A NEW APPROACH FOR CLASSIFICATION OF DOLPHIN WHISTLES

Mahdi Esfahanian, Hanqi Zhuang, Nurgun Erdol

Department of Computer & Electrical Engineering and Computer Science, Florida Atlantic University

ABSTRACT

This paper presents a novel approach to categorize dolphin whistles into various types. Most accurate methods to identify dolphin whistles are tedious and not robust, especially in the presence of ocean noise. One of the biggest challenges of dolphin whistle extraction is the coexistence of short-time duration wide-band echo clicks with the whistles. In this research a subspace of select orientation parameters of the 2-D Gabor wavelet frames is utilized to enhance or suppress signals by their orientation. The result is a Gabor image that contains a noise free grayscale representation of the fundamental dolphin whistle which is resampled and fed into the Sparse Representation Classifier. The classifier uses the ℓ_1-norm to select a match. Experimental studies conducted demonstrate: (a) a robust technique based on the Gabor wavelet filters in extracting reliable call patterns, and (b) the superior performance of Sparse Representation Classifier for identifying dolphin whistles by their call type.

Index Terms— Gabor Wavelets, Sparse Representation Classifier, Whistle Classification.

1. INTRODUCTION

Gabor wavelets [1,2] have been used to extract robust feature for applications such as fingerprint recognition [3], texture segmentation [4], and handwritten numerals recognition [5]. The optimality of Gabor wavelets for extraction of local features can be explained from three aspects: 1) the receptive fields of simple cells in the visual cortex are very similar to the shapes of Gabor wavelets [6], 2) local spatial frequencies can be measured optimally by mathematical means [7], and 3) Gabor wavelets can yield distortion tolerant feature spaces for pattern recognition tasks. Facial images used for recognition [8] are dense in the image frame; i.e., with proper cropping, most of the picture is the face. The focus of this paper is to classify dolphin whistles by their types, types being associated with the shapes of the contours in the time-frequency plane. It is argued by marine mammal biologists that messages are encoded in whistle types. The classifier we attempt to adopt is the Sparse Representation Classifier (SRC) [8] which, under hypothetical conditions, requires no feature extraction. The SRC, developed for face recognition, is based on the premise that if there exists a large dictionary of training data that consists of m sample faces for each subject and there are M subjects, then a test face is sparsely represented in the dictionary if a) it can be reconstructed through its projections on the associated m training faces, and b) if $m \ll Mm$. This one sentence description is over simplified, nevertheless captures the essence of the SRC. In [8] the authors show that if sparse representation accurately models the data then the exact nature of the features is no longer critical as long as they are adequate in quantity. The sparsity exploited in the method is derived from the compressive sensing theory [9, 10, 11] and relies on the training data set to form a near complete, or over complete, dictionary. Claims of both the liberal choice of features and the robustness to occlusion are welcome improvements to the classification of not only images, but also the time-evolutionary spectral representation of non-stationary temporal signals. Image processing tools may be applied to find spectrogram-based features that model short or long-term spectral dynamics of signals. For instance, 2-D spectro-temporal filters have been employed on Mel-scale spectrogram patterns for automatic speech recognition using the SRC technique resulting in an impressive decrease of 28% in word error rate [12, 13].

Application of the SRC to categorize dolphin whistles by their types requires some preprocessing that ensures that a whistle is adequately represented by its projections on the training whistles in its category. In the face recognition problem studied in [8], the authors crop the facial images until the face covers the entire image and there is no space in the image, that is, the face is dense in the image. Spectrograms have been used as images [14] to classify dolphin vocalizations. As can be seen in Fig. 1, the whistles are not dense on the time-frequency plane and they are subject to not only ocean noise but also interference of the echo-clicks. Echo clicks are impulsive signals that dolphins use for navigation. They are frequently produced simultaneously with whistles and they appear as vertical striations in the spectrogram. Clearly the equivalent of cropping of facial images cannot be applied to ensure adequate representation of whistles by others in their own category. We propose a method based on Gabor wavelet filters that generate a simple binary image (see Fig. 4) of the fundamental whistle. The Gabor image is generated by the application of Gabor wavelets with selective orientation

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

parameters. The method not only suppresses the echolocation clicks but also generates whistle images that form, from a small number of training data, a set of near complete exemplars of their own category. It is shown in the paper that Gabor images used as feature vectors to SRC produce superior classification performance for identification of dolphin whistles by their type.

The remainder of the paper is organized as follows. The characteristics of the dolphin vocalization data are described in Section 2. Gabor wavelets and SRC classifier are outlined in Sections 3 and 4, respectively. Section 5 presents the experimental results along with discussion. Finally, we conclude the paper in Section 6.

2. BACKGROUND INFORMATION

Dolphins generate various types of sounds such as whistles, clicks, pulsed tones, and noise. Whistles consist of 2-4 narrow band signals with harmonically related, time-varying frequencies [15]. They are believed to be mostly for the purpose of communication while echolocation clicks are short-time wide-band bursts emitted for object detection and distance measurement. Hydrophone recordings are often contaminated with abundant and diverse noise. Spectrograms in Fig. 2 show four types of whistles embedded in competing trains of wide-band pulses. The classification of whistles according to their type can be described by the manner the fundamental frequency changes over the time as illustrated clockwise from the top left in Fig. 2: upswing, convex-up, convex down and up-and-down [16]. We refer to them as Class 1 through Class 4, respectively. The spectro-temporal filters are designed to limit the frequency range to 4 - 16 kHz and window the whistle into the appropriate time range. In this research, the spectrograms have been built with 13 ms Hamming windows and 50% overlap.

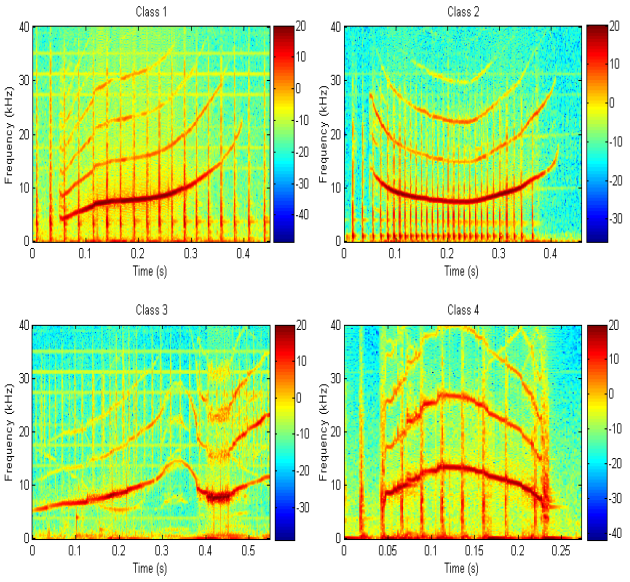

Fig. 1. Spectrograms of four distinct dolphin whistles

In our previous paper [17], the preprocessing involves spectral denoising to remove the vertical striations due to echolocation clicks followed by a contour extraction algorithm based on spectral peak picking. Then two sets of feature vectors known as Fourier descriptors (FDs) and time-frequency parameters (TFPs) consisting of minimum, maximum, start, and end frequencies, frequency range, time duration and number of inflection points were extracted from the contour. Derivation of both sets of features requires tedious, and time consuming computation. However, application of Gabor wavelets eliminates such a need.

3. GABOR WAVELETS

Gabor wavelets are capable of modeling two-dimensional receptive field profiles encountered experimentally in cortical simple cells, which captures their salient tuning properties of spatial localization, spatial orientation and spatial-frequency selectivity. The parameterized family of 2-D Gabor filters [18,19], which is a Gaussian kernel function modulated by a sinusoid plane wave, is defined as follows:

$$\psi_{\mu,\nu}(z) = \frac{\|k_{\mu,\nu}\|^2}{\sigma^2} e^{-\frac{\|k_{\mu,\nu}\|^2 \|z\|^2}{2\sigma^2}} [e^{ik_{\mu,\nu} z} - e^{-\sigma^2/2}] \quad (1)$$

where μ and ν correspond to the orientation and scale of Gabor filter, $z = (x, y)$, and $\|.\|$ denotes the norm operator. The wave vector $k_{\mu,\nu}$ can be defined as:

$$k_{\mu,\nu} = k_\nu e^{i\phi_\mu} \quad (2)$$

Here $k_\nu = k_{\max} / f^\nu$ and $\phi_\mu = \pi\mu / N$, where $k_{\max}$ and f are maximum frequency and spacing factor between kernels in the frequency domain, respectively, and N is number of orientations.

To spot local patterns in the image, the Gabor wavelets at four scales $\nu = \{0,1,2,3\}$ and eight orientations $\mu = \{0,1,...,7\}$ were used. The following values of the parameters have been chosen experimentally for the best performance: $\sigma = \pi$, $k_{\max} = \pi/2$ and $f = \sqrt{2}$. Fig. 2(a) and 2(b) show respectively real components and magnitudes of Gabor kernels at different orientation and scales.

The Gabor wavelet representation is the convolution of the spectrogram with a family of Gabor kernels, which is defined as follows:

$$W_{\mu,\nu}(z) = I(z) * \psi_{\mu,\nu}(z) \quad (3)$$

where $I(z)$ denotes the spectrogram image, and * represents the convolution operator. Each convolution output in (3) exhibits the characteristics of spatial locality, and scale and orientation selectivity. Since $W_{\mu,\nu}(z)$ is a complex function, the magnitude response is used for feature representation. Conventionally, magnitude of each output is vectorized as a column (or row) and then all vectors are concatenated to form a feature vector. However, in this paper, a different approach is taken. In our method, all 32 convolution outputs in (3) are summed up to obtain a single complex image. Then its normalized magnitude is computed to form the so-

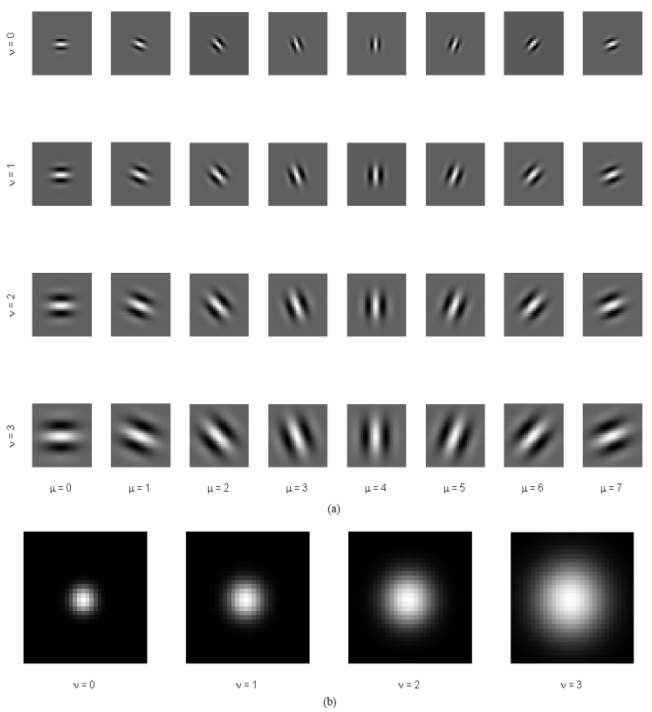

Fig. 2. (a) The real part of the Gabor wavelets at different scales and orientations. (b) The magnitude of the Gabor wavelets at four scales.

called "Gabor image". Subsequently, each Gabor image is vectorized and placed in a dictionary matrix *U*, which is basically a data set in the feature space.

4. SPARSE REPRENTATION CLASSIFIER

The application of compressive sensing to classification has been reported in [8] where sparse representation of the test sample is determined using the training data. A dictionary $U = [U_1, U_2, \ldots, U_d] \in \mathbb{R}^{m \times n}$ is formed by concatenating classes $U_i \in \mathbb{R}^{m \times n_i}$ whose columns consist of training vectors $\{u_i^k, k=1,2,\ldots,n_i\}$. The parameters d, $n = n_1+n_2+\ldots+n_d$ and m refer to the number of classes, total number of whistles in the dictionary, and number of samples per whistle, respectively. The hypothesis is that a test sample $y \in \mathbb{R}^m$ from the i^{th} class can be represented as a linear combination of training samples of the same class. If the representation is exact, then the coefficient vector $\alpha \in \mathbb{R}^n$ in $y = U\alpha$ is $n_i << n$ sparse, meaning that all elements should be zero but those associated with the same class. An algorithm can be applied to obtain the solution under the minimum ℓ_1-norm. However, there is no guarantee that α will have the expected form since the test sample *y* from class *i* may not necessarily be orthogonal to vectors from other classes. The following model introduces the constraint to establish a bound on the residual energy as:

$$\bar{\alpha} = \arg\min \|\alpha\|_1 \quad \textit{subject to} \quad \|y - U\alpha\|_2 \le \varepsilon \qquad (4)$$

with an error tolerance $\varepsilon > 0$. The above equation anticipates the noisy case of cross correlation with the "out of class" training vectors and selects the class that also satisfies the residual constraint of the measurement set. The success of the classifier relies strongly on the effective sparseness of the coefficient set of a test vector. If the test vector belongs to class *i*, then the coefficients should be concentrated in the entries corresponding to the class. Since the way of determining the coefficients is an inner product, the existence of a dictionary vector that is a near match to the test vector is crucial. For this reason, the dictionary matrix must be of a sufficiently size and the noise component of the vectors must be uncorrelated to the signal component and white for optimal results in terms of the residual. The signal component of dolphin whistles are narrow band (cf. Fig. 1), and there are many competing spectral structures such as the wide band echolocation clicks and the monotone anthropomorphic noise (horizontal striations). Gabor features are capable of feeding the classifier with distinguishing information for the recognition of various whistle types even in the present of unwanted interference.

5. RESULTS AND DISCUSSION

Recordings of free-ranging bottlenose dolphins from the resident Sarasota Bay located in north-west of Florida were made nearly annually during brief capture-release events [20, 21]. Custom-built suction-cup hydrophones were attached on the forehead of each individual. Thus the signal-to-noise ratio and general background noise was similar in all the recordings. The hydrophones were not calibrated because amplitude values were not being measured. Whistles were recorded onto various stereo-cassette or video-cassette recorders available on the market at the time.

In this work, features of dolphin whistles extracted from spectrogram images by Gabor wavelets are fed to the SRC algorithm to distinguish various dolphin whistle types. To evaluate our algorithms, a collection of 100 bottlenose dolphin whistles were processed of types one to four. Half of this collection was randomly used for training and the remaining half for testing. All the recordings were sampled at 80 kHz and band-pass filtered between 4 kHz and 16 kHz to restrict the input to the fundamental whistle. The frame length of 1024 samples with 50% overlap followed by cropping to the pass-band frequency range builds spectrograms with 615 frequency bins and 21-42 temporal bins because duration of different whistles is not always equal. To create a dictionary matrix for the classifier, the original spectrogram was resized to a smaller yet informative size 30×30 in order to reduce the amount of data to be processed without loss of salient information. In the next step, 32 Gabor wavelet kernels were convolved with the resized spectrograms individually and their outputs are summed up to obtain the Gabor images whose normalized magnitudes are concatenated column-wisely to create the

columns of matrix *U*. Fig. 3 illustrates the derived Gabor images of four distinct whistles. These choices were empirically selected after the consideration of trade-off between the feature vector length and recognition accuracy. Experiment results in Table 1show the confusion matrix of SRC using features obtained by Gabor wavelets, with an overall classification accuracy of 98%. In this case, just one misclassification was occurred in a very difficult case involving a third-class whistle demonstrated in Fig. 4. As observed, even human eyes may fail to recognize the type of whistle, which shows the effectiveness and robustness of both the Gabor wavelets at whistle detection and representation, and the SRC technique at recognizing different whistle types even when some outliers are present in the Gabor images of Fig. 3. In fact, one of the most important factors for the success of SRC is that whistles being present in the Gabor images are sparse and this characteristic matches the features well with the proposed classifier.

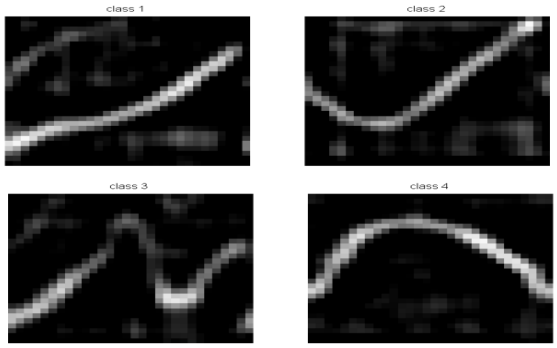

Fig. 3. The Gabor images of different whistle types

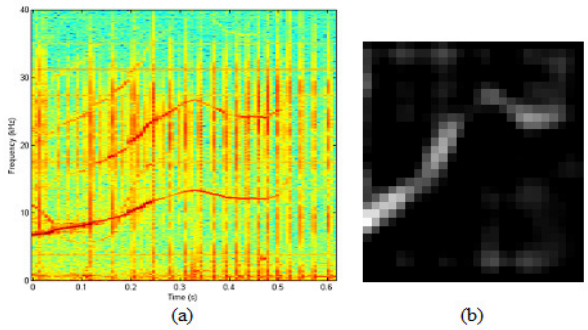

Fig. 4. Third-class misclassified whistle (a) Original spectrogram (b) Gabor image

To evaluate the performance of the Gabor wavelets and SRC combination, two additional feature extraction techniques, named Time-Frequency Parameters (TFPs) and Fourier Descriptors (FDs) [17], were used in place of Gabor wavelets. Tables 2 and 3 present the confusion matrices of SRC using FDs and TFPs, respectively. An accuracy of 88% implies that six whistles were misclassified; refer to Table 2. Table 3 reveals that four misclassifications occurred with TFP features, leading to an accuracy of 92%. The comparison of classification results reveals the capability of Gabor wavelets for extracting relevant features of dolphin whistles in the spectrogram. We believe that since Gabor features extracted from whistle contours are sparse in spectral domain, it meshes well with SRC because sparsity is an important property of SRC based on the compressive sensing principles. The resulting images obtained after application of Gabor wavelets on whistle spectrograms in Fig. 3 illustrates the sparsity of Gabor images.

	1st class (%)	2nd class (%)	3rd class (%)	4th class (%)
1st class	100	0	0	0
2nd class	0	100	0	0
3rd class	0	0	92	8
4th class	0	0	0	100

Table 1. Confusion matrix of SRC + Gabor

	1st class (%)	2nd class (%)	3rd class (%)	4th class (%)
1st class	100	0	0	0
2nd class	0	94	0	6
3rd class	33	0	58	9
4th class	0	0	0	100

Table 2. Confusion matrix of SRC + FDs

	1st class (%)	2nd class (%)	3rd class (%)	4th class (%)
1st class	85	15	0	0
2nd class	0	100	0	0
3rd class	0	0	83	17
4th class	0	0	0	100

Table 3. Confusion matrix of SRC + TFPs

6. CONCLUSION

A new procedure has been proposed, which combines a feature extraction method centered at Gabor wavelets and a compressive-sensing based technique called Sparse Representation Classifier, for the classification of dolphin whistles by call type. It has been shown that the proposed approach avoids the need for tedious preprocessing steps while achieving a superior classification performance in comparison with several other techniques. It is worthwhile to mention that an earlier paper by the authors [21] reported that a feature extraction method named Local Binary Patterns (LBP) had been employed for classification of bottlenose whistles by type. The accuracy performance obtained from the combination of LBP and SVM classifier is comparable with that of Gabor and SRC. For future work, we are eager to test the proposed approach on large data sets of a variety of marine mammal vocalizations.

7. REFERENCES

[1] M. Lades, J. C. Vorbruggen, J. Buhmann, J. Lange, C. Von der Malsburg, R. P. Wurtz, and W. Konen, "Distortion invariant object recognition in the Dynamic Link Architecture," IEEE Transactions on Computers, vol. 42, no. 3, pp. 300-311, 1993.

[2] L. Wiskott, J. M. Fellous, N. Kruger, C. Von der Malsburg, "Face recognition by elastic bunch graph matching," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 19, no. 7, pp. 775-779, 1997.

[3] C. J. Lee, and S. D. Wang, "Fingerprint feature extraction using Gabor filters," Electronics Letters, vol. 35, no. 4, pp. 288–290, 1999.

[4] T. P. Weldon, W. E. Higgins, and D. F. Dunn, "Efficient Gabor filter design for texture segmentation," Pattern Recognition, vol. 29, no. 12, pp. 2005-2015, 1996.

[5] Y. Hamamoto, S. Uchimura, M. Watanabe, T. Yasuda, Y. Mitani, and S. Tomita, "A Gabor filter-based method for recognizing handwritten numerals," Pattern Recognition, vol. 31, no. 4, pp. 395-400, 1998.

[6] J. G. Daugman, "Two-Dimensional Spectral Analysis of Cortical Receptive Field Profiles," Vision Research, vol. 20, pp. 847-856, 1980.

[7] V. Kruger, and G. Sommer, "Gabor wavelet networks for efficient head pose estimation," Image and Vision Computing, vol. 20, no. 9, pp. 665-672, 2002.

[8] J. Wright, A. Y. Yang, A. Ganesh, S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 31, no. 2, pp. 210-227, 2009.

[9] D. L. Donoho, "Compressed Sensing," IEEE Transactions on Information Theory, vol. 52, no. 4, pp. 1289-1306, 2006.

[10] R. G. Baraniuk, "Compresseive sensing," IEEE Signal Processing Magazine, vol. 24, no. 4, pp. 118-121,2007.

[11] E. J. Candès, and M. B. Wakin, "An introduction to compressive sensing," IEEE Signal Processing Magazine, vol. 25, no. 2, pp. 21-30, 2008.

[12] A. Hurmalainen, and T. Virtanen, "Modelling spectro-temporal dynamics in factorisation-based noise-robust automatic speech recognition," in Proc. of the IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), pp. 4113-4116, Japan, 2012.

[13] J. F. Gemmeke, T. Virtanen, and A. Hurmalainen, "Exemplar-based sparse representations for noise robust automatic speech recognition," IEEE Transactions on Audio, Speech and Language Processing, vol. 19, no. 7, pp. 2067-2080, 2011.

[14] M. Esfahanian, H. Zhuang, and N. Erdol, "Using local binary patterns as features for classification of dolphin calls," Journal of Acoustical Society of America, vol. 134, no. 1, 2013.

[15] A. N. Propper, "Sound emission and detection by Delphinids," Cetacean Behavior: Mechanisms and Functions, pp. 1-52, 1980.

[16] L. S. Sayigh, and V. M. Janik, *Signature whistles*, pp. 1014–1016. In: W. F. Perrin, B. Würsig, J. G. M. Thewissen(eds), *Encyclopedia of Marine Mammals*, Elsevier, Inc., San Diego, CA, 2009.

[17] M. Esfahanian, H. Zhuang and N. Erdol, "On contour-based classification of dolphin whistles by type," Journal of Applied Acoustics, vol. 67, pp. 276-279, 2014.

[18] J. G. Daugman, "Complete discrete 2-D Gabor transforms by neural networks for image analysis and compression," IEEE Transactions on Acoustics, Speech, and Signal Processing (ICASSP), vol. 36, no. 7, pp. 1169-1179, 1988.

[19] S. Marcelja, "Mathematical description of the responses of simple cortical cells," Journal of Optical Society of America, vol. 70, no. 11, pp. 1297-1300, 1980.

[20] J. J. Dreher, and W. E. Evans, *Cetacean communication*, pp. 373–393. In: W. N. Tavolga (ed), *Marine Bioacoustics*, Pergamon Press, Oxford, 1964.

[21] M. Janik, L. S. Sayigh, and R. S. Wells, "Signature whistle shape conveys identity information to bottlenose dolphins," in Proceeding of the National Academy of Sciences (PNAS), vol. 103, no. 21, pp. 120-125, 2006.