High Resolution Time-Frequency Generation with Generative Adversarial Networks

Zeynel Deprem, A. Enis Çetin, Fellow, IEEE

Abstract—Signal representation in Time-Frequency (TF) domain is valuable in many applications including radar imaging and inverse synthetic aperture radar. TF representation allows us to identify signal components or features in a mixed time and frequency plane. There are several well-known tools, such as Wigner-Ville Distribution (WVD), Short-Time Fourier Transform (STFT) and various other variants for such a purpose. The main requirement for a TF representation tool is to give a high-resolution view of the signal such that the signal components or features are identifiable. A commonly used method is the reassignment process. Through examples, it is shown that the method generates high-resolution TF representations which are better than the current reassignment methods.

Index Terms—Time-frequency representation, generative adversarial neural networks, CGAN.

I. INTRODUCTION

JOINT Time-Frequency (TF) representation of a signal reveals many features which are valuable in many applications including radar imaging [1] and inverse synthetic aperture radar (ISAR) [2] in which a high-resolution joint TF analysis is necessary to obtain a focused image of the target. TF representation is also valuable in many sound classification [3] and recognition [4] applications. In many Convolutional Neural Network (CNN) sound applications [5], TF representations boosts the performance of the Network.

The classical tool for the TF representation is the Wigner-Ville Distribution (WVD) [6]. Though it has many good features, it suffers from cross terms which makes the readability of different signal parts on TF plane difficult. A generalization of the WVD is the Cohen’s Class [7] distribution given by,

\[
W_x(t, f) = \int_{-\infty}^{+\infty} x(t + \tau/2)x^*(t - \tau/2)e^{-j2\pi f\tau}d\tau.
\] (1)

WVD has many good features, such as high resolution and is a real valued distribution but due to its quadratic nature it produces unwanted artifacts, named as cross terms, on the TF plane. This makes the readability of different signal parts on TF plane difficult. A generalization of the WVD is the Cohen’s Class [11] distribution given by,

\[
P_x(t, f) = \int_{-\infty}^{+\infty} A_x(\theta, \tau)\Phi(\theta, \tau)e^{-j2\pi(\theta t + f\tau)}d\theta d\tau.
\] (2)

where \( A_x(\theta, \tau) \) is the ambiguity function (AF) of the signal \( x(t) \) and has a two-dimensional (2D) Fourier transform relation with WVD. \( \Phi(\theta, \tau) \) is the kernel of the distribution and has the time-frequency shaping or smoothing effect on \( W_x(t, f) \). A proper kernel can attenuate the cross terms and produce high resolution distribution [12], [13], [14].

Another classical tool for the TF analysis is the Short-Time Fourier Transform (STFT) given by

\[
STFT_x(t, f) = \int_{-\infty}^{+\infty} x(\tau)w^*(\tau - t)e^{-j2\pi f\tau}d\tau,
\] (3)

where, \( w(t) \) is the time domain window having the smoothing effect on frequency domain. Spectrogram given by \( S_x(t, f) = |STFT_x(t, f)|^2 \) is a member of Cohen’s Class and depending on the length of the window, can attenuate the cross terms. But there is a trade off between the cross term suppression and the TF resolution or localization.

As a remedy to the resolution problem, the reassignment method is developed [8]. In reassignment, the value of a smoothed version, \( P_x(t, f) \) of WVD is moved to a new location \((t, f)\) which is the center of the gravity of the signal energy distributed around \((t, f)\). In this way, a better localization is obtained. In Figure 1 the ideal TF, the WVD, a smoothed version of WVD, called Smoothed Pseudo WVD (SPWVD) and the reassignment of SPVWD (RSPWVD) are shown for a sample signal. As can be seen WVD is not able to reveal the three TF signal components. The SPWVD is able to show them but with a reduced resolution. But RSPWVD
random vector representing the latent space whose probability distribution function (pdf) \( p_Z(Z) \) is given a priori. The generator network \( G \) is trained to generate an image according to its input vector \( Z \), \( E_{Y \sim p_Y(Y)} \) and \( E_{Z \sim p_Z(Z)} \) are the expectation over data samples and the random variable \( Z \), respectively. The network is trained in an alternating manner. In one turn \( D \) is fixed and parameters of \( G \) are trained, in the next round, \( G \) is fixed and \( D \) is trained. The aim of GAN is to estimate the pdf \( p_Y(Y) \) of data so that, when fed by a random variable \( Z \) the samples generated by \( G \) are indistinguishable from actual data samples. In other words, the generator is trained to fool the discriminator. On the other hand, the discriminator is trained not to be fooled. The beauty of GANs come from the definition of their cost function. Rather than measuring the distance between the desired and the generator \( (G) \) output with a predefined cost function the generator output is classified either as real or fake via the discriminator network \( D \). In this way, the parameters of of the generator network are "learned" during training process.

A variant of the original GAN, which has proved success in image prediction and generation, is the Conditional GAN (CGAN). In CGANs the generation process is based on a given condition. During training, together with the random variable \( Z \), an input \( X \), which define a condition on what is to be generated, is fed both to the generator and the discriminator. In this training approach, the generator is able to generate samples which are restricted by the condition \( X \). In CGAN, rather than the prior distribution \( p_Y(Y) \), the conditional pdf \( p_{Y|X}(Y|X) \) is estimated.

IV. TF GENERATION WITH CGAN

In this section, we consider the TF plots as two-dimensional images and generate reassigned TF plots from SPWVD distributions using CGANs. Even though the aim of CGAN is to estimate a conditional pdf \( p_{Y|X}(Y|X) \), the experimental results shows that, the CGANs produce little stochasticity at the output and the result is nearly deterministic, hence \( p_{Y|X}(Y|X) \) is impulse like \( \delta \). Therefore, in some implementations the stochasticity, rather than using the random variable \( Z \), is introduced in the form of "dropout" in generator implementation. In fact, pix2pix \( \delta \), which is developed for general image to image translation applications, is implemented in this manner. In this CGAN, an image which represent the condition \( X \) is fed to the generator, and the corresponding output is evaluated by the discriminator to be real or fake compared to a ground true or desired image \( \delta \). Starting with this intuition, a high resolution TF generation method is proposed with CGAN as shown in Figure 2 where the condition \( X \) or the input image to the generator, is taken as the TF distribution \( P_X(t,f) \) of time domain signal \( x(t) \). \( P_X(t,f) \) can be any smoothed version of WVD. In this paper, the smoothed pseudo WVD (SPWVD) distribution given by

\[
P_X(t,f) = \int_{-\infty}^{+\infty} h(\tau)g(s-t)R_x(s,\tau)e^{-j2\pi f\tau}dsd\tau, \quad (5)
\]

is used, where \( R_x(s,\tau) = x(s+\tau/2)x^*(s-\tau/2) \) is the instantaneous signal auto correlation, \( g(t) \) and \( h(t) \) are the time and frequency smoothing windows, respectively. SPWV
where \( a_1(t)e^{j\phi_1(t)} \) is the \( k^{th} \) component of the signal, \( a_k(t) \) is the amplitude and \( \phi_k(t) \) is the phase. Though not all the signals with time-varying frequency content can be expressed in this form, most practical ones are in this form. Then the TF trajectory of the \( k^{th} \) component can be expressed as,

\[
I_k(t, f) = a_k^2(t)\frac{1}{2\pi}d\left(f - \frac{d\phi_k(t)}{dt}\right)
\]

where \( \delta() \) is the Dirac delta function and \( \frac{d\phi_k(t)}{dt} \) is the instantaneous frequency (IF). We can express the ideal TF trajectory of the signal \( x(t) \) as

\[
I_x(t, f) = \sum L_{k=1}^L I_k(t, f)
\]

Therefore, we can construct a training set for CGAN based signals whose ideal TF trajectory is known. Together with SPWVD, \( P_x(t, f) \) of the signal, we can train the CGAN setup in Figure 2. After training, the aim is to generate a high resolution TF representation for an arbitrary signal. The benchmark will be the ideal TF trajectory.

The pix2pix model is a modified version of the CGAN. First, it does not use any random variable and uses the condition \( X \) only as the input. The randomization is introduced as dropout in generator. Second, Pix2pix uses regularization by the \( \ell^1 \) norm and uses the cost function given by, \ref{eq:10}

\[
\min_G \max_D V_1(D, G) = V_C(D, G) + \lambda E_{X,Y}[\| Y - G(X) \|_1]
\]

where the last term regulates the generator output with the \( \ell^1 \) norm. The regularization is arranged with the hyper-parameter \( \lambda \). Both the generator and discriminator in pix2pix, uses the U-Net structure \[23\]. Another variation in pix2pix implementation of CGAN is in the discriminator. The discriminator makes the real or fake decision with a PatchGAN structure where the image is divided into \( N \times N \) patches. Each patch is decided whether to be real or fake and then the average of all decisions are used to give final decision about whole image.

V. TRAINING AND TEST RESULTS

In order to test the effectiveness of the proposed method, a training set of 1320 signals and related SPWVD \( P_x(t, f) \) and ideal TF distributions \( I_x(t, f) \) were constructed using the TF toolbox \[24\]. We also used data augmentation to obtain input-output training pairs. The augmentation was done by first obtaining a smaller set and then applying various jittering operations. The time-domain signals have 256 samples. As a result the corresponding TF distributions are 256 \times 256 images. In addition, 120 test signals are generated for the validation purposes. The pix2pix network \[10\] was implemented using the Tensorflow Artificial Intelligence (AI) library \[25\]. The \( \ell^1 \) regularization parameter was set to \( \lambda = 100 \) as selected by \[10\]. Similarly the Adam optimizer was selected with learning rate \( \alpha = 0.0002 \), \( \beta_1 = 0.5 \) and \( \beta_2 = 0.999 \). The training was carried out for 50 epochs over the training set with batches of 10 images. In Figures \[3\] and \[4\] the result for two test signals are shown. Part (a) is the ideal TF representation of the signal \( I_x(t, f) \) and (b) is the SPWVD \( P_x(t, f) \). Part (c) shows the TF obtained by reassignment of SPWVD, namely RSPWVD. Part (d) is the proposed TF distribution obtained by pix2pix CGAN generator. Based on the Figures \[3\] and \[4\] it is clear that pix2pix CGAN produces high resolution TFs whose performance is comparable to the ideal TF. In order to assess the performance of the proposed CGAN based method a quantitative comparison is also carried out and the results are shown in Table 1. The first 10 rows \((x_1, x_2, \ldots, x_{10})\) shows the result for 10 test signals. The bottom line is obtained by averaging the results for 120 test signals. Three metrics are considered. The first metric is the Pearson correlation (pc) coefficient which is given by,

\[
pc = \frac{P_x^T I_x}{\| P_x \|_2 \| I_x \|_2}
\]

where \( P_x \) and \( I_x \) are the vector form of related discrete TF matrices with the mean subtracted. Pearson correlation measures the shape similarity between the vectors. The higher the Pearson correlation the better the similarity. The arrow next to each metric indicates the desired direction. The second metric is the \( \ell^1 \) difference between the method and the ideal TF. The
TABLE I: Performance comparison for SPVWD, RSPWVD and CGAN methods.

|       | Ideal TF | WVD      | SPWVD    | RSPWVD   | CGAN TF   |
|-------|----------|----------|----------|----------|-----------|
|       | pc ↑ ℓ¹ ↓ R ↓ | pc ↑ ℓ¹ ↓ R ↓ | pc ↑ ℓ¹ ↓ R ↓ | pc ↑ ℓ¹ ↓ R ↓ | pc ↑ ℓ¹ ↓ R ↓ |
| x₁    | 1 0 9.82 | 0.27 1735 12.79 | 0.43 3606 12.88 | 0.52 820 10.80 | 0.49 753 10.32 |
| x₂    | 1 0 9.64 | 0.46 816 10.82 | 0.57 1948 12.07 | 0.60 424 9.76 | 0.84 265 9.89 |
| x₃    | 1 0 9.71 | 0.36 1067 11.66 | 0.56 1975 12.12 | 0.54 536 10.06 | 0.78 328 9.85 |
| x₄    | 1 0 8.64 | 0.29 1036 12.52 | 0.34 1823 12.22 | 0.57 392 9.97 | 0.34 361 9.34 |
| x₅    | 1 0 11.02 | 0.27 3757 13.87 | 0.49 5039 13.51 | 0.59 1524 11.37 | 0.67 1325 11.29 |
| x₆    | 1 0 9.49 | 0.20 2052 12.80 | 0.44 2667 12.40 | 0.42 719 10.71 | 0.49 611 9.78 |
| x₇    | 1 0 9.87 | 0.30 1570 12.58 | 0.48 3056 12.64 | 0.47 924 10.72 | 0.73 535 10.09 |
| x₈    | 1 0 10.04 | 0.41 2379 12.80 | 0.55 3225 12.47 | 0.52 846 10.29 | 0.71 656 10.44 |
| x₉    | 1 0 10.07 | 0.39 1927 13.06 | 0.45 4537 13.05 | 0.61 849 10.65 | 0.59 869 10.55 |
| x₁₀   | 1 0 10.10 | 0.31 2381 13.21 | 0.52 3352 12.70 | 0.64 797 10.57 | 0.74 615 10.37 |
|       | ... ... ... ... ... | ... ... ... ... | ... ... ... ... | ... ... ... ... | ... ... ... ... |
| Avg(120) | 1 0 9.82 | 0.33 1852 12.60 | 0.48 3075 12.58 | 0.55 771 10.48 | 0.64 610 10.16 |

Fig. 3: Comparison of TF distributions for a signal $x_1$ with three methods.

Fig. 4: Comparison of TF distributions for a signal $x_2$ with three methods.

Fig. 5: Comparison of TF distributions for a signal obtained from Dolphin’s click signal.

VI. CONCLUSION

A method which generates high resolution TF representations was obtained using CGAN. The method uses SPWVD of the signal. Through examples it was shown that the performance is better than the existing methods, in particular the method gives better results than the reassignment method.
