A Novel Tracker of Adaptive Directional Ridge Separation and Prediction for Detecting Whistles

Yongchun Miao, Jianghui Li, *Member, IEEE*, and Yingsong Li, *Member, IEEE*

***Abstract*—Whistle detection of marine mammal signals with close and overlapping components of varying amplitudes is a key task for overlapping source separation. In this article, we propose a novel tracker, called adaptive directional ridge separation and prediction, for detecting whistles, which are typically analyzed using a time-frequency (TF) representation. Inspired by TF reassignment, a new reassignment scheme based on time-scale changes is developed to acquire instantaneous TF points with high energy concentration. To address the mutual interference among various types of components, a tone-pulse separation model is introduced for the aliased TF components, utilizing these instantaneous TF points and instantaneous rotating operators. An adaptive directional ridge predictor is established for application in automatic overlapping whistle detection, ensuring unbroken detection even when a whistle becomes nearly indistinguishable in the TF representation. Experimental results, obtained using both a simulated signal and recorded calls of marine mammals, demonstrate the superiority of the proposed method compared to other state-of-the-art methods. This method is capable of performing whistle detection and separating overlapping sources even in the presence of splash noises, which may cause partial distortion or disconnection of components from the TF representation.**

***Index Terms*—Adaptive directional ridge prediction (ADRP), instantaneous rotating operator (IRO), time-frequency (TF) representation, tone-pulse separation (TPS), whistle detection.**

I. Introduction

PASSIVE acoustic monitoring (PAM) of marine mammals is a growing research field, which requires the detection and identification of tonal calls of the species of interest, where the main purpose is to ensure the protection of species. A PAM system can provide many different functions, such as underwater acoustic (UWA) signal classification [1], sound event detection [2], [3], and overlapping source separation [4]. Automatic detection methods of calls are used to automatize these functions. These calls can be divided into three types according to their time-frequency (TF) properties: whistles (tonal calls), clicks (pulsed calls), and background noise. Whistles are suitable for communication and social interactions and are used in mammal languages, such as that of whales and dolphins. In this work, we focus on the analysis and detection of whistles mixed with interfering calls of other types, and the proposed method can be applied to overlapping source separation.

Generally, many UWA signals can be modeled as multicomponent (amplitude-modulated and frequency-modulated components) signals of the form: $s(t) = \sum_{m=1}^{M} a_m(t) e^{j2\pi \int_0^t f_m(\tau) d\tau}$ where $a_m(t)$ and $f_m(\tau)$, respectively, stand for instantaneous amplitude and instantaneous frequency (IF) of the mth component, M is the total number of signal components [5]. IFs of UWA signals have a complex structure of many nonlinear TF components. For this reason, TF representations (TFRs) are useful for the analysis of these UWA signals containing multiple time-varying whistles. The appearance of such whistles in the TFR is a frequency peak that varies continuously with time, forming a ridge. To estimate the IF parameters of such whistles, one first needs to extract its associated ridge. Multiple ridges can be directly estimated by using the properties of TFRs obtained by short-time Fourier transform (STFT), Chirplet transform [6], [7], etc. Unfortunately, the aforementioned way applies only to multiple components that are separable in the TF domain.

Recently, most automated algorithms have been developed to extract ridges from TFRs of a signal whose components overlap each other. Some of these are semiautomated methods, such as intrinsic chirp component decomposition [8], random sample consensus [9], relevant ridge portions [10], fast IF tracking algorithm [11], and kernel sparse learning [12], where prior information needs to be provided to assist these methods, such as ridge endpoints and the number of components. This is not a trivial issue, since in real cases the endpoints and number of components often vary. To avoid relying on such prior information, alternative methods, such as automatic ridge extraction [13], [14] and the Kalman filter [15], aim to find as many ridges as possible and then merge them. By using the amplitude and direction information from the TFR, the extracted ridge segments can be effectively merged, especially at intersecting points when overlaps occur [11], [13]. However, if pulsed calls are present in the recorded UWA signals, they can disrupt these methods for detecting whistles. Particle filter and graph search algorithms [2], [16] can provide effective measures

Received 11 September 2023; revised 17 February 2024, 8 April 2024, and 28 April 2024; accepted 30 April 2024. Date of publication 3 October 2024; date of current version 14 January 2025. This work was supported in part by the Key Laboratory of Southeast Coast Marine Information Intelligent Perception and Application, MNR under Grant 220103, and in part by the Anhui Provincial Natural Science Foundation, China under Grant 2208085QF180, and in part by the National Natural Science Foundation of China under Grant U23A20290. *(Corresponding authors: Yingsong Li; Jianghui Li.)*

Associate Editor: W. Xu.

Yongchun Miao and Yingsong Li are with the Key Laboratory of Intelligent Computing and Signal Processing, Ministry of Education, Anhui University, Hefei 230601, China, and also with the Key Laboratory of Southeast Coast Marine Information Intelligent Perception and Application, MNR, Zhangzhou 363000, China (e-mail: sycmiao@gmail.com; yingsong.li@ahu.edu.cn).

Jianghui Li is with the State Key Laboratory of Marine Environmental Science, College of Ocean and Earth Sciences, Xiamen University, Xiamen 361102, China (e-mail: jli@xmu.edu.cn).

Digital Object Identifier 10.1109/JOE.2024.3403255

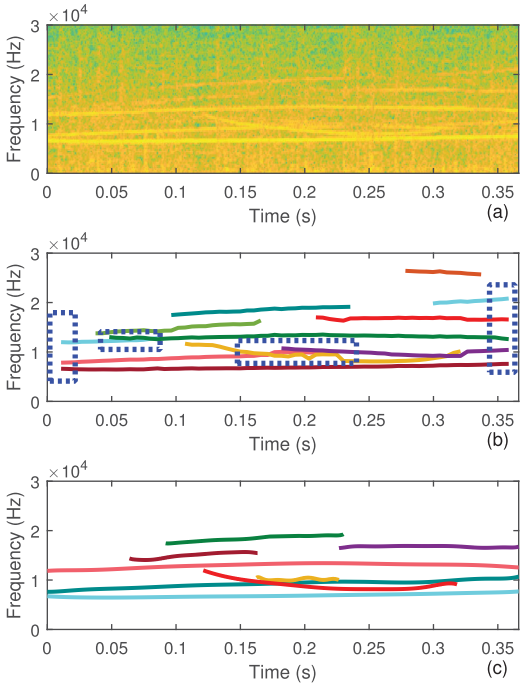

Fig. 1. (a) TFR of Fraser's dolphin recording. Detection of the whistle by using (b) GM-PHD, and (c) proposed method where each whistle is highlighted with a different color. Rectangles with blue dashed lines indicate error regions.

of the whistle ridges in the approximate boundary even though pulsed calls interrupt the continuous tracking. In this way, the whistle ridge is modeled as a 3-D feature, including not only frequency, but also the first and second-order derivatives of the ridge. This can be seen as a different perspective of applying this available information to whistle detection. These methods commonly allow for single-target tracking of IFs of whistles. Some multitarget tracking frameworks, such as Gaussian mixture probability hypothesis density (GM-PHD) filter [17], [18], and sequential Monte Carlo probability hypothesis density [19] allow for simultaneous tracking whistles from highly cluttered measurements in the presence of false alarms and missed detections. However, some failures happen when a part of the whistle becomes nearly indistinguishable in the TFR and overlapping components of whistles are close to each other. For the GM-PHD method, example ridges with errors in the blue dotted rectangle are shown in Fig. 1(b) and the results of detected ridges reveal endpoint effects of the method.

In our earlier study [5], it was demonstrated that an adaptive directional ridge prediction tracker (ADRPT) can guarantee accurate tracking of variations of whistle ridges, which become nearly indistinguishable in the low-resolution TFR. ADRPT is implemented by using prediction and tracking steps without prior knowledge. However, in the prediction step, the ADRPT can lose the whistle track when pulsed components are present in the TFR. Several click detectors [2], [20], [21] are developed as a first step to alleviating the influence of pulsed components. If pulsed components are removed directly from the TF plane with these detectors, the detected whistle ridges will appear discontinuous, which is attributed to the loss of TF points on these ridges. As an alternative approach, a structure-split-merge

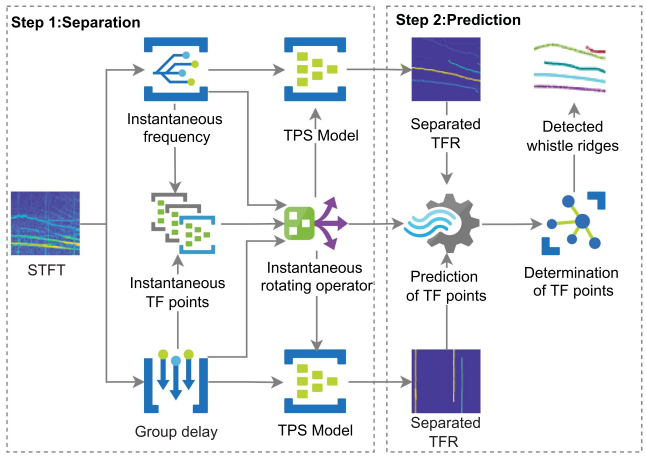

Fig. 2. Block diagram of the proposed tracker. Step 1: Separation of tonal and pulsed components; Step 2: ADRP.

(SSM) algorithm [20] together with the aid of TF methods can effectively separate whistles and clicks of marine mammal signals. However, the separation process of SSM relies on the direction information of TF structures obtained by an anisotropic Chirplet transform. Therefore, for whistle detection, this requires the development of methods that can quickly separate whistles and other types of calls, and handle overlapping whistles in low-resolution TFRs.

In this work, a novel tracker called adaptive directional ridge separation and prediction (ADRSP) is proposed to deal with whistles mixed with pulsed calls in low-resolution TFRs. A flowchart that illustrates the proposed method described in this work is shown in Fig. 2. The advantage of the proposed ADRSP is that it does not require any model training and can be used even when only a low-resolution TFR, namely, STFT, is used. The main contributions of this work are summarized as follows.

1) A new TF reassignment based on time-scale changes initially generates instantaneous TF points with energy concentrated on the ridge corresponding to the IF. Without prior knowledge, the properties of instantaneous TF points, namely, IF and group delay, are exploited to ensure continuity along the ridges of various types of components in a low-resolution TFR.
2) An instantaneous rotating operator (IRO) is defined to compute the direction of each TF point. Using the IRO, ridge points associated with the ridge preference angle are identified at locations where the instantaneous TF points exhibit rapid sign changes.
3) Using the direction of these TF points, a tone-pulse separation (TPS) model is developed for the separation of tonal and pulsed TF structures corresponding to various types of components. In the TPS model, two median filters based on a moving histogram are applied to preserve sharp ridges while removing impulse noise.
4) For instantaneous TF points, an improved adaptive directional ridge prediction (ADRP) method takes into account both the amplitude and direction of the TF points to extract whistle ridges even though they overlap. The

ADRP consists of the prediction and determination of TP points on a ridge.

5) The proposed ADRSP is tested on real UWA data sets to demonstrate its superior performance against existing methods for detecting whistles.

The rest of this article is organized as follows. Section II details the TF reassignment based on time-scale changes to enhance the instantaneous TF points. The two key processes of ADRSP, TPS, and ADRP, are described in Sections III and IV, respectively. Section V provides the summary and computational complexity of the algorithm. Section VI presents the results of performance evaluation experiments for the proposed ADRSP on a data set of UWA signals, which has been hand-annotated. Finally, Section VII concludes this article.

II. Instantaneous TF Point Mapping

Due to its superb efficiency, the STFT is widely used in the analysis of UWA signals, which are characterized by several TF components with nonlinear IF functions. The most significant issue with STFT is the uncertainty principle, which results in low energy concentration of TF points along these nonlinear IFs. TF reassignment allows for an improvement in the energy concentration, thereby enhancing TFRs. However, the goal of this article focuses on using TF reassignment to compute and enhance the instantaneous TF points, rather than TFRs.

A. TF Reassignment

Given an analysis window

$$h(t) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{t}{\sigma})^2}$$

the (modified) STFT of a signal $s(t) \in L^2(R)$ is defined as

$$S(t,f) = \int_{\mathbb{R}} s(\tau) h(\tau - t) e^{-i2\pi f(\tau - t)} d\tau. \tag{1}$$

Starting from the STFT $S(t,f)$, this enables the definition of the local IF $\hat{\omega}(t,f)$ and the local group delay (instantaneous time) $\hat{\tau}(t,f)$ by

$$\begin{aligned} \hat{\omega}(t,f) &= \frac{\partial_t \arg S(t,f)}{2\pi} = f - \Im\left\{\frac{S^{h'}(t,f)}{S(t,f)}\right\} \\ \hat{\tau}(t,f) &= t - \frac{\partial_f \arg S(t,f)}{2\pi} = t + \Re\left\{\frac{S^{\text{th}}(t,f)}{S(t,f)}\right\} \end{aligned} \tag{2}$$

where ∂_t and ∂_f represent first order partial derivatives with respect to t and f, respectively, th stands for the function $\text{th}(t)$, and the terms $S^{h'}(t,f)$ and $S^{\text{th}}(t,f)$ represent the STFTs with $h(t)$ replaced by $(dh(t))/(dt)$ and $\text{th}(t)$, respectively. In addition, $\Im\{\cdot\}$ and $\Re\{\cdot\}$ signify the imaginary and real parts of a complex number, respectively. According to the local energy given by the STFT, each TF pair (t,f) is mapped into a new TF pair $(\hat{\tau}(t,f), \hat{\omega}(t,f))$. As described in [20], TF reassignment enhances the energy concentration along the ridge. Mapping results for three simple signals with a single component are presented in Table I. For a whistle with a constant frequency f_0, the TF pair (t,f) is transformed into (t,f_0), concentrating the energy along the ridge at $f = f_0$. In the STFT analysis, this ridge exhibits a thickness of σ. Similarly, the ridge representing a simple click has a thickness of $1/\sigma$. By adjusting the parameter σ, tonal and pulsed components can be effectively separated in the mapping process. This phenomenon is later corroborated by the experimental results, as shown in Fig. 3.

TABLE I
Mapping Results of TF Reassignment for Simple Signals

	simple click $s(t) = \delta(t - t_0)$	simple whistle $s(t) = e^{i2\pi f_0 t}$	linear whistle $s(t) = e^{i\alpha\frac{t^2}{2}}$
$\hat{\omega}(t,f)$	f	f_0	$\alpha\hat{\tau}(t,f)$
$\hat{\tau}(t,f)$	t_0	t	$\frac{t+\alpha\sigma^4 f}{1+\alpha^2\sigma^4}$

The mapping process $(t,f) \mapsto (\hat{\tau}(t,f), \hat{\omega}(t,f))$, using three STFTs $S(t,f)$, $S^{th}(t,f)$, and $S^{h'}(t,f)$ with distinct windows $h(t)$, $(dh(t))/(dt)$ and $th(t)$, respectively, achieves high energy concentration of TF points. However, to determine the direction information of these points, as discussed in the subsequent Section III-A, similar six STFTs are calculated for second-order partial derivatives. In addition, the algorithmic implementation for nonlinear IF tracking incurs higher complexity due to the use of the second-order Taylor expansion of the phase in the prediction process. Given the STFT defined with a Gaussian window in formula (1), one way to reduce this complexity involves simplifying $\hat{\tau}(t,f)$ and $\hat{\omega}(t,f)$. Therefore, TF reassignment based on time-scale changes is developed to achieve these simplified mapping TF points.

B. TF Reassignment Based on Time-Scale Changes

According to the derivation result

$$\begin{aligned} S^{h'}(t,f) &= \frac{1}{\sigma^2}\int_{\mathbb{R}} s(\tau)(\tau - t)h(\tau - t)e^{-i2\pi f(\tau - t)}d\tau \\ &= \frac{1}{\sigma^2} S^{\text{th}}(t,f) \end{aligned} \tag{3}$$

an auxiliary STFT with a time-shifted window is defined as

$$\hat{S}(t,f) = \frac{1}{\sigma}\int_{\mathbb{R}} s(\tau)(\tau - t)h(\tau - t)e^{-i2\pi f(\tau - t)}d\tau. \tag{4}$$

The mapping in (2) can be simplified to

$$\begin{aligned} \hat{\omega}(t,f) &= f - \frac{1}{\sigma}\Im\left\{\frac{\hat{S}(t,f)}{S(t,f)}\right\} \\ \hat{\tau}(t,f) &= t + \sigma\Re\left\{\frac{\hat{S}(t,f)}{S(t,f)}\right\}. \end{aligned} \tag{5}$$

Computing $\hat{\omega}(t,f)$ and $\hat{\tau}(t,f)$ requires only twice the numerical effort of a single STFT computation. In addition, the formula presented in (5) demonstrates that controlling the parameter σ enables the separation of tonal and pulsed components. A segment of a killer whale signal lasting approximately 0.73 s and sampled at 48 kHz is selected for analysis. Fig. 3 shows the results of applying this TF reassignment procedure to this signal

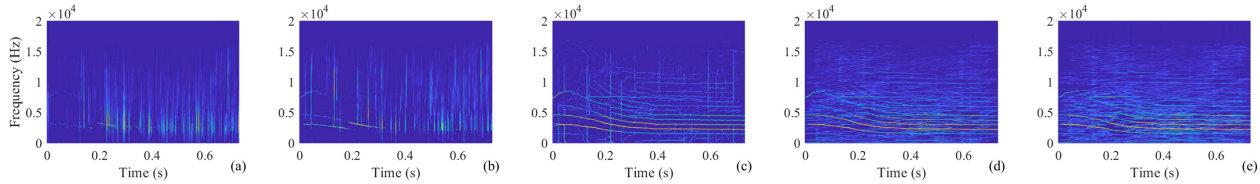

Fig. 3. Results of TF reassignment based on time-scale changes using the different values of the parameter σ. (a) $\sigma = 0.2$. (b) $\sigma = 0.1$. (c) $\sigma = 0.05$. (d) $\sigma = 0.02$. (e) $\sigma = 0.005$.

containing whistles and clicks. It is observed that a smaller value of σ results in a higher energy concentration of TF points on the whistle ridges, whereas it simultaneously decreases the energy concentration on click ridges. Specifically, when the values of the parameter σ are set to 0.02 [see Fig. 3(d)] and 0.005 [see Fig. 3(e)], the fuzzy pulsed component can be considered as noise. This phenomenon can support robust separation of tonal and pulsed components in the TF domain.

Although the expressions $\hat{\omega}(t,f)$ and $\hat{\tau}(t,f)$ for first-order derivatives are obtained and used effectively, the higher order derivatives can provide more accurate detection of whistles ridges with nonlinear IFs. The second-order Taylor expansion of the phase is used to derive the second-order expression for $\hat{\omega}(t,f) = f(t) + f'(t)(t - \hat{\tau}(t,f))$ where $f(t)$ is IF and $f'(t)$ represents the chirp rate, which can be estimated using the IRO $\vartheta(t,f)$ introduced in [5]. This operator aids in determining the direction of the mapped TF points; however, its limitation lies in overlooking the influence of pulsed components. A solution to this issue is discussed in the forthcoming Section III.

III. Separation of Tonal and Pulsed Components

After obtaining the mapped TF points from the TFRs, the goal of this section is to split these points into two distinct TFRs $S_\tau(t,f)$ for tonal components and $S_p(t,f)$ for pulsed components, respectively. To achieve this goal, the SSM algorithm introduced in [20] can effectively separate tonal and pulsed TF structures from the TFR based on a series of predetermined orientation angles. In this work, using the IRO $\vartheta(t,f)$, a novel TPS model is introduced to achieve a separated TFR that more clearly delineates tonal components.

A. Instantaneous Rotating Operator

The IRO $\vartheta(t,f)$ can determine the direction of the TF points along the ridges. Generally, $\vartheta(t,f)$ belongs to a class of mathematical structures that involve second-order derivatives. In mathematics, the Hessian matrix is often used to calculate the second-order derivative of a scalar-valued function. Indeed, the magnitude $|S(t,f)|$ of an STFT represents such a scalar-valued function, whose gradient can be defined as

$$\begin{aligned}\nabla G(t,f) &= [\kappa_t(t,f), \kappa_f(t,f)] \\ &= \left[\sigma \Re\left\{\frac{\hat{S}(t,f)}{S(t,f)}\right\}, -\frac{1}{\sigma}\Im\left\{\frac{\hat{S}(t,f)}{S(t,f)}\right\}\right] \end{aligned} \tag{6}$$

where $\kappa_t(t,f) = \hat{\tau}(t,f) - t$ and $\kappa_f(t,f) = \hat{\omega}(t,f) - f$. To simplify the calculations of second-order partial derivatives, another auxiliary STFT with an analysis window, $t^2 h(t)$, is defined as

$$\hat{\hat{S}}(t,f) = \frac{1}{\sigma^2}\int_{\mathbb{R}} s(\tau)(\tau - t)^2 h(\tau - t) e^{-i2\pi f(\tau - t)} d\tau. \tag{7}$$

Let $\alpha = (\hat{\hat{S}}(t,f))/(S(t,f))$ and $\beta = (\hat{S}(t,f))/(S(t,f))$, the Hessian matrix of the function $|S(t,f)|$ is given by

$$\begin{aligned}\boldsymbol{H} &= \begin{bmatrix} \kappa_{tt}(t,f) & \kappa_{tf}(t,f) \\ \kappa_{ft}(t,f) & \kappa_{ff}(t,f) \end{bmatrix} \\ &= \begin{bmatrix} \Re\left\{-1 + \alpha - \beta^2\right\} & 2\pi\sigma^2 \Im\left\{\alpha - \beta^2\right\} \\ \Im\left\{\frac{1}{\sigma^2}(1 - \alpha + \beta^2)\right\} & 2\pi \Re\left\{-\alpha + \beta^2\right\} \end{bmatrix}. \end{aligned} \tag{8}$$

Finding the eigenvalue and eigenvector of the Hessian matrix is performed stably and efficiently by using the Jacobi algorithm [22]. In the TF domain, the rotation matrix associated with each TF point is defined by specific orientation parameters

$$\boldsymbol{R} = \begin{bmatrix} \cos\vartheta(t,f) & -\sin\vartheta(t,f) \\ \sin\vartheta(t,f) & \cos\vartheta(t,f) \end{bmatrix} = \begin{bmatrix} c & -s \\ s & c \end{bmatrix} \tag{9}$$

where $c = \cos\vartheta(t,f)$ and $s = \sin\vartheta(t,f)$. The direction information for each TF point in the TFR is obtained by calculating the eigenvalues $\lambda_1(t,f)$, $\lambda_2(t,f)$ with $\lambda_1(t,f) \geq \lambda_2(t,f)$. The Hessian matrix can be diagonalized into a form

$$\begin{bmatrix} c & -s \\ s & c \end{bmatrix}^T \boldsymbol{H} \begin{bmatrix} c & -s \\ s & c \end{bmatrix} = \begin{bmatrix} \lambda_1(t,f) & 0 \\ 0 & \lambda_2(t,f) \end{bmatrix}$$
$$\boldsymbol{R}^T \boldsymbol{H} \boldsymbol{R} = \boldsymbol{\Sigma}. \tag{10}$$

When the off-diagonal elements of the Hessian matrix are reduced to zero, the values of the diagonal elements are correspondingly increased, namely

$$(c^2 - s^2)(\kappa_{ft}(t,f) + \kappa_{tf}(t,f)) + 2cs(\kappa_{ff}(t,f) - \kappa_{tt}(t,f)) = 0. \tag{11}$$

From (11), let

$$\mu = \frac{\kappa_{tt}(t,f) - \kappa_{ff}(t,f)}{\kappa_{ft}(t,f) + \kappa_{tf}(t,f)} = \frac{c^2 - s^2}{2cs} = \frac{1 - \tan^2(\vartheta(t,f))}{2\tan(\vartheta(t,f))} \tag{12}$$

the instantaneous chirp rate $\tan(\vartheta(t,f))$ is determined to be

$$\tan(\vartheta(t,f)) = \begin{cases} -\mu + \sqrt{\mu^2 + 1}, & \mu \geq 0 \\ -\mu - \sqrt{\mu^2 + 1}, & \mu < 0. \end{cases} \tag{13}$$

Thus, the IRO at each TF point can be obtained using the $\arctan(\cdot)$ function. Subsequently, this operator will be applied to separate tonal and pulsed components.

B. TPS Model

Using the IRO, ridge points are identified at locations where the instantaneous TF points exhibit rapid sign changes. These ridge points, associated with the ridge preference angle $\vartheta(t,f)$ are defined by the zeros of the inner product, $\Im\{\beta e^{i\vartheta(t,f)}\}=0$, indicating opposite signs on each side of the ridge. However, the presence of noise makes the detection of these ridge points in a discrete setting rather unstable. Direct filtering of the instantaneous TFR could erroneously classify the ridges of weak whistles as noise, resulting in their elimination. To address this challenge, a sign TFR with symbol values of -1 and $+1$ is defined as

$$\vec{S}(t,f)=\begin{cases}1, & \Im\left\{\beta e^{i\vartheta(t,f)}\right\}>0\\ -1, & \Im\left\{\beta e^{i\vartheta(t,f)}\right\}<0\end{cases} \tag{14}$$

to better delineate these ridge points despite noise interference.

To preserve sharp ridges while removing impulse noise, median filters based on a moving histogram are applied to $\vec{S}(t,f)$ in both horizontal and vertical directions. Two enhanced TFRs using median filters are given by

$$\begin{aligned}\bar{S}_\tau(t,f)&:=\text{median}\left(\vec{S}(t-l_\tau,f),\ldots,\vec{S}(t+l_\tau,f)\right)\\ \bar{S}_p(t,f)&:=\text{median}\left(\vec{S}(t,f-l_p),\ldots,\vec{S}(t,f+l_p)\right)\end{aligned} \tag{15}$$

for l_τ, $l_p\in\mathbb{N}$ where $2l_\tau+1$ and $2l_p+1$ are the lengths of the median filters, respectively. The tonal components become more apparent in the enhanced TFR $\bar{S}_\tau(t,f)$, whereas the pulsed components vanish. Conversely, in the enhanced TFR $\bar{S}_p(t,f)$, pulsed components experience similar enhancement effects.

Considering the influence of noise, two enhanced TFRs are not directly used for ridge detection. Although a binary mask can separate different types of components while reducing noise for the TFRs [23], [24], it treats all components equally, potentially removing important information in the intersection regions. Instead, two soft masks based on Wiener filtering [24] are obtained from these enhanced TFRs by

$$\begin{aligned}T_\tau(t,f)&:=\left(\frac{\bar{S}_\tau(t,f)}{\bar{S}_\tau(t,f)+\bar{S}_p(t,f)+\varepsilon}\right)\\ T_p(t,f)&:=\left(\frac{\bar{S}_p(t,f)}{\bar{S}_\tau(t,f)+\bar{S}_p(t,f)+\varepsilon}\right)\end{aligned} \tag{16}$$

where a small positive value ε is added to avoid division by zero. With soft masks, the assignment process for all points is not strict but proportionate, as expressed by the masking weights to prevent breakpoints caused by the mutual influence of components in both directions. By applying two masks to the original TFR $S(t,f)$, the separated TFRs for tonal and pulsed components are obtained

$$\begin{aligned}S_\tau(t,f)&:=S(t,f)T_\tau(t,f)\\ S_p(t,f)&:=S(t,f)T_p(t,f).\end{aligned} \tag{17}$$

Fig. 4 shows the results of applying the TPS model to the real signal mentioned in Fig. 3. The TPS uses directional information to categorize TF points as belonging to either tonal, pulsed, or noise components. The TFR of the signal as shown in Fig. 4(a) is generated using the STFT with a hamming window of length 512. In Fig. 4(b), colored arrows indicate whether the respective TF point is assigned to the tonal (red), the pulsed (blue), or the noise (green) component based on the direction information. The separated TFRs for the tonal components are illustrated in Fig. 4(c), and for the pulsed components in Fig. 4(d). In Fig. 4(c), the tonal components in the separated TFR $S_\tau(t,f)$ [see Fig. 4(b)] become more obvious while pulsed components are eliminated.

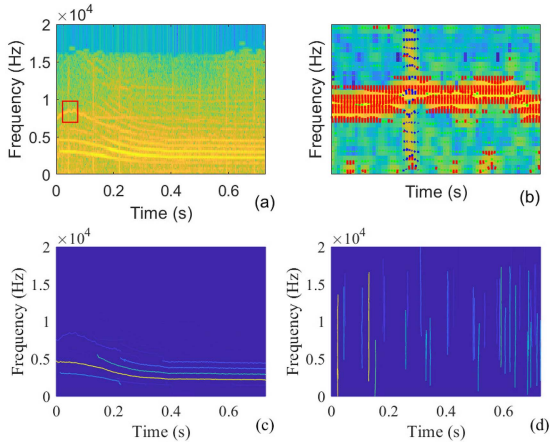

Fig. 4. TFS results for the separation of TF components. (a) Original TFR $S(t,f)$ of a killer whale signal. (b) Direction of zoomed TF structures within the red rectangles. (c) Separated TFR $S_\tau(t,f)$ for tonal components. (d) Separated TFR $S_p(t,f)$ for pulsed components.

IV. Adaptive Directional Ridge Prediction

Based on the TF amplitude and the direction of each TF point, as well as the separated TFR $S_\tau(t,f)$, whistle ridges are estimated using a novel ADRP method. ADRP involves two principal operations: the prediction and the determination of TF points on ridges.

A. Prediction of TF Points on Ridges

The goal of the proposed ADRP is to extract ridges of overlapping components without prior knowledge of the number of components, or the starting and ending points of the whistles. To extract ridges with nonlinear IFs, the second-order expression $\hat{\omega}(t,f)=f(t)+f'(t)(t-\hat{\tau}(t,f))$ where $f(t)$ is IF and $f'(t)$ denotes the chirp rate is derived from the general signal model and the STFT, as introduced in [5]. Diverging from the ridge prediction model presented in [5], the improved prediction of TF points is formulated as follows:

$$\begin{aligned}\tilde{\omega}(t)&=\hat{\omega}(t,f)-\tan(\vartheta(t,f))(t-\hat{\tau}(t,f))\\ &=\hat{\omega}(t,f)+\sigma\tan(\vartheta(t,f))\Re\left\{\frac{\hat{S}(t,f)}{S(t,f)}\right\}\end{aligned} \tag{18}$$

according to IRO $\vartheta(t,f)$ corresponding to a stable orientation. Using the instantaneous TFR, the model can predict ridge points even in the presence of pulsed components, compared to ridge prediction in [5]. The key steps of the prediction of TF points on ridges are elaborated as follows.

1) For the mth component, an initialized point (t_0, f_0) of the highest energy can be modeled as

$$(t_0, f_0) := \arg\max_{t,f} \{|S_\tau(t, f)|\} \tag{19}$$

and its direction $\vartheta_m(t_0, f_0)$ is calculated using the inverse tangent function of the value given in (13).

2) For $\tau = \{t_0 + 1, \ldots, N_t\}$, the next TF points of the mth ridge is predicted to be

$$\tilde{\omega}_m(\tau) = \hat{\omega}_m(\tau, f) + \sigma \tan(\vartheta_m(\tau, f))\Re\left\{\frac{\hat{S}(\tau, f)}{S(\tau, f)}\right\}. \tag{20}$$

Although the prediction process can ensure an unbroken detection of whistles, the predicted point may not always lie on a ridge. For instance, if the mth ridge, with a length of N_m, is shorter than the signal length N_t, a predicted point beyond the two endpoints may not exist on this ridge. Therefore, it is essential to determine whether the predicted point is part of this ridge.

B. Determination of TF Points on Ridges

One adaptive method involves penalizing the frequency difference between the predicted points and the previous points. To circumvent the influence of the pulsed and noise components, instead of tracking TF points in the same TFR as described in [5], another separated TFR $S_\tau(t, f)$, which contains only tonal components, is utilized for the accurate determination of TF points on ridges.

The mth ridge can be parameterized as $\bar{\omega}_m(t) = \varphi_{p_m(t)}(t)$ where $p_m(t)$ represents the frequency index of the determined TF point at each time t, and $\varphi_m(t)$ denotes the frequency value corresponding to this index, which is calculated by

$$\varphi_m(t) : \begin{cases} \left[\partial_f |S_\tau(t, f)|\right]_{f=\varphi_m(t)} & = 0 \\ \left[\partial_f^2 |S_\tau(t, f)|\right]_{f=\varphi_m(t)} & < 0 \end{cases} \tag{21}$$

where ∂_t^2 represents the second order partial derivative with respect to t. According to the prediction model derived in (20), assuming that $\bar{\omega}_m(\tau - 1)$ represents the frequency of the determined previous point, the frequency index of the next point may be determined by $p_m(\tau) = p_m(\tau - 1) + \kappa_m(t)$, where the integer $\kappa_m(t)$ stands for a time-varying slope. Depending on the chirp rate of the previous point and the frequency resolution Δf of the TFR obtained by the STFT, this slope is calculated as

$$\begin{aligned} \kappa_m(t) = \lfloor (\Delta f)^{-1} \tan(\vartheta_m(\tau - 1, \bar{\omega}_m(\tau - 1))) \\ \times (\tau - 1 - \hat{\tau}(\tau - 1, \bar{\omega}_m(\tau - 1))) \rfloor \end{aligned} \tag{22}$$

where $\lfloor \cdot \rfloor$ denotes rounding down to the nearest integer. Given that a constant frequency bandwidth $[-\Delta f, \Delta f]$ is not ideally suited for components with both fast and slow varying IFs, the upper and lower boundaries of a new frequency-varying bandwidth, $\Delta\omega$, are defined within the range $[\Delta\omega^-, \Delta\omega^+]$. Using the frequency index of the previous point, these boundaries are calculated as follows:

$$\begin{cases} \Delta\omega^- = \max(1, p_m(\tau - 1) + \kappa_m(t) - \Delta f) \\ \Delta\omega^+ = \min(N_f, p_m(\tau - 1) + \kappa_m(t) + \Delta f) \end{cases} \tag{23}$$

where N_f represents the size of the frequency dimension in the TFR.

To suppress absolute frequency jumps, as previously done, an adaptive penalty function can then be constructed by the relative deviations of the ridge frequency of components and its derivative from these typical values

$$\begin{aligned} p_m(\tau) = \arg\min_f \{|S_\tau(\tau, \Delta\omega)| - \tilde{\omega}_m(\tau)\} \\ + \Delta\omega^- - 1, |S_\tau(\tau, \Delta\omega)| \geq \delta_S \\ \bar{\omega}_m(\tau) = \begin{cases} \tilde{\omega}_m(\tau), p_m(\tau) \in [\frac{\Delta\omega^-}{2}, \frac{\Delta\omega^+}{2}] \\ \varphi_{p_m(\tau)}(\tau), \text{otherwise} \end{cases} \end{aligned} \tag{24}$$

where the threshold $\delta_S = \mu \max\{|S_\tau(t, f)|\}$ is determined by the parameter $\mu = \min\{E(f)\}/\max\{E(f)\}$, with the concentration measure $E(f)$ introduced in [10]. The concentration measure $E(f) = \int_{\mathbb{R}} |S_\tau(t, f)| dt$ is used for every frequency considered in the separated TFR.

Immediately following the second step of the prediction operation, the ADRP carries out this penalty function. For $\tau = \{t_0 - 1, \ldots, 1\}$, a similar procedure is applied for the prediction and determination of TF points on the mth ridge. Due to potential interference from noise, the extracted ridge might correspond to a noise component. To eliminate these ridges, the length of the ridge should not be less than $N_t^2/(\alpha f_s)$ where the parameter $\alpha = 32$ introduced in [10].

Using the extracted ridge, denoted as $\bar{\omega}_m(t)$, the corresponding component is removed from the mixed signal utilizing the TF filtering method [3], [4]. This process is repeated until no additional ridges are extracted in the TF plane. To achieve separation of overlapping sources, techniques, such as correlation analysis or graph clustering [16], are then applied to isolate single source ridges from these extracted ridges.

V. Summary and Computational Complexity of the Algorithm

In this section, we present the summary and computational complexity of the proposed ADRSP algorithm. The key steps of the overall algorithm are summarized in Algorithm 1.

The proposed method involves the computation of the STFT, TPS, and ADRP. For a signal of length N, the computational cost of applying the STFT with N_h windows is $O(3N \log N_h)$, where the factor of 3 represents the use of three different windows to compute IRO and the instantaneous TF points. The resulting TFR has dimensions $N_t \times N_f$. In the TPS model, the main complexity of applying two median filters based on moving histograms to the TFR is $O(2N_tN_f(l_\tau + l_p + 1))$, which simplifies to $O(2N_tN_f)$ when l_τ and l_p are significantly smaller than N_t and N_f, respectively. For the ADRP, where prediction and determination operations are executed

Algorithm 1: Proposed ADRSP Algorithm.

Input: A signal $s(t)$, the sampling frequency f_s
Output: All extracted ridges and the number of ridges m
Initialize $m = 0$ and compute three different STFTs $S(t, f)$, $\hat{S}(t, f)$, and $\hat{\hat{S}}(t, f)$;
The Hessian matrix of the function $|S(t, f)|$ is given in Eq.(8);
The instantaneous chirp rate $\tan(\vartheta(t, f))$ is determined by Eq.(13);
A sign TFR with symbol values of -1 and +1 is computed using Eq.(14);
The separated TFRs, $S_\tau(t, f)$ and $S_p(t, f)$, are obtained using median filters;
Calculate the concentration measure $E(f)$ for $S_\tau(t, f)$ and the threshold δ_S;
while True **do**
  Find the highest energy point (t_0, f_0);
  Initialize the number of points, $i = 0$, for the m^{th} ridge;
  for $\tau \leftarrow t_0 + 1$ **to** N_t **do**
    The next TF point, $\tilde{\omega}_m(\tau)$, of the m^{th} ridge is predicted using Eq.(20);
    Compute a new frequency-varying bandwidth, $\Delta\omega$;
    if $|S_\tau(\tau, \Delta\omega)| \geq \delta_S$ **then**
      The point $\bar{\omega}_m(\tau)$ of the m^{th} ridge is determined in the separated TFR $S_\tau(t, f)$ using an adaptive penalty function, as described in Eq.(24);
      $i = i + 1$;
    else
      break;
    end
  end
  for $\tau \leftarrow t_0 - 1$ **to** 1 **do**
    A similar procedure is applied to predict and determine TF points on the m^{th} ridge;
  end
  if $i \geq \frac{N_t^2}{\alpha f_s}$ **then**
    $m = m + 1$;
    Using the extracted ridge, $\bar{\omega}_m(t)$, the corresponding component is removed from $S_\tau(t, f)$ by utilizing the TF filtering method;
  end
  Compute the mean of the filtered TFR $S_\tau(t, f)$, denoted as $S_{\tau,\text{mean}}$;
  if $S_{\tau,\text{mean}} < \delta_S$ **then**
    break;
  end
end

in parallel, the computational cost for tracking ridges is $O(MN_mN_p) \approx O(M \log N_t \log N_f)$, where M is the number of components, $N_m \in [1, N_t]$ is the length of the mth ridge, $N_p \in [1, N_f]$ is the number of frequency bins. Therefore, the total computational cost of the method can be summarized as $O(3N \log N_h) + O(2N_tN_f) + O(M \log N_t \log N_f)$, indicating that the proposed method is computationally feasible for real-world applications.

To further validate the practical applicability of the proposed method, the subsequent Experiment 4 in Section VI is dedicated to comparing its execution time and resource consumption with established methods under equivalent conditions. These empirical evaluations will complement the theoretical complexity analysis and provide insights into the method's performance and feasibility in real-world scenarios.

VI. Experimental Results and Performance Analysis

In this section, the performance of the proposed ADRSP for whistle detection is evaluated by the application of both

synthetic and real UWA signals. To quantify the detection performance, point-based evaluation metrics are used by comparing manually annotated reference information with ridges detected by the algorithms. The references were sourced from the MobySound archive. Three different methods: graph-based clustering (GC) [16], GM-PHD [17], and SMC-PHD [19] are extended for the comparative analysis of performance and complexity.

A. Evaluation Metrics

In this work, detection accuracy is quantified using four metrics: *precision* (P), *recall* (R), F-score, and *error rate* (ER). Mathematically, precision (P) and recall (R) are calculated as follows:

$$P = \frac{\sum_{k=1}^{K} N_{\text{TP}}(k)}{\sum_{k=1}^{K} N_{\text{TP}}(k) + \sum_{k=1}^{K} N_{\text{FP}}(k)} \times 100\% \tag{25}$$

$$R = \frac{\sum_{k=1}^{K} N_{\text{TP}}(k)}{\sum_{k=1}^{K} N_{\text{TP}}(k) + \sum_{k=1}^{K} N_{\text{FN}}(k)} \times 100\% \tag{26}$$

where the number of true positives $N_{\text{TP}}(k)$ is the total number of TF points that are active in both reference and detection for the kth frame, the number of false positives $N_{\text{FP}}(k)$ is the total number of TF points that are active in the detection but are not present in the reference for the same frame, the number of false negatives $N_{\text{FN}}(k)$ is the total number of TF points that are not present in the detection but are active in the reference. The F-score metric, which combines the harmonic mean of precision (P) and recall (R), is formulated as $(2PR)/(P+R)$.

For ER, the number of substitutions $S(k)$ is obtained by merging $N_{\text{FN}}(k)$ and $N_{\text{FP}}(k)$ without correlating which false positive substitutes for which false negative, defined as $S(k) = \min(N_{\text{FN}}(k), N_{\text{FP}}(k))$. The remaining numbers of insertions $I(k)$ and deletions $D(k)$, if any, are counted as $I(k) = \max(0, N_{\text{FP}}(k) - N_{\text{FN}}(k))$ and $D(k) = \max(0, N_{\text{FN}}(k) - N_{\text{FP}}(k))$. Then, the total ER is calculated as

$$ER = \frac{\sum_{k=1}^{K} S(k) + \sum_{k=1}^{K} I(k) + \sum_{k=1}^{K} D(k)}{\sum_{k=1}^{K} N(k)} \tag{27}$$

where $N(k)$ is the total number of TF points in the reference.

B. Experiments

In simulations, unless otherwise noted, the STFT of a signal is computed by using a Kaiser window with a length of 512. In the TPS model, to balance noise reduction and computational cost, the parameters l_τ and l_p are set to 2. This setting ensures that the lengths $2l_\tau + 1$ and $2l_p + 1$ of the median filters are odd numbers, namely, 5. In addition, a small value of ε is set to 1e-12, considering the algorithm runs on MATLAB 2021b on a 64-bit system.

Experiment 1: To investigate the sensitivity of the proposed ADRSP to the values of the input parameter σ, the best-performing version is evaluated on a mixed signal comprising three types of components. Given the formulas for computing $\hat{\omega}(t,f)$ and $\hat{\tau}(t,f)$ [refer to (5)], it is evident that σ should

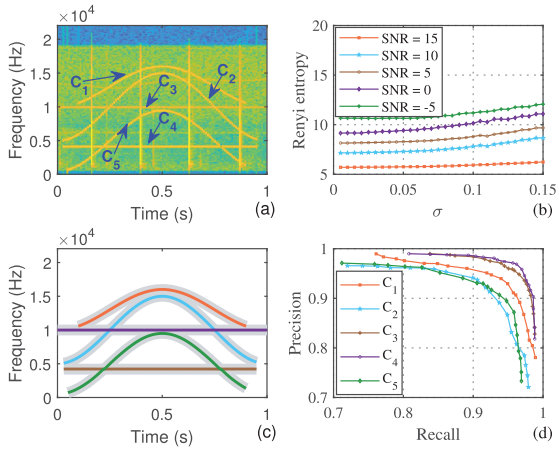

Fig. 5. (a) TFR of the signal. (b) Evolution of Rényi entropies with respect to σ in different SNR cases. (c) Whistle detection from ADRSP. Each detected whistle ridge is highlighted with a different color and gray indicates the true whistle ridge referenced. (d) Precision and recall for whistle detection.

be carefully selected to achieve a balance between localization and separation of different types of components. This balance requires σ to be sufficiently large for pulsed components, and small for tonal components. Nevertheless, by examining the time-varying term in (20), it can be demonstrated that this term depends on $\sigma \tan(\vartheta_m(t,f))$, indicating that σ should be small enough to ensure accurate reconstruction step for whistle ridges.

To deal with this issue, an improved R é nyi entropy metric is defined as

$$R = \frac{1}{1-\rho} \log_2 \left(\frac{\iint_{\mathbb{R}^2} |S(t,f)\delta(\omega - \hat{\omega}(t,f))|^\rho \, df\, dt}{\iint_{\mathbb{R}^2} |S(t,f)\delta(\omega - \hat{\omega}(t,f))| \, df\, dt} \right) \tag{28}$$

with integer orders $\rho = 3$ being recommended in [25]. The larger the Rényi entropy, the less energy is concentrated along the ridges of tonal components in the TFR.

A signal composed of five tonal components ($\text{C}_1, \text{C}_2, \text{C}_3, \text{C}_4$, and C_5) is corrupted with noise at different signal-to-noise ratios (SNRs) and strong pulsed components from the clicks of a real whale signal. The signal length is 44 101, with a sampling rate of 44.1 kHz. The TFR of the signal with an SNR of 0 dB is obtained by using STFT and is illustrated in Fig. 5(a). In Fig. 5(b), the evolution of Rényi entropy versus the variable σ from 0.005 to 0.15 in steps of 0.005 at different noise levels (SNR $= -5$, 0, 5, 10, and 15 dB). It can be observed that for smaller σ, one acquires the higher energy concentration of tonal components, and conversely, for larger σ, one acquires the lower energy concentration. This phenomenon is independent of the noise level. When $\sigma < 0.05$, the energy concentration of TF points is close to the optimum, such that the sensitivity to σ is very low.

The parameter σ is set to 0.02. Fig. 5(c) shows the results of whistle detection from ADRSP compared to the true whistles. ADRSP successfully detects all five whistles, even those with overlapping components, and remains unaffected by the strong pulsed components due to the effective separation of tonal and pulsed components using the TPS model.

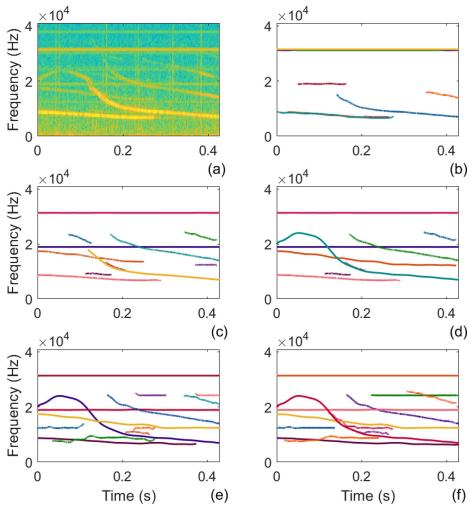

Fig. 6. Results of whistle detection with different values of the parameter σ. (a) TFR obtained by STFT. (b) $\sigma = 0.2$. (c) $\sigma = 0.1$. (d) $\sigma = 0.05$. (e) $\sigma = 0.02$. (f) $\sigma = 0.005$.

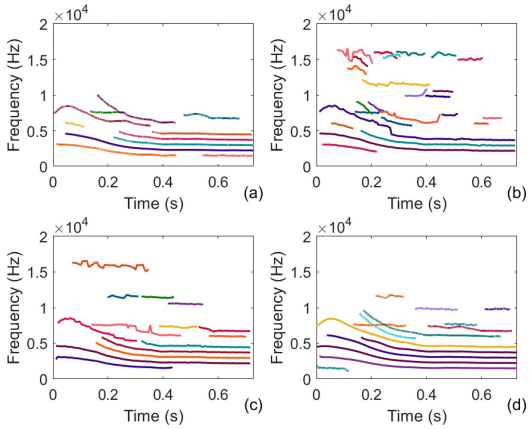

Fig. 7. Results of whistle detection of a killer whale signal whose TFR is presented in Fig. 3(a). (a) GC. (b) GM-PHD. (c) SMC-PHD. (d) ADRSP.

To evaluate the performance of each parameter set with σ ranging from 0.003 to 0.06 in steps of 0.003, the precision and recall of each whistle are computed. For values in the range of 0.018 to 0.045, this region of the precision–recall curve yields both high recall and high precision. Selecting values that are too low causes noise components to be extracted from the TFR, leading to decreased performance. This phenomenon is demonstrated in Fig. 6, where one displays the results of ADRSP for a TFR of a common dolphin signal, using five different values, 0.2, 0.1, 0.05, 0.02, and 0.005, of the input parameter σ. Comparing results in the value of $\sigma = 0.02$ and $\sigma = 0.005$ [see Fig. 6(e) and (f)], it is observed that ADRSP, with the smaller σ, detects more whistle ridges with some interference. If the value of σ is large, the energy of the mapped TF points spreads across a ridge with width σ resulting in the algorithm detecting fewer TF points on a ridge [see Fig. 6(c)], or two TF points in the same position on a ridge [see Fig. 6(b)]. Therefore, unless otherwise specified, the parameter σ is set to 0.02 in all subsequent experiments to achieve optimal performance.

Experiment 2: To illustrate the performance of the proposed method for the signal in strong background noise, the recorded signal shown in Fig. 3(a) is selected. Fig. 7 shows the results when the methods of GC, GM-PHD, SMC-PHD, and ADRSP are applied to the TFR of the signal.

Fig. 7(a) shows that GC can detect only part of the whistle ridges and missed part of the whistle with close components. However, it can extract nonlinear TF structures between 0 and 0.2 s. The detection of a tonal component around 0.5 s is difficult to discern due to a very weak whistle in the TFR. Although GC can extract whistle ridges with nonlinear TF structures, it struggles with weak whistles, which may then be misclassified as noise or outliers, leading to misleading clusters. In Fig. 7(b), detection errors are observed between 0.2 and 0.3 s. The GM-PHD method fails to detect whistle ridges in low-resolution TFRs, particularly when these components overlap. SMC-PHD, as illustrated in Fig. 7(c), can detect more ridges than both GC and GM-PHD. Despite improved performance in detecting strong overlapping components, this method has a limitation: it does not consistently succeed with weak components that are obscured by underwater background noise. Since the SMC-PHD detector cannot continue to track a target if the measurements are absent for several continuous time steps, the whistles are mainly detected as a single ridge but are occasionally "broken" into multiple fragments. Fig. 7(d) shows that the proposed ADRSP correctly detects most of the whistle ridges present and uses the TPS model to eliminate clicks (pulsed components). ADRSP can still predict TF points according to the direction and amplitude of the whistle ridges, such that continuous TF points are obtained during whistle detection, even in cases of component overlap. The experimental results indicate that ADRSP outperforms GC, GM-PHD, and SMC-PHD.

Experiment 3: To evaluate the overall performance of the proposed method, a data set of whistles was downloaded from the MobySound Archive. This data set includes audio recordings of five whale species; however, only recordings of melon-headed whales (denoted by s_2), long-beaked common dolphins (s_3), bottlenose dolphins (s_4), and spinner dolphins (s_5) are used in this study. Recordings of short-beaked common dolphins were excluded due to annotation errors in some files. Whistle ridges were annotated by trained analysts, as described in [26]. A total of 32 signals (16 recordings from the four selected species and synthesized signals denoted by s_1 at different SNRs $\{-5, 0, 5, 10, 15\}$ dB) are selected to compare metrics for detection performance.

Table II presents the evaluation metric scores for the GC, GM-PHD, SMC-PHD, and the proposed ADRSP methods applied to the MobySound data set. The GM-PHD method achieves higher overall precision P compared to overall recall R. In contrast, the SMC-PHD method demonstrates lower precision but higher recall in this comparison. For the GC and ADRSP methods, the results indicate a tradeoff between precision and recall; however, the precision and recall of ADRSP are significantly improved

TABLE II
PERFORMANCE COMPARISON OF WHISTLE DETECTION METHODS USING FOUR ESTIMATION METRICS ON THE MOBYSOUND DATA SET

Method	Signal	P	R	F	ER
GC	s_1	83.1	81.2	82.1	0.162
	s_2	82.4	81.1	81.7	0.161
	s_3	79.2	75.8	77.5	0.202
	s_4	78.6	78.3	78.5	0.179
	s_5	77.2	76.5	76.9	0.191
GM-PHD	s_1	87.7	83.9	85.8	0.144
	s_2	86.3	83.2	84.7	0.149
	s_3	84.7	81.8	83.3	0.158
	s_4	85.2	82.0	83.6	0.157
	s_5	85.5	82.5	83.9	0.154
SMC-PHD	s_1	85.8	90.4	88.0	0.131
	s_2	84.9	89.5	87.1	0.137
	s_3	83.5	86.8	85.1	0.147
	s_4	84.2	87.4	85.7	0.141
	s_5	83.3	88.4	85.8	0.150
ADRSP	s_1	93.3	91.7	92.5	0.077
	s_2	91.3	90.9	91.1	0.084
	s_3	89.3	88.5	88.9	0.104
	s_4	89.8	90.3	90.1	0.093
	s_5	89.6	89.1	89.3	0.099

compared to GC. For the GC, GM-PHD, and SMC-PHD methods, the F-scores are lower for species with larger group sizes, such as the signals denoted by s_3, s_4, and s_5, due to the closer and more overlapping components in the TFRs of these signals. Rather than directly extracting parameters, such as IF, from the TFR, the ADRSP method enhances the process by separating and predicting overlapping components, thereby preserving the flexibility required for accurate whistle detection, even in cases where components in the TFR are overlapped. High F-scores are still achieved by ADRSP for signals from these species. The average ER of the GC, GM-PHD, SMC-PHD, and ADRSP methods across five categories of signals (s_1, s_2, s_3, s_4, and s_5,) are 0.179, 0152, 0.141, and 0.091, respectively. ADRSP exhibits a lower average ER compared to the other methods. Therefore, Table II demonstrates the superior performance of ADRSP in accurately detecting whistles in real UWA signals.

Experiment 4: To illustrate the computational cost of the proposed method in detecting whistles, the comparison between the GC, GM-PHD, SMC-PHD, and ADRSP methods is presented using 32 signals from *Experiment 3*. All signals were clipped to a uniform length of 1.12 s. The two time intervals used in this experiment are 5.6 ms and 2.8 ms. The programs are run offline on an 11th Gen Intel Core i5 at 2.4 GHz with 16 GB of RAM, using MATLAB 2021b. To estimate the average computational time, one hundred trials of each simulation are performed.

Fig. 8(a) and (b) are scatterplots visualizing the ER versus the computational time required by all methods, using time intervals of 5.6 and 2.8 ms, respectively. Not surprisingly, all methods perform better with shorter time intervals, albeit at the cost of longer

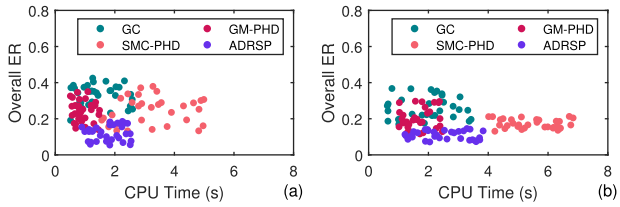

Fig. 8. ER versus time, with colors indicating whistle detector. (a) Time interval of 5.6 ms. (b) Time interval of 2.8 ms.

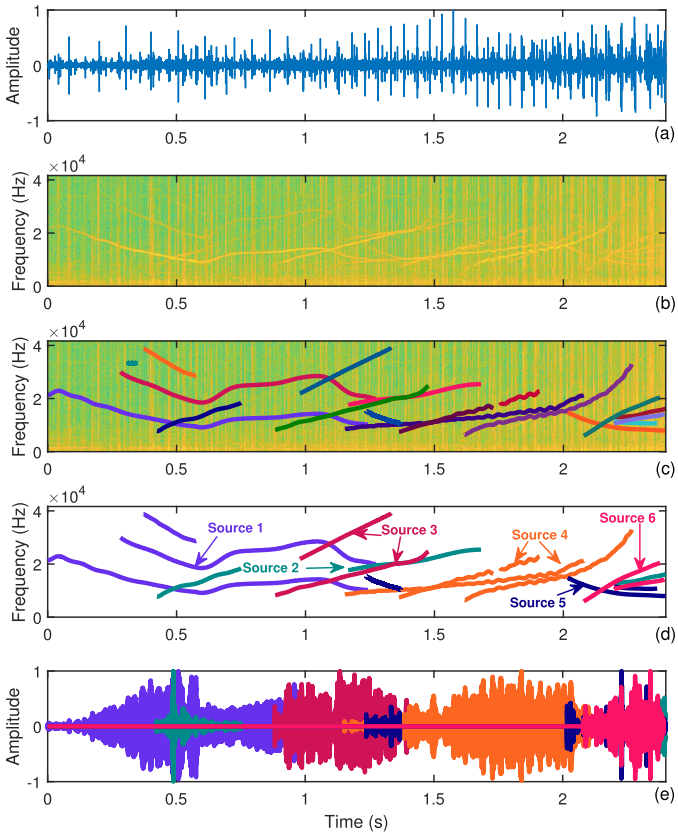

Fig. 9. Illustration of the utility of the proposed method in automatically analyzing overlapping source separation using a recording with mixed types of calls. (a) Signal waveform. (b) TFR obtained by STFT. (c) Results of whistle detection. (d) Results of the correlation analysis on these whistles and color-coded by the same source results. (e) Waveforms of the reconstructed signal from the different sources.

running times. It is observed that the GC and GM-PHD methods are faster on average while ADRSP tends to yield slower but more accurate results, requiring at least 1.77 and 2.53 s per signal for the time intervals of 5.6 and 2.8 ms, respectively. SMC-PHD can match the speed when the number of TF regions is limited but underperforms with longer time intervals. The experimental results indicate that the ADRSP method balances computational cost and accuracy, standing out for its accuracy and robustness in analyzing complex and overlapping signal components, despite its slightly slower processing speed compared to the GC and GM-PHD.

Experiment 5: To illustrate the utility of the proposed method in automatically analyzing PAM recordings, an example is provided in Fig. 9. The results demonstrate the capability of

the ADRSP to retrieve and regroup tonal components for the reconstruction of independent UWA signal types. The output of the whistle detection is shown in Fig. 9(c) where different colors represent the extracted components. Using a correlation clustering method introduced in [27], components originating from the same source are identified. The results of this cluster analysis are displayed in Fig. 9(d), where different colors represent the separated different sources. Fig. 9(e) presents the reconstructed waveforms of these sources. It is observed that interfering components, such as those of the UWA clicks and noise, have been successfully suppressed in the resulting signals.

VII. Conclusion

In this article, we introduced a novel tracker called ADRSP, designed for the detection of whistles intertwined with mixed pulsed calls in low-resolution TFRs. Using a TPS model, this method effectively removes mixed pulsed and noise components from the TFR. By shifting the focus from simply finding peaks in the TFR to a more sophisticated prediction of TF points, ADRSP uses the amplitude and directional information of ridge points to reduce the ER associated with spurious peaks. Demonstrated to effectively process real UWA signals with overlapping components at low SNR, ADRSP proves to be suitable for tasks requiring the automatic separation of overlapping sources.

The performance of the proposed method underwent comparative analysis against the GC, GM-PHD, and SMC-PHD techniques. The results demonstrated that the proposed method balances computational cost and accuracy, standing out for its accuracy and robustness in analyzing complex and overlapping components, despite its slightly slower processing speed compared to the GC and GM-PHD techniques. When applied to real UWA signals, the results also showed that the method can generate an unbroken whistle detection and reconstruct separated signals from overlapping UWA sources. This enables researchers to explore not only the similarities among whistles but also their intricate compositions. Unfortunately, this method cannot currently be directly applied to TF transformation methods other than the STFT. Addressing this limitation is one of the directions for future efforts. In future research, potential directions also involve improving the runtime efficiency and the setting of the parameter σ for this method in various scenarios, as well as developing a novel approach for the simultaneous detection of whistles and clicks.

References

[1] Y. Miao, Y. V. Zakharov, H. Sun, J. Li, and J. Wang, "Underwater acoustic signal classification based on sparse time–frequency representation and deep learning," *IEEE J. Ocean. Eng.*, vol. 46, no. 3, pp. 952–962, Jul. 2021.

[2] A. T. Johansson and P. R. White, "An adaptive filter-based method for robust, automatic detection and frequency estimation of whistles," *J. Acoust. Soc. Amer.*, vol. 130, no. 2, pp. 893–903, Jun. 2011.

[3] M. A. Roch, T. Scott Brandes, B. Patel, Y. Barkley, S. Baumann-Pickering, and M. S. Soldevilla, "Automated extraction of odontocete whistle contours," *J. Acoust. Soc. Amer.*, vol. 130, no. 4, pp. 2212–2223, Oct. 2011.

[4] A. O. T. Hogg, C. Evers, A. H. Moore, and P. A. Naylor, "Overlapping speaker segmentation using multiple hypothesis tracking of fundamental frequency," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 1479–1490, Mar. 2021.

[5] Y. Miao, Z. A. Qasem, and Y. Li, "Adaptive directional ridge prediction tracker for instantaneous frequency estimation," *Signal Process.*, vol. 209, 2023, Art. no. 109035.

[6] G. Yu and Y. Zhou, "General linear chirplet transform," *Mech. Syst. Signal Process.*, vol. 70, pp. 958–973, Mar. 2016.

[7] X. Tu, Y. Hu, F. Li, S. Abbas, and Y. Liu, "Instantaneous frequency estimation for nonlinear FM signal based on modified polynomial chirplet transform," *IEEE Trans. Instrum. Meas.*, vol. 66, no. 11, pp. 2898–2908, Nov. 2017.

[8] O. A. Alkishriwo, A. Akan, and L. F. Chaparro, "Intrinsic mode chirp decomposition of non-stationary signals," *IET Signal Process.*, vol. 8, no. 3, pp. 267–276, May 2014.

[9] I. Djurović, "QML-RANSAC instantaneous frequency estimator for overlapping multicomponent signals in the time-frequency plane," *IEEE Signal Process. Lett.*, vol. 25, no. 3, pp. 447–451, Mar. 2018.

[10] N. Laurent and S. Meignen, "A novel ridge detector for nonstationary multicomponent signals: Development and application to robust mode retrieval," *IEEE Trans. Signal Process.*, vol. 69, pp. 3325–3336, May 2021.

[11] N. A. Khan and S. Ali, "A robust and efficient instantaneous frequency estimator of multi-component signals with intersecting time-frequency signatures," *Signal Process.*, vol. 177, 2020, Art. no. 107728.

[12] H. Sun, Y. Miao, and J. Qi, "Intrinsic mode chirp multicomponent decomposition with kernel sparse learning for overlapped nonstationary signals involving Big Data," *Complexity*, vol. 2018, 2018, Art. no. 8426790.

[13] Y. Miao, "Automatic instantaneous frequency estimator for multicomponent signals with the variable number of components," *Signal Process.*, vol. 197, 2022, Art. no. 108541.

[14] B. Mohammad and R. McHugh, "Automatic detection and characterization of dispersive north atlantic right whale upcalls recorded in a shallow-water environment using a region-based active contour model," *IEEE J. Ocean. Eng.*, vol. 36, no. 3, pp. 431–440, Jul. 2011.

[15] A. Mallawaarachchi, S. Ong, M. Chitre, and E. Taylor, "Spectrogram denoising and automated extraction of the fundamental frequency variation of dolphin whistles," *J. Acoust. Soc. Amer.*, vol. 124, no. 2, pp. 1159–1170, Aug. 2008.

[16] D. Kipnis and R. Diamant, "Graph-based clustering of dolphin whistles," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 2216–2227, Jun. 2021.

[17] P. Gruden and P. R. White, "Automated tracking of dolphin whistles using Gaussian mixture probability hypothesis density filters," *J. Acoust. Soc. Amer.*, vol. 140, no. 3, pp. 1981–1991, Sep. 2016.

[18] T. Zhou, Y. Wang, L. Zhang, B. Chen, and X. Yu, "Underwater multitarget tracking method based on threshold segmentation," *IEEE J. Ocean. Eng.*, vol. 48, no. 4, pp. 1255–1269, Oct. 2023.

[19] P. Gruden and P. R. White, "Automated extraction of dolphin whistles—a sequential monte carlo probability hypothesis density approach," *J. Acoust. Soc. Amer.*, vol. 148, no. 5, pp. 3014–3026, Nov. 2020.

[20] Y. Miao, J. Li, and H. Sun, "Multimodal sparse time–frequency representation for underwater acoustic signals," *IEEE J. Ocean. Eng.*, vol. 46, no. 2, pp. 642–653, Apr. 2021.

[21] H. Lohrasbipeydeh, D. T. Dakin, T. A. Gulliver, H. Amindavar, and A. Zielinski, "Adaptive energy-based acoustic sperm whale echolocation click detection," *IEEE J. Ocean. Eng.*, vol. 40, no. 4, pp. 957–968, Oct. 2015.

[22] R. Byers, "A Hamiltonian-Jacobi algorithm," *IEEE Trans. Autom. control*, vol. 35, no. 5, pp. 566–570, May 1990.

[23] U. Kjems, J. B. Boldt, M. S. Pedersen, T. Lunner, and D. Wang, "Role of mask pattern in intelligibility of ideal binary-masked noisy speech," *J. Acoust. Soc. Amer.*, vol. 126, no. 3, pp. 1415–1426, Sep. 2009.

[24] M. H. Radfar and R. M. Dansereau, "Single-channel speech separation using soft mask filtering," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 8, pp. 2299–2310, Nov. 2007.

[25] D.-H. Pham and S. Meignen, "High-order synchrosqueezing transform for multicomponent signals analysis—with an application to gravitational-wave signal," *IEEE Trans. Signal Process.*, vol. 65, no. 12, pp. 3168–3178, Jun. 2017.

[26] P. Li et al., "Learning stage-wise GANs for whistle extraction in time-frequency spectrograms," *IEEE Trans. Multimedia*, vol. 25, pp. 9302–9314, Mar. 2023.

[27] S. Kim, C. D. Yoo, S. Nowozin, and P. Kohli, "Image segmentation using higher-order correlation clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 9, pp. 1761–1774, Sep. 2014.

Yongchun Miao received the B.S. and M.Sc. degrees in software engineering from the Jiangxi Normal University, Nanchang, China, in 2013 and 2016, respectively, and the Ph.D. degree in communication engineering from Xiamen University, Xiamen, China, in 2021.

From 2019 to 2020, she visited the Department of Electronic Engineering, University of York, York, U.K. She is currently a Lecturer with the School of Electronic and Information Engineering, Anhui University, Hefei, China. Her research interests include underwater acoustic signal processing and image processing.

Jianghui Li received the B.S. degree in communications engineering from the Huazhong University of Science and Technology, Wuhan, China in 2011, the M.Sc. degree in communications engineering, and the Ph.D. degree in electronics engineering from the University of York, York, U.K., in 2013 and 2017, respectively.

From 2017 to 2021, he was a Research Fellow with the University of Southampton, Southampton, U.K. Since 2021, he has been a Professor with Xiamen University, Xiamen, China. His current research interests include offshore carbon capture, utilization and storage, underwater acoustics, adaptive signal processing, and ocean engineering.

Yingsong Li (Senior Member, IEEE) received the B.S. degree in electrical and information engineering and the M.S. degree in electromagnetic field and microwave technology from Harbin Engineering University, Harbin, China, in 2006 and 2011, respectively, and the Ph.D. degree in fundamental engineering from the Kochi University of Technology (KUT), Kochi, Japan, and in communication and information system from Harbin Engineering University, Harbin, China, in 2014.

He was a Professor with Harbin Engineering University from 2014 to 2022, a Visiting Scholar with the University of California at Davis, Davis, CA, USA, from March 2016 to March 2017, and a Visiting Professor with the University of York, York, U.K., in 2018, and Far Eastern Federal University, Vladivostok, Russia, and KUT. From 2016 to 2020, he was a Postdoctoral Fellow with the Key Laboratory of Microwave Remote Sensing, Chinese Academy of Sciences, Beijing, China. He has been a Full Professor with the School of Electronic and Information Engineering, Anhui University, Hefei, China, since March 2022. He is currently a Visiting Professor with the School of Information, KUT, which he began in 2018. He has authored or coauthored 300 journal articles and conference papers in various areas of electrical engineering. His research interests include remote sensing, underwater communications, signal processing, adaptive filters, metasurface designs, and microwave antennas.

Dr. Li is a Fellow of the Applied Computational Electromagnetics Society (ACES) and a Senior Member of the Chinese Institute of Electronics. He is a Reviewer for numerous IEEE, IET, Elsevier, and other international journals. He was the General Co-Chair of ICEICT in 2020 and the General Chair of IEEE 9th International Conference on Computer Science and Network Technology (ICCSNT 2021) and ICCSNT in 2022. He is the TPC Co-Chair of the 2019 IEEE International Workshop on Electromagnetics (iWEM 2019.2020), 2019 IEEE 2nd International Conference on Electronic Information and Communication Technology (ICEICT 2019), 2019 International ACES Symposium-China, and 2019 Cross Strait Quad-regional Radio Science and Wireless Technology Conference (2019 CSQRWC), and the TPC Chair of ICEICT during 2021.2022. He is also the Session Chair or an Organizer for many international and domestic conferences, including the WCNC, AP-S, and ACES. He was an Area Editor for the *AEÜ-International Journal of Electronics and Communications* from 2017 to 2020. He is an Associate Editor for IEEE ACCESS, ACES Journal, and *Alexandria Engineering Journal*.