Crossterm-Free Time-Frequency Analyses Exploiting Deep Neural Networks

S. Zhang, S. Pavel and Y. Zhang

Department of Electrical and Computer Engineering, Temple University, Philadelphia, PA, 19122, USA
{shuimei.zhang, pavel.saidur, ydzhang}@temple.edu

Nonstationary signals are naturally found and exploited in various applications, such as radar, sonar, radio astronomy, seismology, and electroencephalogram (EEG) [1–6]. Time-frequency (TF) analysis of nonstationary signals is a key enabling technology for radar-based human-machine interface through gesture recognition using hands and arms [7, 8]. Many nonstationary signals can be described as frequency modulated (FM) signals and characterized by their time-varying instantaneous frequencies (IFs). Compared to single-domain signal representations with respect to either time or frequency, joint TF-domain representations are better suited for the analyses and classification of such signals as they provide a time-varying spectrum [1, 2].

Commonly used TF representations (TFRs) can be classified into linear and bilinear approaches. Compared to their linear counterparts, bilinear TFRs generally provide higher TF concentration. However, they suffer from the existence of crossterms, which are artifacts arising from their bilinearity. To address this issue, various fixed and adaptive TF kernels have been developed to obtain reduced crossterm interference while preserving autoterms [1, 2, 11, 12]. The Wigner-Ville distribution (WVD) is often referred to as the prototype bilinear TFR because it does not apply a TF kernel. Essentially, a TF kernel acts as a two-dimensional low-pass filter multiplied in the ambiguity function domain, expressed with respect to time-lag and frequency-shift. TF kernels can be broadly classified into two types, namely, fixed (data-independent) kernels and adaptive (data-dependent) kernels. For example, the Choi-Williams distribution (CWD) [11] is a popular data-independent TF kernel, whereas the adaptive optimal kernel (AOK) [12] is a commonly used data-dependent TF kernel. It is noted that there is a trade-off between autoterm preservation and crossterm mitigation, i.e., crossterm mitigation is often achieved at a cost of compromised autoterm preservation. Designing a TF kernel function with satisfactory autoterm preservation and crossterm suppression performance has been a challenging task for the past several decades [1, 2]. Recently, sparsity-based methods also find attractive for TF analyses [9, 10].

In this poster abstract, we develop a novel approach to obtain high-resolution TFRs with crossterm effectively suppressed. In particular, we exploit a deep neural network (DNN) [13] which is trained to achieve crossterm-free TFRs. DNN has been successfully exploited in many applications, such as image recognition [14], natural language understanding [15], human motion recognition [7, 8], EEG interpretation [16], crack detection [17], and radar waveform recognition [18].

DNN Structures. The generic block diagram of the proposed DNN structures is depicted in Figure 1. The input to the DNN is a two-dimensional WVD image $\mathbf{X}$, whereas the corresponding crossterm-free TFR $\mathbf{Y}$, constructed from the actual instantaneous frequency law of the signal components scaled by their respective magnitudes, is used as the training label. The difference in the form of the mean square error between the DNN output $\hat{\mathbf{Y}}$ and the TFR label $\mathbf{Y}$ is taken as the loss function.

In this work, the following two neural network architectures are considered:

- **Fully convolutional neural network (FCNN):** Except for the last layer, $N-1$ convolutional layers are utilized, each exploiting C filters of size $D \times D$ and a rectified linear unit (ReLU) activation function. The last layer forms a single channel by performing $D \times D$ convolution without performing an activation operation. The convolutional stride is fixed as 1, and zero-padding is employed to keep the size of the feature maps unchanged after each convolution step. Increasing the number of layers can increase the receptive field and enable a larger TFR region. At

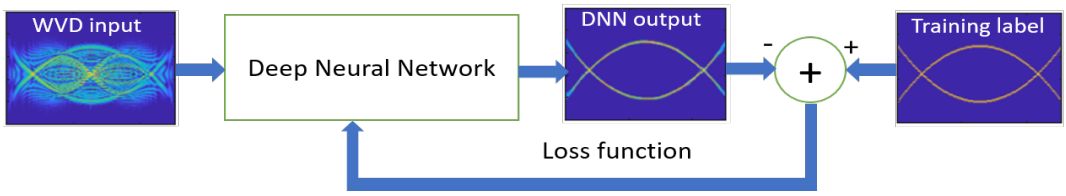

Figure 1. Proposed DNN architecture to achieve crossterm-free TFR.

the same time, it may cause the over-fitting problem and results in high computational burden. We select $N = 11$, $C = 40$, and $D = 5$ to well balance the complexity and the performance for the proposed FCNN.

- **Convolutional autoencoder (CAE):** The CAE consists of two parts, i.e., encoder and decoder. The encoder consists of N convolutional layer, each using C filters of size $D \times D$ followed by a ReLU and a max-pooling layer. Each decoder layer is a combination of a deconvolutional layer followed by a ReLU and an upsampling layer. The convolutional layers capture the abstraction of autoterms while eliminating crossterms, whereas the decovolutional layers upsample the feature maps and recover the autoterms details. The encoder and the decoder are symmetric and use the same hyperparameters. They both use $N = 3$ layers and each layer consists of 40 filters of size 5×5.

Network Training. We consider a two-component FM signal model, expressed as:

$$x(t) = e^{j\phi_1(t)} + e^{j\phi_2(t)}, \tag{1}$$

where $t = 0, 1, \cdots, T-1$. We assume $T = 128$, and the size of each input TFR image is 128×128. Two types of signals are considered: (i) two-component nonlinear FM signals and (ii) signals consisting of one linear FM component and one sinusoidal FM component. $2,000$ samples are randomly generated for each class with different parameters. 90% of the samples are utilized for training and the remaining 10% are utilized for validation. In both architectures, the optimizer implements the Adam algorithm with a learning rate of 0.001. No noise is considered in this poster abstract.

Numerical Examples. To demonstrate the effectiveness of the proposed method, we compare the results obtained using the FCNN and CAE networks with those obtained from the existing TF kernels, CWD and AOK. Both synthetic and real-world signals are considered. For synthetic signals, we consider the same two types of signals used for network training. For real-world signal, we consider the bat echolocation signal.

Two nonlinear parallel FM: In this case, the instantaneous phase laws of the two FM components are given as:

$$\phi_1(t) = 2\pi\left(0.25t - 0.15t^2/T + 0.15t^3/T^2\right), \quad \phi_2(t) = 2\pi\left(0.15t - 0.15t^2/T + 0.15t^3/T^2\right). \tag{2}$$

Figure 2(a) shows the TFR label image, and the corresponding WVD is shown in Figure 2(b). All the TFR results are depicted in dB level normalized to the peak values. It is clear that severe crossterms exist in the WVD. The TFRs obtained by the CWD and the AOK are respectively shown in Figures 2(c) and 2(d). It is clear that these kernels substantially mitigate the effects of crossterms, but there are still noticeable residual crossterm effects and, at the same time, the resolution of the autoterm TFR is compromised. Figures 2(e) and 2(f) present the result of the proposed TFRs using the FCNN and CAE, respectively. Both results provide high-resolution and crossterm-free TFR, and are very close to their respective label image. Compared to the FCNN in Figure 2(e), the TFR obtained from the CAE depicted in Figure 2(f) provides a slightly better IF reconstruction near the ends.

One sinusoidal FM and a linear FM: In this case, the instantaneous phase laws of the two signal components are given as:

$$\phi_1(t) = 2\pi\left((3T/50)\pi\cos\left(2\pi t/T + \pi\right) + 0.22t\right), \quad \phi_2(t) = 2\pi\left(0.10t + 0.12t^2/T\right). \tag{3}$$

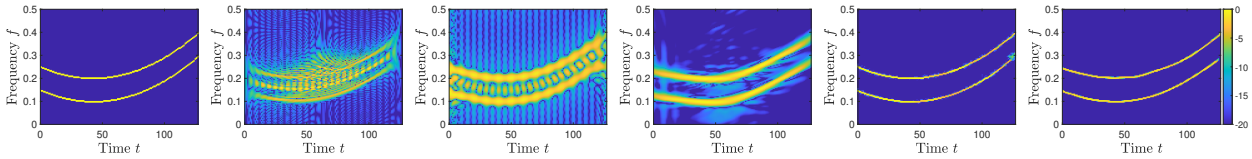

Figure 2. TFRs for a two-component nonlinear FM signal. (a) Label; (b) WVD; (c) CWD; (d) AOK; (e) Proposed (FCNN); (f) Proposed (CAE).

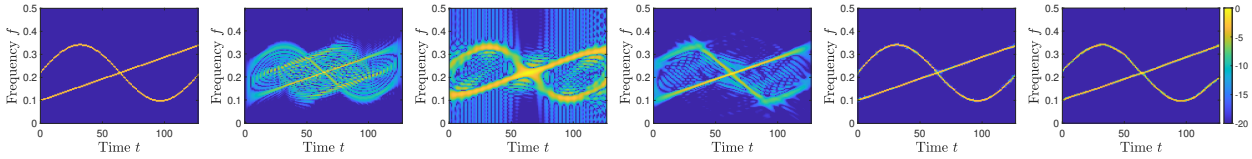

Figure 3. TFRs for a signal consisting of one sinusoidal FM component and one linear FM component. (a) Label; (b) WVD; (c) CWD; (d) AOK; (e) Proposed (FCNN); (f) Proposed (CAE).

Figures 3(a)–3(e) respectively present the label, the WVD, the CWD, the AOK and the proposed TFRs. Compared to the WVD shown in Figure 3(b), the crossterms are substantially suppressed after the kernels are applied, as shown in Figures 3(c) and 3(d). Compared to the case of the two-component linear FM signal depicted in Figure 2, the crossterms are much more complicated and difficult to handle. In this case, besides the existence of residual artifacts, the autoterms are distorted in both CWD and AOK results. In particular, for the CWD shown in Figure 3(c), the two signal components are distorted at their intersection. For the AOK shown in Figure 3(d), on the other hand, the sinusoidal FM signal component is distorted at the crest and the trough. As shown in Figures 3(e) and 3(f), the proposed method using both FCNN and CAE obtains high-fidelity TFRs with crossterm completely eliminated. Their results are very similar to the training label shown in Figure 3(a).

Bat Echolocation Signal: The recorded echolocation exponential chirp signal emitted by the bat Eptesicus fucus[1] consists of 400 samples with a sampling period of 7 μs. Figure 4 presents the TFRs obtained via different methods. It is noted that there is no label for this real-world signal. It is observed that the kerneled TFRs with CWD and AOK and the proposed TFRs all detect the three harmonics of the signal. However, for CWD shown in Figure 4(b), there are strip-like artifacts due to the windowing effect. Figure 4(c) depicts the TFR obtained by applying the AOK, which shows reduced energy preservation around the two ends of the autoterms. As shown in Figures 4(d) and 4(e), the proposed method with both network architectures not only detects all harmonic components but also maintains the signal energy levels.

Conclusion. We proposed a novel method to obtain crossterm-free TFR using a DNN-based model. Two network architectures are considered. Compared with the FCNN, the CAE achieves comparable performance with less number of layers, and the use of max-pooling further reduces the computational complexity. The proposed method offers desirable autoterm preservation and crossterm mitigation capabilities and significantly outperforms the commonly used fixed and adaptive kernels, such as the CWD and AOK.

REFERENCES

[1] L. Stankovic, M. Dakovic, and T. Thayaparan, *Time-frequency Signal Analysis with Applications*. Artech House, 2013.

[1] The authors wish to thank C. Condon, K. White, and A. Feng of the Beckman Institute of the University of Illinois for the bat data and for permission to use it in this paper.

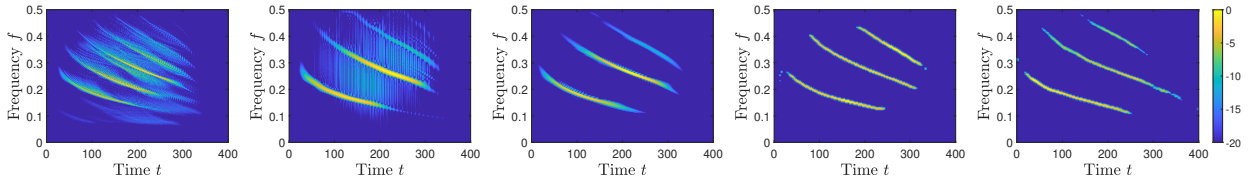

Figure 4. TFRs for a bat echolocation signal. (a) WVD; (b) CWD; (c) AOK; (d) Proposed (FCNN); (e) Proposed (CAE).

[2] B. Boashash (ed.), *Time-Frequency Signal Analysis and Processing, 2nd Ed.* Academic Press, 2015.

[3] G. Liu, S. Fomel, and X. Chen, "Time-frequency analysis of seismic data using local attributes," *Geophysics*, vol. 76, no. 6, pp. 23–34, 2011.

[4] B. Boashash, G. Azemi, and J. M. O'Toole, "Time-frequency processing of nonstationary signals: Advanced TFD design to aid diagnosis with highlights from medical applications," *IEEE Signal Process. Mag.*, vol. 30, pp. 108-–119, Nov. 2013.

[5] M. G. Amin, Y. D. Zhang, F. Ahmad, K. C. D. Ho, "Radar signal processing for elderly fall detection: The future for in-home monitoring," *IEEE Signal Process. Mag.*, vol. 33, no. 2, pp. 71–80, 2016.

[6] V. Shah, R. Anstotz, I. Obeid, and J. Picone, "Adapting an automatic speech recognition system to event classification of electroencephalograms," in *Proc. IEEE Signal Process. Medicine and Biology Symp.*, Philadelphia, PA, 2018, pp. 1–5.

[7] M. Wang, Y. D. Zhang, and G. Cui, "Human motion recognition exploiting radar with stacked recurrent neural network," *Digital Signal Process.*, vol. 87, pp. 125–131, April 2019.

[8] S. Gurbuz and M. G. Amin, "Radar-based human-motion recognition with deep learning: Promising applications for indoor monitoring," *IEEE Signal Process. Mag.*, vol. 36, no. 4, pp. 16–28, July 2019.

[9] M. G. Amin, B. Jokanovic, Y. D. Zhang, and F. Ahmad, "A sparsity-perspective to quadratic time-frequency distributions," *Digital Signal Process.*, vol. 46, pp. 175–190, Nov. 2015.

[10] S. Zhang and Y. D. Zhang, "Low-rank Hankel matrix completion for robust time-frequency analysis," *IEEE Trans. Signal Process.*, vol. 68, pp. 6171–6186, Oct. 2020.

[11] H. I. Choi and W. J. Williams, "Improved time-frequency representation of multi-component signals using exponential kernels," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 6, pp. 862–871, June 1989.

[12] D. L. Jones and R. G. Baraniuk, "An adaptive optimal-kernel time-frequency representation," *IEEE Trans. Signal Process.*, vol. 43, no. 10, pp. 2361–2371, Oct. 1995.

[13] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[14] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, "Learning hierarchical features for scene labeling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1915–1929, Aug. 2013.

[15] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," *J. Mach. Learn. Res.*, vol. 12, pp. 2493–2537, 2011.

[16] I. Obeid and J. Picone, "Machine learning approaches to automatic interpretation of EEGs," in E. Sejdik and T. Falk (Eds.), *Biomedical Signal Processing in Big Data*. CRC Press, 2017.

[17] L. Zhang, F. Yang, Y. D. Zhang, and Y. J. Zhu, "Road crack detection with deep convolution neural network," in *Proc. IEEE Int. Conf. Image Process.*, Phoenix, AZ, Sept. 2016.

[18] S. Zhang, A. Ahmed, and Y. D. Zhang, "Sparsity-based time-frequency analysis for automatic radar waveform recognition," in *Porc. IEEE Int. Radar Conf.*, Rockville, MD, April-May 2020.

Crossterm-Free Time-Frequency Analyses Exploiting Deep Neural Network

Shuimei Zhang, Md. Saidur Rahman Pavel, Yimin D. Zhang

Advanced Signal Processing Laboratory, Temple University, Philadelphia, PA, 19122

Advanced Signal Processing Laboratory
http://asplab.net

College of Engineering
Temple University

Introduction

- Nonstationary are commonly encountered in various applications, such as speech, biomedicine, vibration, and radar:
  - Finite-duration or transient
  - Signal with time-varying frequencies
- Time-frequency (TF) domain representation is an important tool to analyze nonstationary signals with time-varying spectrum.
- TF analysis methods can be classified into linear and bilinear: Bilinear TF analysis generally provides higher TF concentration but suffers from the existence of crossterms.

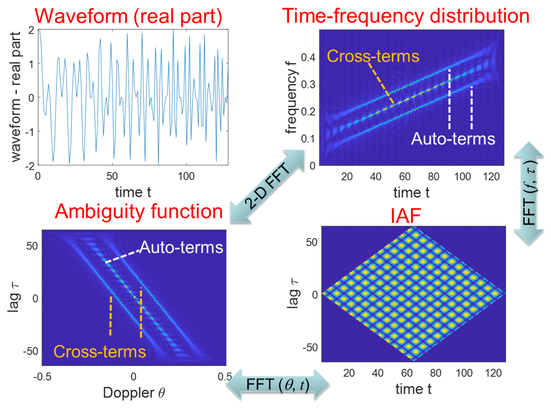

Fig.1 Illustration of autoterms and crossterms in bilinear TF representations

- Autoterms are concentrated around the origin of the ambiguity function (AF) domain, whereas crossterms tend to be scattered away from the origin.
- TF kernels function as two-dimensional low-pass filters to preserve autoterms while mitigating crossterms:
  - Data-independent kernels: such as Choi-Williams distribution (CWD)
  - Data-dependent kernels: such as adaptive optimal kernel (AOK)

Fig.2 Conflicting objectives between autoterm preservation and crossterm mitigation

Proposed Method

- Existing kernels minimize crossterms without assurance.
- We exploit a deep neural network (DNN) to train for high-resolution crossterm-free TF representations.
- Two neural network architectures are considered and compared: fully convolutional neural network (FCNN) and convolutional autoencoder (CAE).

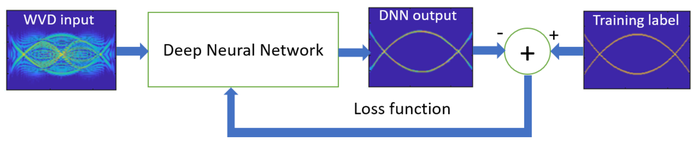

Fig.3 Generalized DNN architecture achieving crossterm-free TF representations

FCNN

- **Conv + ReLU**:
  - For layer $n = 1, \cdots, N-1$, C filters of size $D \times D$ are used to generate C feature maps.
  - Rectified linear unit (ReLU, $\max(0, \cdot)$) is followed to introduce nonlinearity and enhance TF-domain sparsity by mitigating negative outputs.
- **Conv :**
  - For the last layer, only a single filter of size $D \times D$ is utilized to reconstruct the output TF representation.
  - No active function is included in the last layer, so as not to limit the range of the output values.

CAE

- **Encoder (Conv + ReLU + Max-Pooling):**
  - For each convolutional layer $n = 1, \cdots, N$, C filters of size $D \times D$ are used to generate C feature maps.
  - ReLU helps enhance TF-domain sparsity.
  - Max-pooling reduces the input image size, adds translational invariancy to the feature maps and keeps the most prominent features by avoiding trivial solutions.
- **Decoder (Deconv + ReLU + Up-Sampling):**
  - Deconvolution performs the reverse option of the convolution layer.
  - Up-sampling layer restores the size of the input image.
  - The network parameters are the same as those in the encoder.

Network Training

We consider a two-component frequency-modulated (FM) signal model as

$$x(t) = \exp\left(j\phi_1(t)\right) + \exp(j\phi_2(t))$$

where $t = 0,1,\cdots,T-1$ with $T = 128$. Each input TF representation image has a size of 128×128.

- Two types of signal are considered: (a) two-component nonlinear FM signals, and (b) signals consisting of one linear FM and one sinusoidal FM.
- The training label is constructed from the actual instantaneous frequency law of the signal components scaled by their respective magnitudes.
- 2,000 samples for each class with different parameters. 90% for training and 10% for validation.
- For FCNN, number of layers: $N = 11$, number of filters: $C = 40$, filter size: $D = 5$.
- For CAE, number of layers: $N = 3$, number of filters: $C = 40$, filter size: $D = 5$.

Simulation Results

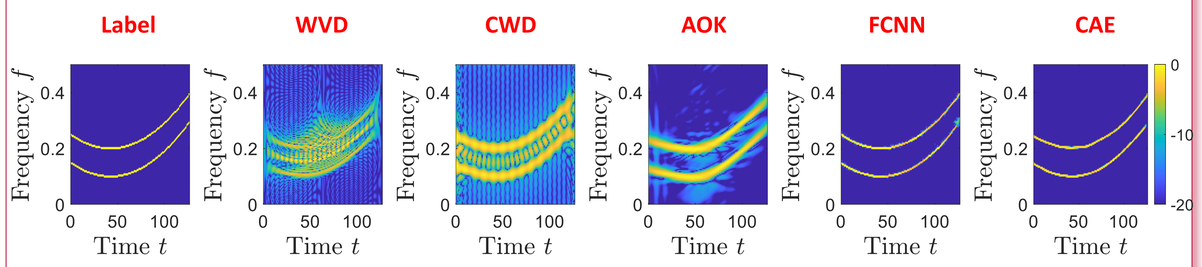

Fig. 4 Two nonlinear FM components

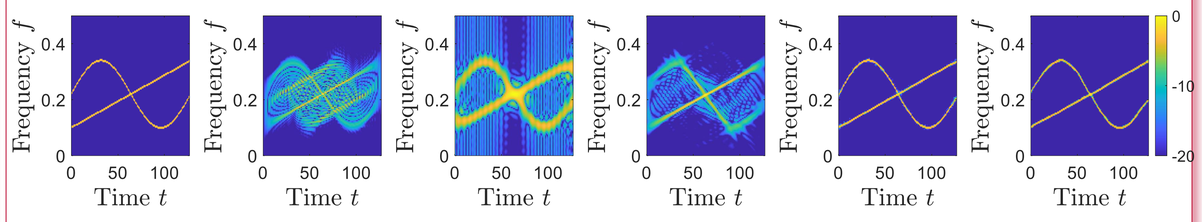

Fig. 5 One sinusoidal FM and one linear FM components

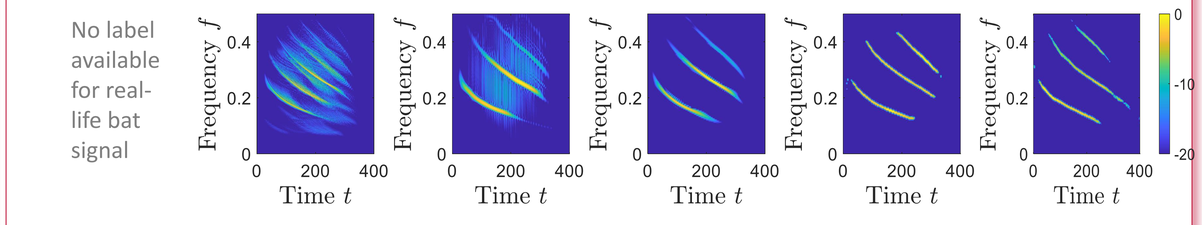

Fig. 6 Bat echolocation signal

Conclusion

- The proposed method offers desirable autoterm preservation and crossterm mitigation capabilities that are unmatched by existing kernels, such as the CWD and AOK.
- Compared with the FCNN, the CAE architecture achieves comparable performance with a smaller number of layers, and the use of max-pooling further reduces the computational complexity.