. Specifically, the challenge includes the following three sub-tasks:

- Food-type Sub-Challenge: Perform a seven-class food classification per utterance.
- Likability Sub-Challenge: Recognise the subjects' food likability rating.
- Chew and Speak Difficulty Sub-Challenge: Recognise the level of difficulty to speak while eating.

These topics are interesting given that eating is one of the basic activities of human beings, and speaking while eating is very common human behaviour in real life [8]. In this work, we focus on the food likability classification based on the audio modality, i. e., recognising the subjects' food likability rating from speech signals.

In the literature, there are merely a handful of studies on this research [7, 8]. For example, in [8] and [7], the acoustic features were extracted over the whole frequency bands, with an assumption that all frequency bands hold an equal importance. However, some research in the speech domain suggest that different frequency subbands indeed contribute differently for certain speech processing tasks, such as speech enhancement [1] and emotion recognition [4]. Motivated by these studies, we aim to decompose the original audio signals into five sub-audios covering from high to low frequency subbands by means of a 4-level Double Tree Complex Wavelet Transformation (DT-CWT). Then, we select the most informative sub-audio for a consecutive feature extraction.

Inspired by the effectiveness of deep scalogram representation for sound event classification [19], we further extract deep representations [19] as a new feature set, which is obtained by feeding the SliCQ-nonstationary Gabor transformation (SliCQ-NSGT) and cochleagram maps [30] that are transformed from sub-audio signals into pretrained Convolutional Neural Networks (CNNs) [15].

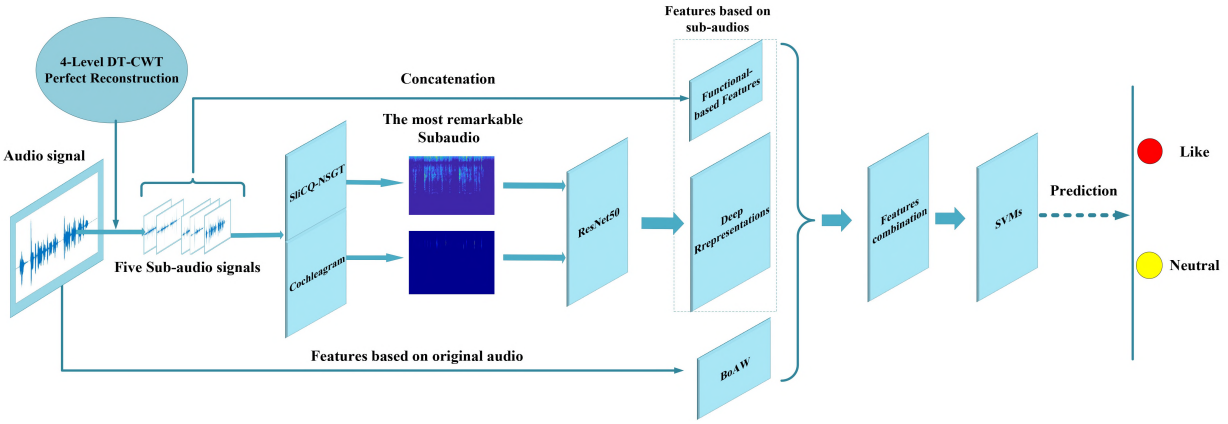

Figure 1: Flowchart of the proposed model.

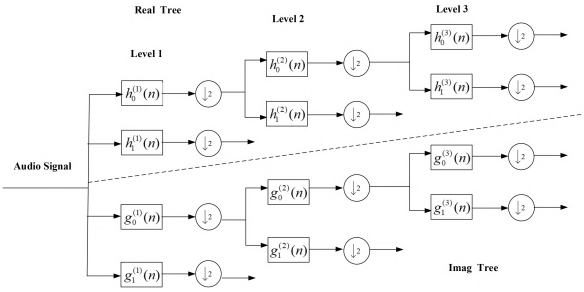

Figure 2: DT-CWT of a 1-D audio signal. $h_0(n)$ and $h_1(n)$ denote the low-pass and high-pass filter pair for the upper filter bank, respectively; $g_0(n)$ and $g_1(n)$ represent the low-pass and high-pass filter pair for the lower filter bank, respectively.

2 METHODOLOGY

Figure 1 illustrates the pipeline of our proposed model. Firstly, the original audio is decomposed into five sub-audios by DT-CWT approach. Then, the high-level deep representations are distilled by feeding the SliCQ-NSGT and cochleagram maps of remarkable sub-audio into a pre-trained Resnet50 CNN model. Meanwhile, traditional functional-based features are extracted on the sub-audio signals as well. Above two features sets as well as BoAW extracted on the original audio signals are finally fused and fed into SVMs for prediction. More details of the framework are described in the following sections.

2.1 Dual-Tree Complex Wavelet Transform

Wavelet-based features have already been applied successfully for many acoustic signal analysis tasks such as [3, 18, 25], because of their ability to better localise transients and higher common time-frequency resolution. Along with several further advantages, such features potentially suffer from some deficiencies such as lack of time invariance and directionality for two and higher dimensional signals, as well as oscillatory behaviour. Similarly, Wavelet packet transformation (WPT) [28] also suffers from two main problems [12]: 1) The shift-invariance of WPT is completely fulfilled only in a particular case, so it becomes obvious that the WPT cannot achieve shift-invariance; 2) the order of the packet is not the same as the order in the frequency area. That is, the features based on WPT are not always as robust and stable as expected. Even more, the greater effort for calculating wavelet coefficients is considered not worth the extra effort if gains are not outstanding. In order to cope with the above problems, complex wavelet transformation (CWT) has been proposed [11]. DT-CWT, which is a specific case of CWT proposed by Kingsbury in 1998, can be realised by different low-pass and high-pass filters based on two parallel Discrete Wavelet Transformations (DWT) [22]. Thus, it is shift-variance, and provides a low-directional selectivity in high dimensions [27], as well as a perfect reconstruction and an efficient computation. The wavelet function with a dual-tree complex wavelet is expressed as:

$$\psi(t) = \psi_h(t) + i\psi_g z(t)\,, \tag{1}$$

where $\psi_h(t)$ and $\psi_g z(t)$ indicate two real wavelets, i denotes the complex unit, and $\psi_h(t)$ and $\psi_g z(t)$ are a pair of Hilbert transforms. The complex function guarantees its better performance for non-stationary signals.

Figure 2 exhibits the DT-CWT decomposition of an audio signal, it can lead to twice the number of DWT wavelet coefficients and these wavelet coefficients are almost shift invariant. This means that a small change on the input signal cannot change the distribution of the energy of DT-CWT coefficients at different scales [29], so it holds promises for audio feature extraction [5]. In this work, we recommend 1-D DT-CWT for the food likability sub-task. The specific process is to divide the original audio into several frequency bands and let the audio information uniformly distribute at different sub-audios, effectively avoiding the flooding of useful information.

2.2 SliCQ-nonstationary Gabor Transform

As often observed, constant-Q non-stationary Gabor transform (CQ-NSGT) usually outperforms the state-of-the-art implementations of the 'classical' constant-Q transform for audio signals [13, 26], but it requires a super-linear complexity. The SliCQ transformation greatly improves the efficiency of the CQ-NSGT for sufficiently long signals, especially for audio signal processing, which was demonstrated in [27] by many experiments, we name this competitive transform as SliCQ-NSGT [1]. It is assumed that SliCQ-NSGT

[1] http://www.univie.ac.at/nonstatgab/slicq

coefficients can enhance the acoustic representations at a certain degree.

2.3 Cochleagram

Motivated by [23], cochleagram images are introduced in our model. A cochleagram is similar to a spectrogram, but it is built based on the human auditory model, so it is supposed to have potentially a better representation of auditory phenomena than other representations. It is a gamma-tone filter that determines the centre frequencies and bandwidth, which is expressed as [17]:

$$g(r) = Ar^{j-1}e^{-2\pi Br}cos(2\pi f_c r + \phi), \quad (2)$$

where A denotes the amplitude, B is the filter bandwidth, j stands for the filter order, r denotes the time, f_c and ϕ represent the centre frequency and phase respectively.

2.4 Extraction of Deep Representations

To extract the proposed deep representations from sliCQ-NSGT and cochleagram maps, transfer learning [16, 19, 24] is our choice, which can provide a trade-off between time and performance. Here, a pretrained ResNet50 [9, 10] model is introduced to extract deep representations because of its excellent feature extraction performance for images. The architecture of ResNet50 [14] is illustrated in Figure 3. Herein, we are interested in the outputs of the fully connected layer of 'fc1000' with 1 000 attributes. The final deep representations are the concatenation of most 'distinguished' sub-audio deep features in the SliCQ-NSGT and Cochleagram domains. The details on how to determine the 'distinguished' sub-audios can be found in Section 3.2.

2.5 Functional-based Features and Bag-of-Audio-Words

To obtain the segment-level functional-based features (aka OS features), we firstly extract the 35 frame-level Low-Level Descriptors (LLDs) on the sub-audio signals, including 3 temporal-domain features (i. e. zero crossing rate, energy, and entropy of energy), 19 spectral-domain features (i. e. Spectral & Spectral Centroid, Spectral Spread, Spectral Entropy, Spectral Flux, Spectral Roll-off, Harmonic Ratio, Fundamental Frequency, Chroma Vector, and etc.), 13 cepstral-domain features (i. e. Mel Frequency Cepstral Coefficients [MFCCs]). After that, four types of functional (i. e., mean, variance, skewness and kurtosis) are applied to the sequential frame-level LLDs, leading to 140 dimensional statistic features. Therefore, the final functional-based features contain 700=140*5 dimensional attributes given an audio segment.

Different from the functional-based feature extraction process, the BoAW extraction is performed on the original audios rather than the decomposed sub-audios. For the sake of comparison, we kept in line with the BoAW extraction procedure with the benchmark [7]. In more detail, the toolkit of openSMILE [6] is firstly utilised to extract the acoustic attributes (i. e., LLDs). After that, our another open toolkit openXBOW [20] is used to generate BoAW based on the sequential LLDs, with the codebook size of 1 200.

3 EXPERIMENTS

In our work, the deep representations, the functional-based features as well as the BoAW features were fed into SVM models. Before the training phase, the original features values were standardised. Furthermore, to reduce the feature dimension and select the most effective features, a feature selection approach called Chi-Squared test [2] was performed, and the feature selection information of the training set was applied to the test set. To keep in line with the benchmark [7], we used the linear kernel SVMs with the complexity of $[1e{-}4, 1e{-}3, \ldots, 1e{+}2]$ SVMs. Furthermore, Leave-One-Speaker-Out Cross-Validation (LOSO-CV) strategy was utilised to evaluate the model performance.

To evaluate the system performance, the UAR was officially recommended as the metric for the Likability Sub-Challenge because it is meaningful for highly unbalanced distributions of instances among classes compared with the weighted average recall (i. e., accuracy).

3.1 Database

In this challenge, the audio tracks of the audio-visual iHEARu-EAT database [21] were used, in which 30 subjects were recorded including 15 female and 15 male, with a mean age of 26.1 years (standard deviation: 2.7 years). Before the recording process, subjects performed practice trials to familiarise themselves with the recording procedure [7]. For the sub-task2 (i. e. Likability Sub-Challenge), the likability labels have been mapped to two discrete categories (i. e., 'Neutral', 'Like') considering the range of the individual subjects' ratings. For more details of this database, please refer to [7].

3.2 Experimental Results and Discussion

To select well-behaved sub-audios for the deep representation extraction, we separately collected the experimental results for SliCQ-NSGT and Cochleagram maps in Figures 4 and 5. Here, *sub*1-*sub*5 denotes the sub-audios with the frequency range from high to low; the bold values denote the highest UARs over the evaluated complexities of SVMs. From this table we can see that, not only for SliCQ-NSGT maps, but also for cochleagram maps, the sub-audio *sub*1 can reach the higher UAR values in all cases, and the other subbands appear to follow no discernible at all. Based on this observation, we use the combination of *sub*1 deep representations learnt from the SliCQ-NSGT and Cochleagram maps respectively as final deep-learnt representations. The results shown in the two figures explicitly tell us that not all the sub-audio signals hold the equal importance. In contrast, the most distinguished characteristic description might occur at a certain subband.

Table 1 shows the performance of three feature sets (i. e., BoAW features, functional-based features, and the deep representations) and their combinations based on a linear kernel SVM model with seven complexities. Notice that the BoAW features results deliver the best results in terms of UAR, yielding the highest UAR of 76.2 %. The performance gain demonstrates the effectiveness of the Chi-Squared feature selection method. However, the functional-based features based on sub-audios are not good enough, just reaching 61.6 % of UAR. However, when they are combined with BoAW, the obtained UAR can be up to 79.1 %, getting almost 3.0 % gain. Moreover, the proposed deep representations can only deliver 59.5 %

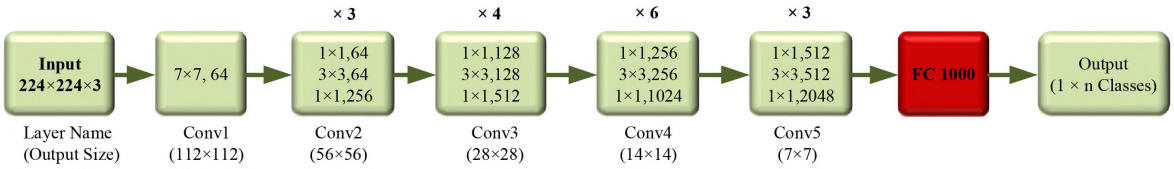

Figure 3: The ResNet-50 [14] architecture is shown with the residual units, the size of the filters and the outputs of each convolutional layer, and the 'fc1000' layer is our interest.

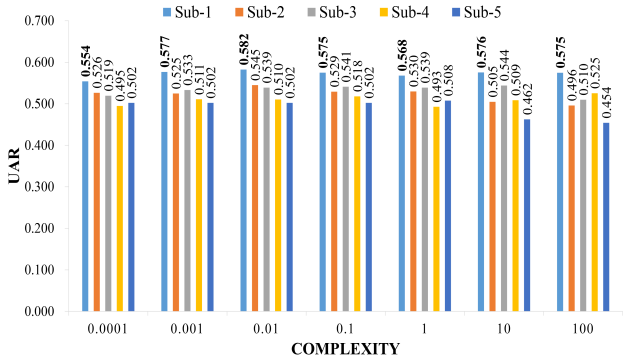

Figure 4: The obtained UARs with *Cochleagram* maps over five decomposed sub-audios in a variety of SVMs complexities.

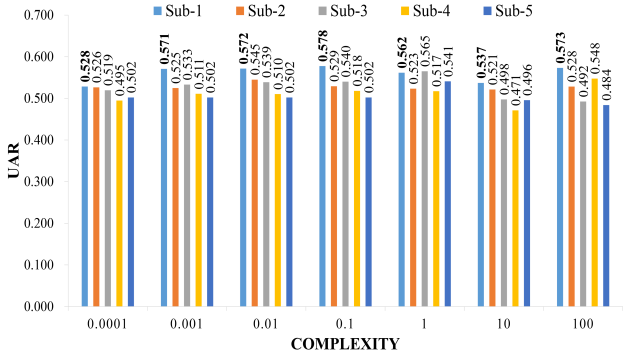

Figure 5: The obtained UARs with *SliCQ-NSGT* maps over five decomposed sub-audios in a variety of SVMs complexities.

in term of UAR, whereas it also achieves 3.0% improvement when combined with BoAW. The best performance of 79.2% is obtained where three kinds of features including functional-based features, deep representations, as well as BoAW are fused, which is 12.7% absolutely higher than the baseline of 66.5 %.

Although the prediction result of our model is higher than the baseline, it is the selected BoAW features that contribute most to the prediction results. According to above analysis, it seems that the proposed sub-audio-based features (i. e. the functional-based features and the learnt deep representations) in DT-CWT domain just can be regarded as the auxiliary features, rather than the core features, to improve the system performance for the food likability rating. Several reasons might lead to these findings. First, the employed Resnet50 CNN model is specifically trained on natural images and is designed to classify images into 1 000 object categories, e. g. the keyboard, mouse, pencil, and many animals, which is different from our task of a binary problem on audio data. Second, limited audio data are provided in this challenge, just 945 instances of the training set. Thus, it is really difficult for us to get robust representations through deep neural networks. Third, many parameters of the model need to be adjusted, and various parameters setting can make a great difference in the system performance.

Table 1: The obtained results in terms of UAR when using different feature types and their combinations. A: BoAW features; B: functional-based features; C: deep representations; '+': feature fusion, best results are highlighted.

Feature types	complexity						
UARs [%]	.0001	.001	.01	.1	1	10	100
A	67.9	74.7	**76.2**	76.0	75.6	75.1	75.0
B	49.9	52.2	59.4	60.3	**61.6**	60.9	60.0
C	54.9	59.2	**59.5**	58.1	59.0	58.7	55.9
A+B	68.9	75.9	77.7	**79.1**	78.3	78.5	78.5
A+C	68.6	76.1	**79.1**	79.0	78.1	78.2	78.2
B+C	57.7	59.7	**62.1**	62.0	59.1	58.2	59.0
A+B+C	68.8	75.8	78.6	**79.2**	78.8	78.9	78.9

With regard to above dilemmas, on the one hand, the multimodal method should be considered to improve our model; after all, the video can provide complementary information, and the facial expression that is very important to Likability Sub-Challenge. On the other hand, some burgeoning optimisation methods such as particle swarm optimisation and cat swarm optimisation algorithm should be introduced to overcome the parameter setting dilemma. Besides, re-investigating our model by combining some augmented data before training might lead to better system performance.

4 CONCLUSIONS

In summary, we have performed both an experimental and theoretical study of a food likability rating task. Several feature sets as well as their combinations are evaluated, and the experimental results have shown the effectiveness of the proposed features. The highest UAR is achieved at 79.2 %, which is 12.7 % absolutely higher than the baseline. From this challenge, we also find that it is very necessary to train CNN models based on audio and speech databases, because almost all the mature CNN models are designed and trained on natural images. Therefore, these CNN models might be not suitable for audio data upon most scenarios. Besides, parameters setting and feature selection methods should be further optimised in future.

ACKNOWLEDGMENTS

This work is partially supported by the China Scholarship Council (CSC).

REFERENCES

[1] Mohammed Bahoura and Jean Rouat. 2001. Wavelet speech enhancement based on the Teager energy operator. *IEEE Signal Processing Letters* 8, 1 (Jan. 2001), 10–12.

[2] RalphB D'Agostino. 2017. *Goodness-of-fit-techniques*. Routledge, Abingdon, UK.

[3] Charalampos A. Dimoulas and George M. Kalliris. 2013. Investigation of Wavelet approaches for joint temporal, spectral and cepstral features in audio semantics. In *Proc. 134th International Audio Engineering Society (AES) Convention*. Rome, Italy. no pagination.

[4] Bin Dong, Zixing Zhang, and Björn Schuller. 2016. Empirical Mode Decomposition: A Data-Enrichment Perspective on Speech Emotion Recognition. In *Proc. the 6th International Workshop on Emotion and Sentiment Analysis (ESA), satellite of the 10th Language Resources and Evaluation Conference (LREC)*. Portoroz, Slovenia, 71–75.

[5] Timur Düzenli and Nalan Özkurt. 2011. Discrete and dual tree Wavelet features for real-time speech/music discrimination. *ISRN Signal Processing* 2011 (Mar. 2011), 10 pages.

[6] Florian Eyben, Felix Weninger, Florian Gross, and Björn Schuller. 2013. Recent developments in openSMILE, the Munich open-source multimedia feature extractor. In *Proc. 21st ACM International Conference on Multimedia (MM)*. Barcelona, Spain, 835–838.

[7] Simone Hantke, Maximilian Schmitt, Panagiotis Tzirakis, and Björn Schuller. 2018. EAT – The ICMI 2018 eating analysis and tracking challenge. In *Proc. 20th ACM International Conference on Multimodal Interaction (ICMI)*. Boulder, CO. 5 pages.

[8] Simone Hantke, Felix Weninger, Richard Kurle, Fabien Ringeval, Anton Batliner, Amr El-Desoky Mousa, and Björn Schuller. 2016. I hear you eat and speak: Automatic recognition of eating condition and food types, use-cases, and impact on ASR performance. *PLoS ONE* 11, 5 (May 2016), e0154486.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, NV, 770–778.

[10] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, and Bryan Seybold. 2017. CNN architectures for large-scale audio classification. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. New Orleans, LA, 131–135.

[11] Nick G. Kingsbury. 1998. The dual-tree complex Wavelet transform: A new technique for shift invariance and directional filters. In *Proc. 8th IEEE Digital Signal Processing (DSP) Workshop*. Salt Lake City, UT, 120–131.

[12] Makoto Kobayashi and Kazushi Nakano. 2013. Two problems of Wavelet packet transform. In *Proc. 10th International Conference on Information Technology: New Generations (ITNG)*. Las Vegas, NV, 153–159.

[13] Tao Liu, Shaoze Yan, and Wei Zhang. 2016. Time–Frequency analysis of nonstationary vibration signals for deployable structures by using the constant-Q nonstationary Gabor transform. *Mechanical Systems and Signal Processing* 75 (June 2016), 228–244.

[14] Ammar Mahmood, Mohammed Bennamoun, Senjian An, and Ferdous Ahmed Sohel. 2017. ResFeats: Residual network based features for image classification. In *Proc. 24th IEEE International Conference on Image Processing (ICIP)*. Beijing, China, 1597–1601.

[15] Dimitri Palaz, Mathew Magimai.-Doss, and Ronan Collobert. 2015. Analysis of CNN-based speech recognition system using raw speech as input. In *Proc. 16th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. Dresden, Germany, 11–15.

[16] Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering* 22, 10 (Oct. 2010), 1345–1359.

[17] Roy D Patterson, KEN Robinson, John Holdsworth, Denis McKeown, C Zhang, and Michael Allerhand. 1992. Complex sounds and auditory images. In *Proc. 9th International Symposium on Hearing: Auditory physiology and perception*. Carcens, France, 429–446.

[18] Kun Qian, Zhao Ren, Vedhas Pandit, Zijiang Yang, Zixing Zhang, and Björn Schuller. 2017. Wavelets revisited for the classification of acoustic scenes. In *Proc. IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE)*. Munich, Germany, 1597–1601.

[19] Zhao Ren, Kun Qian, Zixing Zhang, Vedhas Pandit, Alice Baird, and Björn Schuller. 2018. Deep scalogram representations for acoustic scene classification. *IEEE/CAA Journal of Automatica Sinica* 5, 3 (Apr. 2018), 662–669.

[20] Maximilian Schmitt and Björn Schuller. 2017. OpenXBOW - Introducing the Passau open-source crossmodal bag-of-words toolkit. *Journal of Machine Learning Research* 18 (Jan. 2017), 1–5.

[21] Björn Schuller, Stefan Steidl, Anton Batliner, Simone Hantke, Florian Hönig, Juan Rafael Orozco-Arroyave, Elmar Nöth, Yue Zhang, and Felix Weninger. 2015. The INTERSPEECH 2015 computational paralinguistics challenge: Nativeness, Parkinson's & eating condition. In *Proc. 16th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. Dresden, Germany, 478–482.

[22] Ivan W Selesnick, Richard G Baraniuk, and Nick C Kingsbury. 2005. The dual-tree complex Wavelet transform. *IEEE Signal Processing Magazine* 22, 6 (Nov. 2005), 123–151.

[23] Roneel V Sharan and Tom J Moir. 2015. Cochleagram image feature for improved robustness in sound recognition. In *Proc. 20th International Conference on Digital Signal Processing (DSP)*. Singapore, 441–444.

[24] Aäron Van Den Oord, Sander Dieleman, and Benjamin Schrauwen. 2014. Transfer learning by supervised pre-training for audio-based music classification. In *Proc. 15th International Society for Music Information Retrieval Conference (ISMIR)*. Taipei, Taiwan, 29–34.

[25] V Nivitha Varghees and KI Ramachandran. 2017. Effective heart sound segmentation and murmur classification using empirical Wavelet transform and instantaneous phase For electronic stethoscope. *IEEE Sensors Journal* 17, 12 (Apr. 2017), 3861–3872.

[26] Gino Angelo Velasco, Nicki Holighaus, Monika Dörfler, and Thomas Grill. 2011. Constructing an invertible constant-Q transform with non-stationary Gabor frames. In *Proc. 14th International Conference on Digital Audio Effects*. Paris, France, 19–23.

[27] Yancai Xiao, Yi Hong, Xiuhai Chen, and Weijia Chen. 2017. The application of dual-tree complex Wavelet transform (DTCWT) energy entropy in misalignment fault diagnosis of doubly-fed wind turbine (DFWT). *Entropy* 19, 11 (Sep. 2017), 587.

[28] Gao Huan Xu and Jun Xiang Ye. 2012. Feature extraction of fault engine audio signal based on Wavelet packet transform. *Electrical Insulating Materials and Electrical Engineering* 546 (Sep. 2012), 675–679.

[29] Xiao-Chen Yuan, Chi-Man Pun, and CL Philip Chen. 2015. Robust Mel-frequency cepstral coefficients feature detection and dual-tree complex Wavelet transform for digital audio watermarking. *Information Sciences* 298 (Mar. 2015), 159–179.

[30] Xiao-Lei Zhang and DeLiang Wang. 2014. Boosted deep neural networks and multi-resolution cochleagram features for voice activity detection. In *Proc. 15th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. Singapore, 1534–1538.