Perceptually motivated wavelet packet transform for bioacoustic signal enhancement

Yao Ren[a)], Michael T. Johnson and Jidong Tao
Speech and Signal Processing Laboratory, Marquette University, P.O. Box 1881, Milwaukee, Wisconsin 53233-1881

[a)] Electronic-mail: yao.ren@marquette.edu

Bioacoustic signal enhancement

A significant and often unavoidable problem in bioacoustic signal processing is the presence of background noise due to an adverse recording environment. This paper proposes a new bioacoustic signal enhancement technique which can be used on a wide range of species. The technique is based on a perceptually scaled wavelet packet decomposition using a species-specific Greenwood scale function. Spectral estimation techniques, similar to those used for human speech enhancement, are used for estimation of clean signal wavelet coefficients under an additive noise model. The new approach is compared to several other techniques, including basic band-pass filtering as well as classical speech enhancement methods such as spectral subtraction, Wiener filtering, and Ephraim Malah filtering. Vocalizations recorded from several species are used for evaluation, including the ortolan bunting (*Emberiza hortulana*), rhesus monkey (*Macaca mulatta*), and humpback whale (*Megaptera novaeanglia*), with both additive white Gaussian noise and environment recording noise added across a range of Signal-to-Noise Ratios (SNR). Results, measured by both SNR and Segmental SNR (SSNR) of the enhanced waveforms, indicate that the proposed method outperforms other approaches for a wide range of noise conditions.

PACS numbers: 43.60.Hj, 43.50.Rq, 43.80.-n

I. INTRODUCTION

The presence of background noise and interfering signals is a fundamental problem in the collection and analysis of bioacoustic data, regardless of the specific species under study or the type of environment. This noise takes a variety of forms, including ambient background noise due to weather conditions, continuous interference from nearby vehicular or boat traffic, or the presence of numerous non-target vocalizations from other species and individuals. Since the distance from the acoustic recording device to the individuals under study can be quite large leading to significant signal attenuation, interfering noise can create a substantial obstacle to analysis and understanding of the desired vocalization patterns.

Common techniques to reduce noise artifacts in bioacoustic signals include basic band-pass filters and related frequency-based methods for spectrogram filtering and equalization, often incorporated directly into acquisition and analysis tools (Mellinger, 2002). Other approaches in recent years have included spectral subtraction (Liu *et al.*, 2003), Minimum Mean-Squared Error (MMSE) estimation (Álvarez and García, 2004), adaptive line enhancement (Yan *et al.*, 2005; Yan *et al.*, 2006) and denoising using wavelets (Gur and Niezrecki, 2007).

In comparison, there are a wide variety of advanced techniques used for human speech enhancement, some of which form the basis for the more recent bioacoustic enhancement methods cited above. Historically the most common approaches for speech enhancement have focused on spectral subtraction (Boll, 1979), Wiener filtering (Lim and Oppenheim, 1978), and MMSE and log-MMSE estimation using Ephraim Malah filtering (Ephraim and Malah, 1984; 1985). Added to this in recent years are newer methods based on sub-space estimation and filtering (Ephraim and Trees, 1995) and wavelet decomposition (Johnson *et al.*, 2007).

In this paper, we introduce a new bioacoustic signal enhancement technique which is based on a perceptually-scaled wavelet packet decomposition, using spectral estimation methods

similar to those used for human speech enhancement. The underlying goal is to obtain higher quality and more intelligible enhanced signals through the use of more perceptually meaningful frequency representations. This method is robust across a wide range of species, needing only $f_{\min}$ and $f_{\max}$ frequency boundary parameters to generalize for application to a new species of interest.

The new method is compared to a variety of other enhancement and denoising techniques, including simple band-pass filtering, spectral subtraction, Wiener filtering, and the Ephraim-Malah log-MMSE estimation. To evaluate and compare its applicability across a variety of species, the method is applied to the animals of the order Passeriformes (ortolan bunting), Primates (rhesus monkey) and Cetaceans (humpback whale). Evaluation is done using both Signal-to-Noise Ratio (SNR) and Segmental SNR (SSNR), which is known to be a more perceptually relevant quality measure for human speech (Deller *et al.*, 2000).

II. CURRENT ENHANCEMENT METHODS

A. Band-pass filtering

Band-pass filtering removes signal energy outside of a specified frequency range. This can be applied in either the time-domain or the frequency domain (e.g. applied to a spectrogram), and is effective primarily in cases where signals are predominately narrow-band and are well separated from the noise spectrum.

B. Spectral subtraction

Spectral subtraction (Boll, 1979) was one of the first algorithms applied to the problem of speech enhancement. It is based directly on the additive noise model:

$$y(n) = x(n) + d(n)\,, \tag{1}$$

where $y(n)$, $x(n)$ and $d(n)$ denote the noise-corrupted input signal, clean signal and additive noise signal, respectively. The noise spectrum is estimated from the Fourier Transform magnitude of a silent region in the waveform, so that for each frame of the signal an estimate for the clean signal in the frequency domain can be given directly as

$$\hat{X}(\omega) = \left[|Y(\omega)| - |\hat{D}(\omega)| \right] e^{j\phi_y(\omega)}, \tag{2}$$

where $\phi_y(\omega)$ is the phase component of the noisy signal, used under the assumption that spectral phase is much less important than spectral magnitude for reconstruction.

Note that application of equation (2) may result in negative magnitude values, which are typically set to zero. This often results in some processing artifacts that are usually described by listeners as "musical tones". The presence of such artifacts is one disadvantage of the spectral subtraction approach.

C. Wiener filtering

Wiener filtering is conceptually similar to spectral subtraction, but replaces the direct subtraction with a mathematically optimal estimate for the signal spectrum in a MMSE sense (Lim and Oppenheim, 1978).

The frequency domain formulation of the Wiener filter is given as

$$H(\omega) = \frac{S_{xx}(\omega)}{S_{xx}(\omega) + S_{dd}(\omega)}, \tag{3}$$

where $H(\omega)$ is the desired filter response and $S_{xx}(\omega)$ and $S_{dd}(\omega)$ are power spectral densities (PSD) of the desired clean signal and noise. Since these two PSDs are unknown, this filter cannot be determined directly and instead needs to be realized in an iterative fashion. In particular, $S_{dd}(\omega)$ is estimated from a silence region and $S_{xx}(\omega)$ is initialized from the noisy waveform and

then updated from the output of the filter after each iteration. This process is repeated either a fixed number of times or until a convergence criteria is reached.

D. Ephraim Malah filtering

The Wiener filter is an optimal linear estimator of the clean signal spectrum in an MMSE sense. Ephraim and Malah extended this idea by deriving an optimal nonlinear estimator of the clean spectral amplitude. This estimator assumes the real and imaginary parts of the spectral magnitude have a zero-mean Gaussian probability density distribution and are statistically independent. Under this statistical model, a Short Time Spectral Amplitude (STSA) estimator was derived using the MMSE optimization criteria (Ephraim and Malah, 1984). This work was then modified to use log-spectra rather than spectra as an optimization criteria (Ephraim and Malah, 1985), since log spectral distance is a more perceptually relevant distortion criteria, resulting in improved overall enhancement results. This estimator, known as the Ephraim Malah (EM) filter, can be summarized using the following estimation formula for the clean signal Fourier transform coefficient $\hat{A}_k$ in each frequency bin:

$$\hat{A}_k = \frac{\xi_k}{1+\xi_k} e^{\left(\frac{1}{2}\int_{\upsilon_k}^{\infty} \frac{e^{-t}}{t} dt\right)} R_k , \tag{4}$$

In this equation, $\xi_k = \frac{\lambda_x(k)}{\lambda_d(k)}$, $\upsilon_k = \frac{\xi_k}{1+\xi_k}\gamma_k$ and $\gamma_k = \frac{R_k^2}{\lambda_d(k)}$, where R_k^2 is the noisy speech Fourier transform magnitude in the *k*th frequency bin, and $\lambda_d(k)$ and $\lambda_x(k)$ are the average noise and signal powers in each bin. Similar to the spectral subtraction method, the noise power is estimated from silence regions in the waveform, while $\lambda_x(k)$ is a moving average of spectrally subtracted noisy spectra $\left(R_k^2 - \lambda_d(k)\right)$. The *a priori* SNR ξ_k is estimated via Ephraim Malah's

well-known "decision-directed method", which is updated from the previous amplitude estimate using a forgetting factor α as follows:

$$\hat{\xi}_k(n) = \alpha \frac{\hat{A}_k^2(n-1)}{\lambda_d(k,n-1)} + (1-\alpha) P\left[\gamma_k(n)-1\right], \tag{5}$$

where the indicator function P is given by

$$P\left[\gamma_k(n)-1\right] \triangleq \begin{cases} \gamma_k(n)-1 & \gamma_k(n)-1 \geq 0 \\ 0 & otherwise \end{cases} \tag{6}$$

The key characteristics of this estimator are that it tends to do less enhancement (i.e., less change to the noisy signal spectrum) when the SNR is high, and that musical noise artifacts are significantly reduced.

E. Wavelet denoising

Spectral subtraction, Wiener filtering, and Ephraim Malah filtering are all based on the same mathematical tool, the Short Time Fourier Transform (STFT), with the waveform divided into short frames during which the signal is assumed to be stationary. The STFT is a compromise between time resolution and frequency resolution: a shorter frame length results in a better time resolution but poorer frequency resolution. The Wavelet Transform (WT) by comparison has the advantage of implicitly using a variable window size for different frequency components. This often results in better handling of broadband non-stationary signals, including speech and bioacoustic data.

Whereas the STFT is a function of frequency for each individual signal frame, the WT is a function of two variables, time and scale. Scale is used rather than frequency because depending on the wavelet basis being used each scale may actually represent information across a range of frequencies. Like the Fourier Transform, the Wavelet Transform has both continuous (CWT) and discrete (DWT) implementations. A DWT can be efficiently implemented using a Quadrature

Mirror Filter (QMF) decomposition, resulting in scales that are powers of two, called a dyadic transform. A further generalization of the DWT is the Wavelet Packet Transform (WPT). In the WPT, the filtering process is iterated on both the low frequency and high frequency components, whereas the DWT iterates only on the low frequency components. Filter decomposition structures for the DWT and WPT are shown in Fig. 1. In the decomposition tree, each node is labeled (l,n) where l is the decomposition level and n represents a subband node index. The root of the tree $(l,n)=(0,0)$ refers to the entire signal space. The left and right branches denote low-pass and high-pass filtering followed by 2:1 down-sampling, respectively.

The application of wavelets for signal enhancement, sometimes referred to as denoising, is a three step procedure involving wavelet decomposition, wavelet coefficient thresholding and wavelet reconstruction. Given an appropriate choice of wavelet basis function, signal energy will be concentrated in a small number of relatively large coefficients while ambient noise will be spread out, allowing coefficients to be thresholded.

Threshold selection and implementation are two factors which significantly impact wavelet denoising methods. Common methods include hard, soft and non-linear thresholding approaches. Hard thresholding sets all coefficient values beneath the threshold to zero, leaving the others unchanged (Jansen, 2001); soft thresholding additionally reduces all coefficients value to maintain continuity; while nonlinear thresholding typically enforces a smoothness constraint on the coefficient mapping function as well. Typical threshold selection methods include Universal thresholding and the Stein unbiased risk estimator (SURE) (Donoho, 1995), both implemented using soft thresholding.

Recently, the Ephraim Malah suppression rule (Ephraim and Malah, 1984) for speech enhancement has been applied to the wavelet domain as a more advanced time-varying

thresholding approach (Cohen, 2001). This method helps to reduce the "musical noise" artifacts caused by uniformly applied thresholds.

III. PROPOSED METHOD

The method introduced here is based on a modified Wavelet Packet Decomposition using an MMSE coefficient estimation for thresholding. The key element of the technique is the use of the Greenwood warping function to determine the WPT decomposition structure, based on a perceptually-motivated frequency axis.

Greenwood (1961) has shown that many land and aquatic mammals perceived frequency on a logarithmic scale along the cochlea, which corresponds to a non-uniform frequency resolution. This relationship can be modeled by the equation

$$A(10^{\alpha x} - k), \tag{7}$$

where α, A and k are species-specific constants and x is the cochlea position. Transformation between true frequency f and perceived frequency f_p can be obtained through the following equation pair:

$$F_p(f) = (1/\alpha)\log_{10}(f/A + k), \tag{8}$$

$$F_p^{-1}(f_p) = A(10^{\alpha f_p} - k). \tag{9}$$

The constants α, A and k can be found if frequency-cochlear position data is available. However, since cochlear information has never been measured for many species, an approximate solution is needed. Lepage (2003) has shown that k can be estimated as 0.88, based on both theoretical justification and experimental data acquired on a number of mammalian species. Assuming this value for k, α and A can be solved for given approximate hearing range $f_{\min} \sim f_{\max}$ of the species (Clemins, 2005; Clemins and Johnson, 2006; Clemins *et al.*, 2006):

$$A = \frac{f_{\min}}{1-k}, \tag{10}$$

$$\alpha = \log_{10}\left(\frac{f_{\max}}{A} + k\right). \tag{11}$$

Thus, a frequency warping function can be constructed by using the species specific values of $f_{\min}$ and $f_{\max}$.

A perceptually-motivated wavelet transform can be designed to mimic the auditory frequency scale by using decomposition critical bands. This implementation was originally proposed by Black for coding (Black and Zeytinoglu, 1995), and has been widely used for perceptual speech enhancement (Cohen, 2001; Fu and Wan, 2003; Shao and Chang, 2006). To generalize this technique to bioacoustic signal enhancement, we propose to decompose wavelet packet tree into the critical bands with respect to the species specific Greenwood frequency warping curve.

Fig. 2 shows an approximation of the Greenwood scale by critical-band WPD for three distinct species: ortolan bunting (*Emberiza hortulana*) downsampled to 20kHz, rhesus monkey (*Macaca mulatta*) downsampled to 20KHz, and the humpback whale (*Megaptera novaeanglia*) sampled at 4kHz. The corresponding decomposition trees are illustrated in Fig. 3. The perceptual WPD splits the frequency range corresponding to different species data into critical bands: ortolan bunting, 0Hz~10kHz, 36 critical bands; rhesus monkey, 0Hz~10kHz, 30 critical bands; humpback whale, 0Hz~2kHz, 31 critical bands. The bands are established automatically by optimally matching the subband center frequencies to the perceptual scale curve, in the mean error sense. For Greenwood scale calculation, the $f_{\min}$ and $f_{\max}$ used in (10) and (11) are 400Hz and 7200Hz for the ortolan bunting (Edward, 1943), 20Hz and 42000Hz for the rhesus monkey (Heffner, 2004) and 2Hz and 6000Hz for the humpback whale (Helweg, 2000).

Given this perceptual decomposition structure, an MMSE estimator for performing thresholding can be derived in the wavelet domain (Cohen, 2001; Cohen and Berdugo, 2001). Using an additive time-domain model, the resulting wavelet domain model is

$$Y_{l,n}(k) = X_{l,n}(k) + D_{l,n}(k), \tag{12}$$

where $Y_{l,n} =< y, \varphi_{l,n,k} >$, $X_{l,n}(k) =< x, \varphi_{l,n,k} >$, $D_{l,n}(k) =< d, \varphi_{l,n,k} >$, k is the index of the coefficients in each subband, l is the decomposition level, n is the node index, and $\varphi_{l,n,k}$ is the scaled and shifted mother wavelet. The notation $< x, \varphi >$ represents the wavelet transform of signal x using φ as the mother wavelet.

The optimally modified LSA estimator (Cohen and Berdugo, 2001) is used to perform wavelet denoising. Under this approach, the clean speech wavelet packet coefficients are estimated using a minimum mean square error (MMSE) criteria, under the assumptions that both speech and noise are complex Gaussian variables. Speech presence uncertainty is also incorporated, using the hypothesis testing framework given by:

$$H_0 = D_{l,n}(k) \tag{13}$$

$$H_1 = X_{l,n}(k) + D_{l,n}(k) \tag{14}$$

Under this framework, a parameter of signal presence uncertainty is calculated through the equation (Cohen and Berdugo, 2001)

$$p_{l,n}(k) = \left\{ 1 + \frac{1+\xi_{l,n}(k)}{q_{l,n}^{-1}(k) - 1} \exp(-\upsilon_{l,n}(k)/2) \right\}^{-1}, \tag{15}$$

where $\xi_{l,n}(k)$ is the *a priori* SNR, $\upsilon_{l,n}(k)$ is from equation (4), and $q_{l,n}(k)$ is the a priori probability for signal absence, which is estimated by

$$\hat{q}_{l,n}(k) = 1 - \begin{cases} \dfrac{\log(\xi_{l,n}(k)/\xi_{\min})}{\log(\xi_{\max}/\xi_{\min})} & if\ \xi_{\min} \le \xi_{l,n}(k) \le \xi_{\max} \\ 0 & if\ \xi_{l,n}(k) \le \xi_{\min} \\ 1 & otherwise \end{cases}, \tag{16}$$

where $\xi_{\min}$ and $\xi_{\max}$ are empirical constants, $\xi_{\min} = -10\text{dB}$, $\xi_{\max} = -5\text{dB}$.

An estimate for the clean speech, which minimizes the mean-square error, results in

$$\hat{X}_{l,n}(k) = \frac{\lambda_{l,n}(k) \cdot p_{l,n}(k)}{\lambda_{l,n}(k) + \sigma_{l,n}^2(k)} Y_{l,n}(k), \tag{17}$$

where the signal variance is given using the Decision-Directed method of Ephraim Malah:

$$\hat{\lambda}_{l,n}(k) = \alpha \left| \hat{X}_{l,n}(k-1) \right| + (1-\alpha) \max\{ \left| Y_{l,n}(k) \right| - \sigma_{l,n}(k), 0 \}. \tag{18}$$

IV. EXPERIMENTAL SETUP AND RESULTS

The proposed method and comparative baseline approaches were applied to ortolan bunting (*Emberiza hortulana*), rhesus monkey (*Macaca mulatta*) and humpback whale (*Megaptera novaeanglia*). Norwegian ortolan bunting vocalization data was collected from County Hedmark, Norway in May of 2001 and 2002 (Osiejuk *et al.*, 2003). Rhesus data was recorded on the island of Cayo Santiago, Puerto Rico by Joseph Solitis and John D. Newman (Li *et al.*, 2007). Humpback whale data (Payne and McVay, 1971) was provided by MobySound (Mellinger and Clark, 2006), a database for research in automatic recognition of marine animal calls. This data was collected in March 1994 off the north coast of the island of Kauai, Hawaii. Ten clean vocalizations from each species were segmented from the original recording data.

Both white noise and true environment noise were added to the clean data at SNR levels of -15, -10, -5, 0, +5, and +10 dB. The environment noise came from ambient noise regions of appropriate domain recordings for each species, spectrally flattened with a low order filter to preserve the basic noise characteristics while ensuring that the energy is spread through the entire

frequency band. For the rhesus monkey vocalizations, background noise was taken from a Vervet monkey data set (Seyfarth and Cheney, 2004). For the ortolan bunting vocalizations, background noise came directly from the dataset. For the humpback whale, marine noise was taken from a Beluga whale vocalization data set (Scheifele *et al.*, 2005), downsampled to 4000Hz.

Based on visual examination of the clean data from Fig. 4 – Fig. 6, tight passbands are chosen around the vocalizations. Selected ranges are 2600-5600Hz, 1000-10000Hz and 200-2000Hz for the ortolan bunting, rhesus monkey and humpback whale data, respectively. For the spectral subtraction, Wiener Filter, and Ephraim Malah filter approaches, the signal is divided into 32ms windows with 75% overlap between frames. This frame length was chosen empirically, as it is sufficiently long for good spectral estimation in each frame but not so long as to affect temporal change in the signals, and adjustments to this value cause only minor changes to the overall enhancement results. Frequency analysis is done using a Hanning window and noise estimation is accomplished using the first three frames of the signal. For wavelet analysis, the discrete Meyer wavelet is used as the mother wavelet, which was chosen to provide good separation of subbands due to their regularity property (Cohen, 2001). The decomposition was done as illustrated in Fig. 3. The forgetting factor α used in equations (5) and (18) is set to 0.98 for the EM filter and 0.92 for the Wavelet denoising.

Signal-to-noise ratio (SNR) and Segmental signal-to-noise ratio (SSNR) are used as objective measurement criteria for all sets of experiments. SSNR is computed by calculating the SNR on a frame-by-frame basis over the signal and averaging theses values. This permits the measure to assign equal weights to the loud and soft portions of the signal, which has been shown to have a

higher correlation with perceived quality in human speech evaluation (Deller *et al.*, 2000). The formulas for SNR and SSNR are

$$SNR = 10\log_{10}\frac{\sum_{n} x^2(n)}{\sum_{n}\left[x(n)-\hat{x}(n)\right]^2} \tag{19}$$

$$SSNR = \frac{1}{M}\sum_{j=0}^{M-1} 10\log_{10}\left[\sum_{n=Nj+1}^{N(j+1)}\frac{x^2(n)}{\left[x(n)-\hat{x}(n)\right]^2}\right] \tag{20}$$

where M is the number of frames, each of length N, and $x(n)$ and $\hat{x}(n)$ are the original and enhanced signals, respectively.

For visualization, spectrograms of the enhanced signals for the white noise and environment noise conditions at -10dB SNR can be seen in Fig. 4 through Fig. 6.

SNR and SSNR results for the white noise and environment noise are shown in Fig. 7 and Fig. 8. The SNR and SSNR values are given as amount of improvement over original input noisy values. Methods shown in these figures include band-pass filtering, spectral subtraction, Wiener filtering, Ephraim Malah filtering, the proposed Perceptual Wavelet Packet Transform (P–WPT) method, as well as a Uniform band Wavelet Packet Transform (U–WPT), which is identical to the proposed method except utilizing uniformly-spaced frequency bands rather than the perceptual scaling.

From reviewing the spectrograms and the SNR and SSNR plots, several conclusions can be drawn. It is clear that the proposed perceptual wavelet denoising method and the Ephraim Malah filtering method have the best overall performance in both the white noise and the environment noise conditions. The proposed method shows better enhancement performance for the higher noise (lower original SNR) cases in particular. Comparing the SNR improvement to the SSNR improvement in Fig. 7 and Fig. 8, it can be seen that the SSNR, which is generally considered to

be a more perceptually meaningful metric, shows greater superiority for the proposed method over the other methods than does SNR. Wiener filtering and spectral subtraction have moderate enhancement performance overall, while band-pass filtering results are a little sporadic, giving generally moderate results with good results in a few specific environment cases. Specifically, as expected, band-pass filtering works relatively well in the ortolan case where the vocalization frequency range is narrow and has limited overlap with the environment noise spectrum. Comparing the P-WPT and U-WPT results, it can be seen that the use of perceptual scale has little overall impact. In the white noise case, the SNR is slightly higher for the uniform scaling, and SSNR measures show little difference. For environmental noise, the SNR is again slightly higher for the uniform scaling, and SSNR is again similar, showing a slight benefit for the perceptual scaling in two of the three examples. Under the noisiest conditions, the two wavelet-based enhancement techniques significantly outperform all of the baseline methods.

One interesting thing to note is that each of the different enhancement methods has unique characteristics, as seen in the spectrograms of Fig. 4 through Fig. 6. Band-pass filtering has the expected look, keeping all noise in the target range and eliminating nearly everything out-of-band. Spectral subtraction shows some temporal streaking due to the fact that the noise spectrum being removed is fixed. Wiener filtering and Ephraim Malah filtering have similar looks, except with Ephraim Malah providing better overall results. The proposed method has the best noise removal, but can also be seen to possess an artifact (most noticeable in Fig. 5), seen as a faint reflection of the primary signal. This artifact, which is not audible and does not contain enough energy to significantly impact the SNR or SSNR metrics, illustrates some of the processing differences between a frequency domain approach such as Ephraim Malah and a wavelet domain approach such as the proposed method. Because the mother wavelet used for analysis is

somewhat broadband, each of the nodes in the decomposition trees shown in Fig. 3 contains more than a single frequency component. Thus the nodes that are given primary emphasis for reconstruction have energy at more than one frequency. However, since the nature of this wavelet representation is also more compact, coefficients not given primary emphasis can be more strongly thresholded, yielding less energy throughout the entire background frequency range, as can also be seen in the spectrograms. The selection of mother wavelet also impacts the degree of this artifact. The overall effect is that while the residual noise for the Ephraim Malah and perceptual wavelet approaches have similar total energy (with the perceptual wavelet having a little less in high noise situations), this residual noise in the Ephraim Malah approach is spread more evenly across the frequency range, while in the perceptual wavelet approach it is more concentrated.

V. CONCLUSIONS

Enhancement techniques taken from the field of speech processing have been generalized and applied to noise reduction of bioacoustic vocalizations. Four baseline methods, including spectral subtraction, Wiener filtering, and Ephraim Malah filtering, as well as simple band-pass filtering, were compared to a new technique based on perceptual wavelet decomposition. Results indicate improved performance of the new method, particularly for the most noisy conditions. The new approach can be easily applied to any species, requiring only upper and lower frequency limits for the species to create the appropriate Greenwood function frequency warping curve.

ACKNOWLEDGMENTS

This material is based on work supported by National Science Foundation under Grant No. IIS-0326395. We also want to express our thanks to Joseph Solitis and John D. Newman for providing the rhesus monkey vocalizations; T. S. Osiejuk for providing ortolan bunting vocalizations, and Mobysound for providing humpback whale vocalizations.

REFERENCES

Álvarez, B. D., and García, C. F. (**2004**). "System architecture for pattern recognition in eco systems," in Proc. ESA Special Publication No. 553, Madrid, Spain.

Black, M., and Zeytinoglu, M. (**1995**). "Computationally Efficient Wavelet Packet Coding of Wide-band Stereo Audio Signals," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., Detroit, MI, **5**, 3075-3078.

Boll, S. F. (**1979**). "Suppression of acoustic noise in speech using spectral subtraction," IEEE Trans. on Acoust., Speech, Signal Process. **ASSP-27**(2), 113-120.

Clemins, P., and Johnson, M. T. (**2006**). "Generalized perceptual linear prediction (gPLP) features for animal vocalization analysis," J. Acoust. Soc. Am. **120**(1), 527-534.

Clemins, P. J. (**2005**). "(Doctoral Dissertation) Automatic Speaker Identification and Classification of Animal Vocalizations," Marquette University.

Clemins, P. J., Trawicki, M. B., Adi, K., Tao, J., and Johnson, M. T. (**2006**). "Generalized perceptual feature for vocalization analysis across multiple species," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., Paris, France, **1**, 253 - 256.

Cohen, I. (**2001**). "Enhancement of speech using bark-scaled wavelet packet decomposition," in Proc. Eurospeech, Aalborg, Denmark, 1933-1936.

Cohen, I., and Berdugo, B. (**2001**). "Speech enhancement for non-stationary noise environments," Signal Processing **81**(11), 2403-2418.

Deller, J. R., Hansen, J. H. L., and Proakis, J. G. (**2000**). "Chap. 9: Speech Quality Assessment," in *Discrete-Time Processing of Speech Signals* (IEEE Press, Piscataway, NJ), pp. 584-587.

Donoho, D. L. (**1995**). "De-noising by soft-thesholding," IEEE Trans. Info. Theory, **41**(3), 613-627.

Edward, E. P. (**1943**). "Hearing ranges of four species of birds," The Auk **60**, 239-241.

Ephraim, Y., and Malah, D. (**1984**). "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," IEEE Trans. on Acoust., Speech, Signal Process. **ASSP_32**(6), 1109-1121.

Ephraim, Y., and Malah, D. (**1985**). "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," IEEE Trans. on Acoust., Speech, Signal Process. **ASSP-33**(2), 443-445.

Ephraim, Y., and Trees, H. L. V. (**1995**). "A signal subspace approach for speech enhancement," IEEE Trans. Speech Audio Process. **3**(4), 251-266.

Fu, Q., and Wan, E. A. (**2003**). "Perceptual wavelet adaptive denoising of speech," in Proc. EuroSpeech, Geneva, Switzerland, 577-580.

Greenwood, D. D. (**1961**). "Critical bandwidth and the frequency coordinates of the basilar membrane," J. Acoust. Soc. Am. **33**(10), 1344-1356.

Gur, B. M., and Niezrecki, C. (**2007**). "Autocorrelation based denoising of manatee vocalizations using the undecimated discrete wavelet transform," J. Acoust. Soc. Am. **122**(1), 188-199.

Heffner, R. S. (**2004**). "Primate hearing from a mammalian perspective," The anatomical record (part A) **281A**, 1111-1122.

Helweg, D. A. (**2000**). "An integrated approach to the creation of a humpback whale hearing model," in *Technical report 1835* (San Diego).

Jansen, M. (**2001**). *Noise reduction by wavelet thresholding* (Springer Verlag).

Johnson, M. T., Yuan, X., and Ren, Y. (**2007**). "Speech signal enhancement through adaptive wavelet thresholding," Speech Communication **49**(2), 123-133.

Lepage, E. L. (**2003**). "The mammalian cochlear map is optimally warped," J. Acoust. Soc. Am. **114**(2), 896-906.

Li, X., Tao, J., Johnson, M. T., Soltis, J., Savage, A., Leong, K. M., and Newman, J. D. (**2007**). "Stress and emotion classification using jitter and shimmer features," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., Honolulu, Hawaii, **IV**, 1081-1084.

Lim, J., and Oppenheim, A. V. (**1978**). "All-pole modeling of degraded speech," IEEE Trans. on Acoust., Speech, Signal Process. **26**(3), 197-210.

Liu, R. C., Miller, K. D., Merzenich, M. N., and Schreiner, C. E. (**2003**). "Acoustic variability and distinguishability among mouse ultrasound vocalizations," J. Acoust. Soc. Am. **114**(6), 3412-3422.

Mellinger, D. K. (**2002**). "Ishmael 1.0 User's Guide," (Pacific Marine Enviromental Laboratory, Seattle).

Mellinger, D. K., and Clark, C. W. (**2006**). "MobySound: A reference achive for studying automatic recognition of marine mammal sounds," Applied Acoustics **67**(11-12), 1226-1242.

Osiejuk, T. S., Ratynska, K., Cygan, J. P., and Svein, D. (**2003**). "Song structure and repertoire variation in ortolan bunting (Emberiza hortulana L.) from isolated Norwegian population," Annales Zoologici Fennici **40**, 3-19.

Payne, R. S., and McVay, S. (**1971**). "Songs of Humpback Whales," Science **173**(3397), 585-597.

Scheifele, P. M., Andrew, S., Cooper, R. A., and Darre, M. (**2005**). "Indication of a Lombard vocal response in the St. Lawrence River beluga," J. Acoust. Soc. Am. **117**(3), 1486-1492.

Seyfarth, R. M., and Cheney, D. L. (**2004**). "TalkBank Ethology Data: Field Recordings of Vervet Monkey Calls," (Linguistic Data Consortium, Philadelphia).

Shao, Y., and Chang, C-H. (**2006**). "A generalized perceptual time-frequency subtraction method for speech enhancement," in Proc. ISCAS 2006, 2537-2540.

Yan, Z., Niezrecki, C., and Beusse, D. O. (**2005**). "Background noise cancellation for improved acoustic detection of manatee vocalizations," J. Acoust. Soc. Am. **117**(6), 3566-3573.

Yan, Z., Niezrecki, C., L. N. III, C., and Beusse, O. D. (**2006**). "Background noise cancellation of manatee vocalizations using an adaptive line enhancer," J. Acoust. Soc. Am. **120**(1), 145-152.

FIGURE CAPTIONS

FIG. 1. (a) Discrete wavelet transform. (b) Wavelet packet decomposition tree.

FIG. 2. Center frequencies of Greenwood Scale (solid line) and WPD critical bands. (a) Ortolan bunting. (b) Rhesus monkey. (c) Humpback whale.

FIG. 3. Perceptual wavelet decomposition tree. (a) Ortolan bunting. (b) Rhesus monkey. (c) Humpback whale.

FIG. 4. Spectrograms of ortolan bunting signals: Clean signal, -10dB SNR noisy signals and signals enhanced by band-pass filtering, spectral subtraction, Wiener filtering, Ephraim Malah log-MMSE filtering and perceptual WPT filtering (left column is for white noise, right for environment noise).

FIG. 5. Spectrograms of rhesus monkey signals: Clean signal, -10dB SNR noisy signals and signals enhanced by band-pass filtering, spectral subtraction, Wiener filtering, Ephraim Malah log-MMSE filtering and perceptual WPT filtering (left column is for white noise, right for environment noise).

FIG. 6. Spectrograms of humpback whale: Clean signal, -10dB SNR noisy signals and signals enhanced by band-pass filtering, spectral subtraction, Wiener filtering, Ephraim Malah log-MMSE filtering and perceptual WPT Filtering (left column is for white noise, right for environment noise).

FIG. 7. SNR and SSNR results for white noise at -15, -10, -5,0, +5, +10 dB SNR levels.

FIG. 8. SNR and SSNR results for environment noise at -15, -10, -5,0, +5, +10 dB SNR levels.

FIG. 1

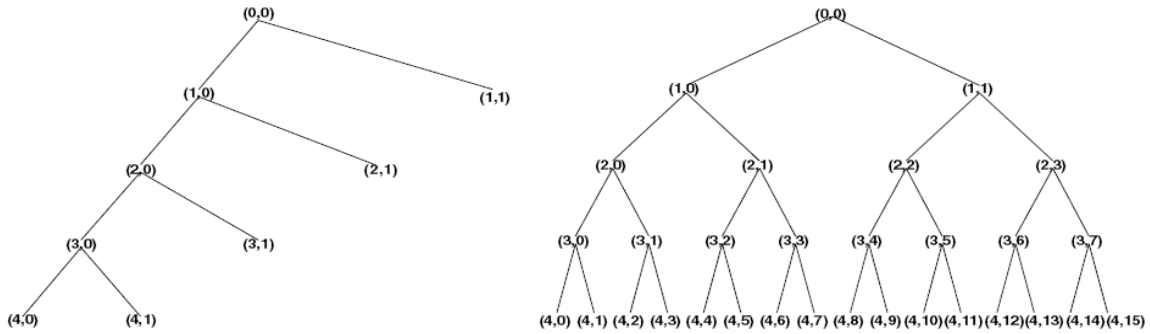

(0,0)
(1,0)
(1,1)
(2,0)
(2,1)
(3,0)
(3,1)
(4,0)
(4,1)
(a)
(0,0)
(1,0)
(1,1)
(2,0)
(2,1)
(2,2)
(2,3)
(3,0)
(3,1)
(3,2)
(3,3)
(3,4)
(3,5)
(3,6)
(3,7)
(4,0) (4,1) (4,2) (4,3) (4,4) (4,5) (4,6) (4,7) (4,8) (4,9) (4,10) (4,11) (4,12) (4,13) (4,14) (4,15)
(b)

FIG. 2

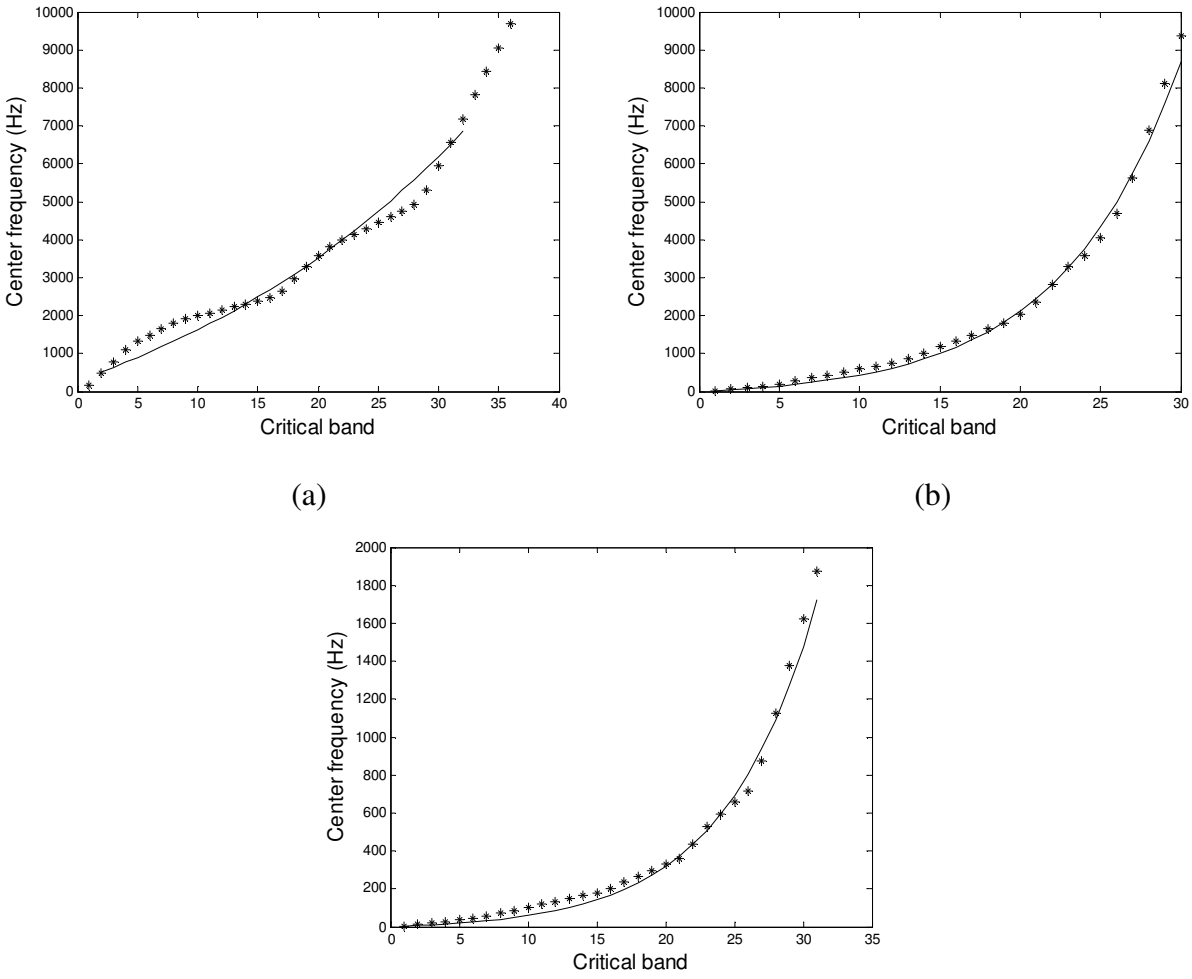
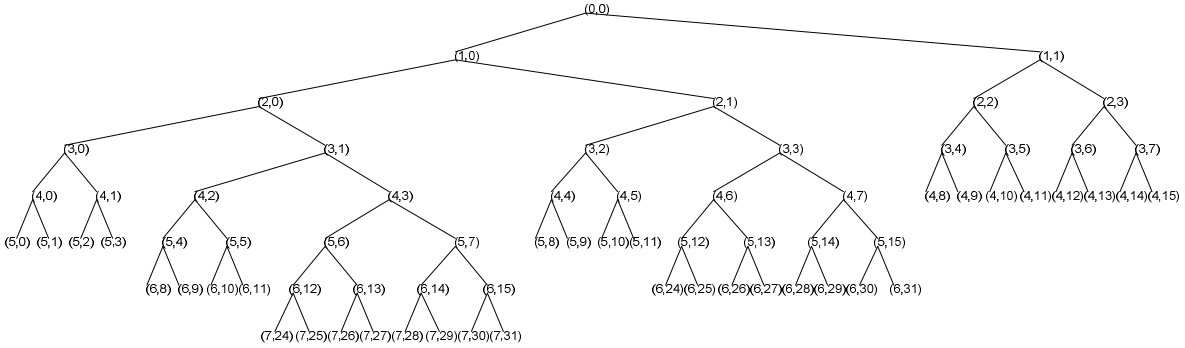
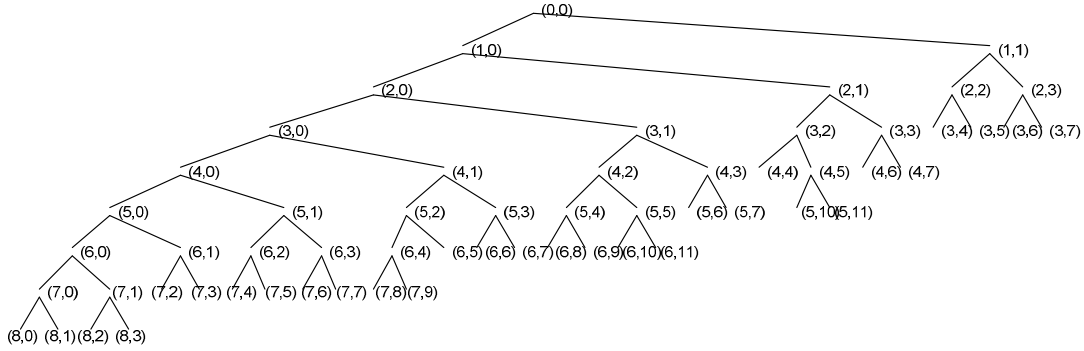

Center frequency (Hz)
Critical band
(a)
Center frequency (Hz)
Critical band
(b)
Center frequency (Hz)
Critical band
(c)

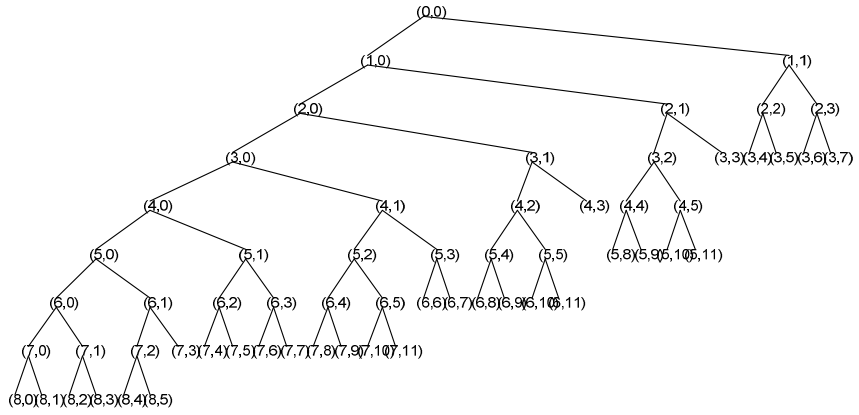
FIG. 3


(0,0)
(1,0)
(1,1)
(2,0)
(2,1)
(2,2)
(2,3)
(3,0)
(3,1)
(3,2)
(3,3)
(3,4)
(3,5)
(3,6)
(3,7)
(4,0)
(4,1)
(4,2)
(4,3)
(4,4)
(4,5)
(4,6)
(4,7)
(4,8)
(4,9)
(4,10)
(4,11)
(4,12)
(4,13)
(4,14)
(4,15)
(5,0)
(5,1)
(5,2)
(5,3)
(5,4)
(5,5)
(5,6)
(5,7)
(5,8)
(5,9)
(5,10)
(5,11)
(5,12)
(5,13)
(5,14)
(5,15)
(6,8)
(6,9)
(6,10)
(6,11)
(6,12)
(6,13)
(6,14)
(6,15)
(6,24)
(6,25)
(6,26)
(6,27)
(6,28)
(6,29)
(6,30)
(6,31)
(7,24)
(7,25)
(7,26)
(7,27)
(7,28)
(7,29)
(7,30)
(7,31)

(a)


(0,0)
(1,0)
(1,1)
(2,0)
(2,1)
(2,2)
(2,3)
(3,0)
(3,1)
(3,2)
(3,3)
(3,4)
(3,5)
(3,6)
(3,7)
(4,0)
(4,1)
(4,2)
(4,3)
(4,4)
(4,5)
(4,6)
(4,7)
(5,0)
(5,1)
(5,2)
(5,3)
(5,4)
(5,5)
(5,6)
(5,7)
(5,10)
(5,11)
(6,0)
(6,1)
(6,2)
(6,3)
(6,4)
(6,5)
(6,6)
(6,7)
(6,8)
(6,9)
(6,10)
(6,11)
(7,0)
(7,1)
(7,2)
(7,3)
(7,4)
(7,5)
(7,6)
(7,7)
(7,8)
(7,9)
(8,0)
(8,1)
(8,2)
(8,3)

(b)


(0,0)
(1,0)
(1,1)
(2,0)
(2,1)
(2,2)
(2,3)
(3,0)
(3,1)
(3,2)
(3,3)(3,4)(3,5)(3,6)(3,7)
(4,0)
(4,1)
(4,2)
(4,3)
(4,4)
(4,5)
(5,0)
(5,1)
(5,2)
(5,3)
(5,4)
(5,5)
(5,8)(5,9)
(6,0)
(6,1)
(6,2)
(6,3)
(6,4)
(6,5)
(6,6)(6,7)(6,8)(6,9)
(7,0)
(7,1)
(7,2)
(7,3)(7,4)(7,5)(7,6)(7,7)(7,8)(7,9)(7,10)(7,11)
(8,0)(8,1)(8,2)(8,3)(8,4)(8,5)

(c)

FIG. 4

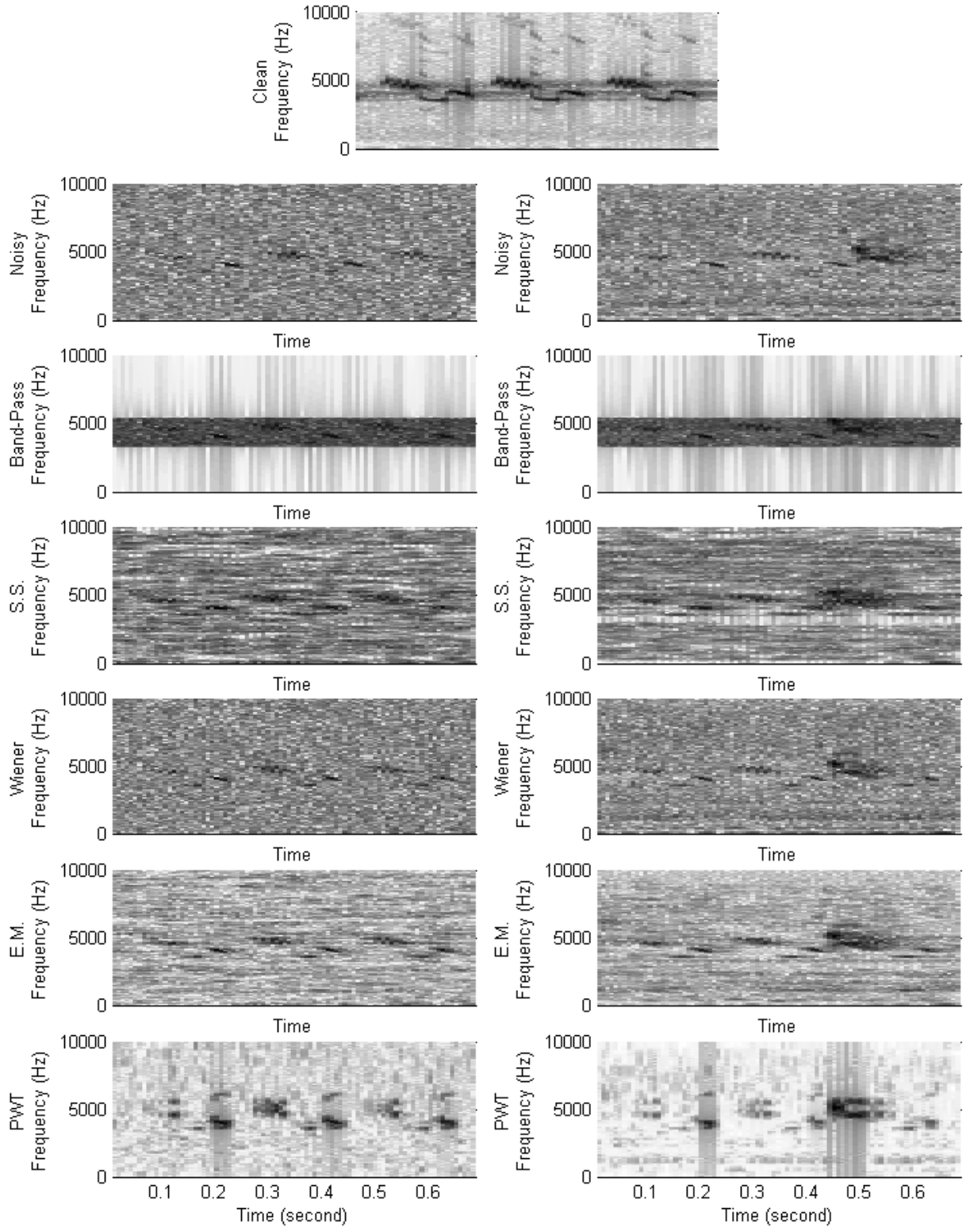

Clean
Frequency (Hz)
10000
5000
0
Noisy
Band-Pass
S.S.
Wiener
E.M.
PWT
Time
0.1
0.2
0.3
0.4
0.5
0.6
Time (second)

FIG. 5

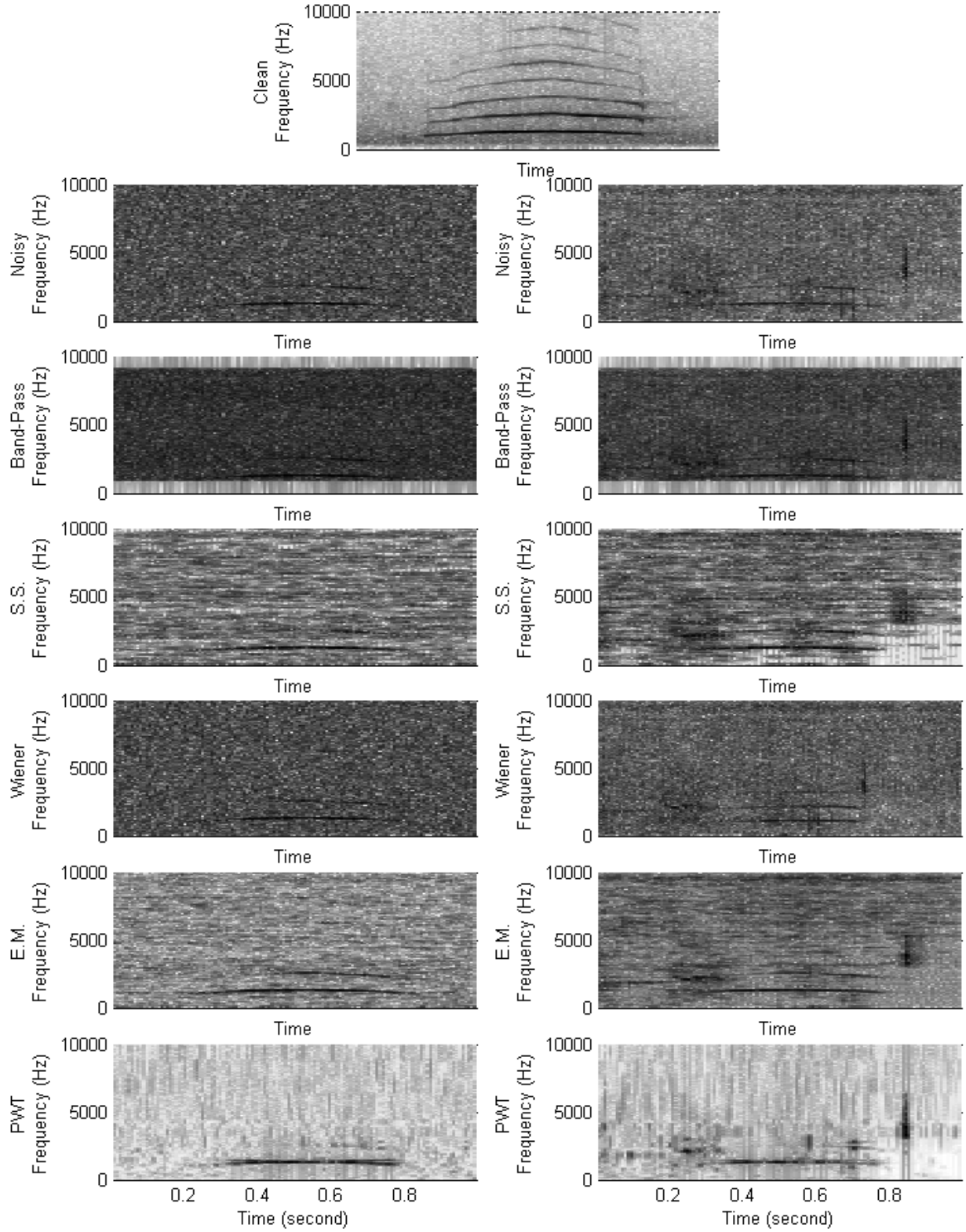

Clean
Frequency (Hz)
10000
5000
0
Time
Noisy
Band-Pass
S.S.
Wiener
E.M.
PWT
0.2
0.4
0.6
0.8
Time (second)

FIG. 6

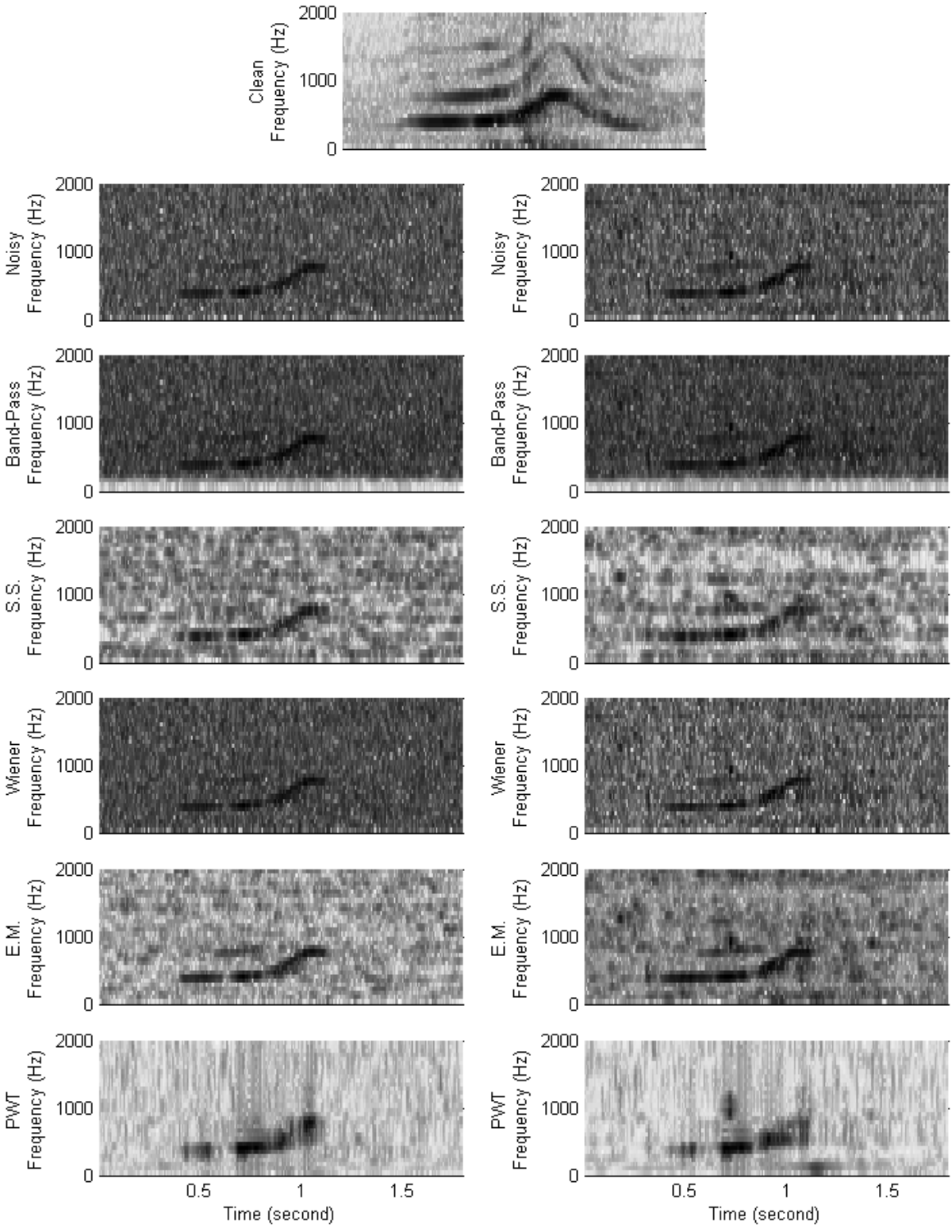

Clean
Noisy
Band-Pass
S.S.
Wiener
E.M.
PWT
Frequency (Hz)
2000
1000
0
0.5
1
1.5
Time (second)

FIG. 7

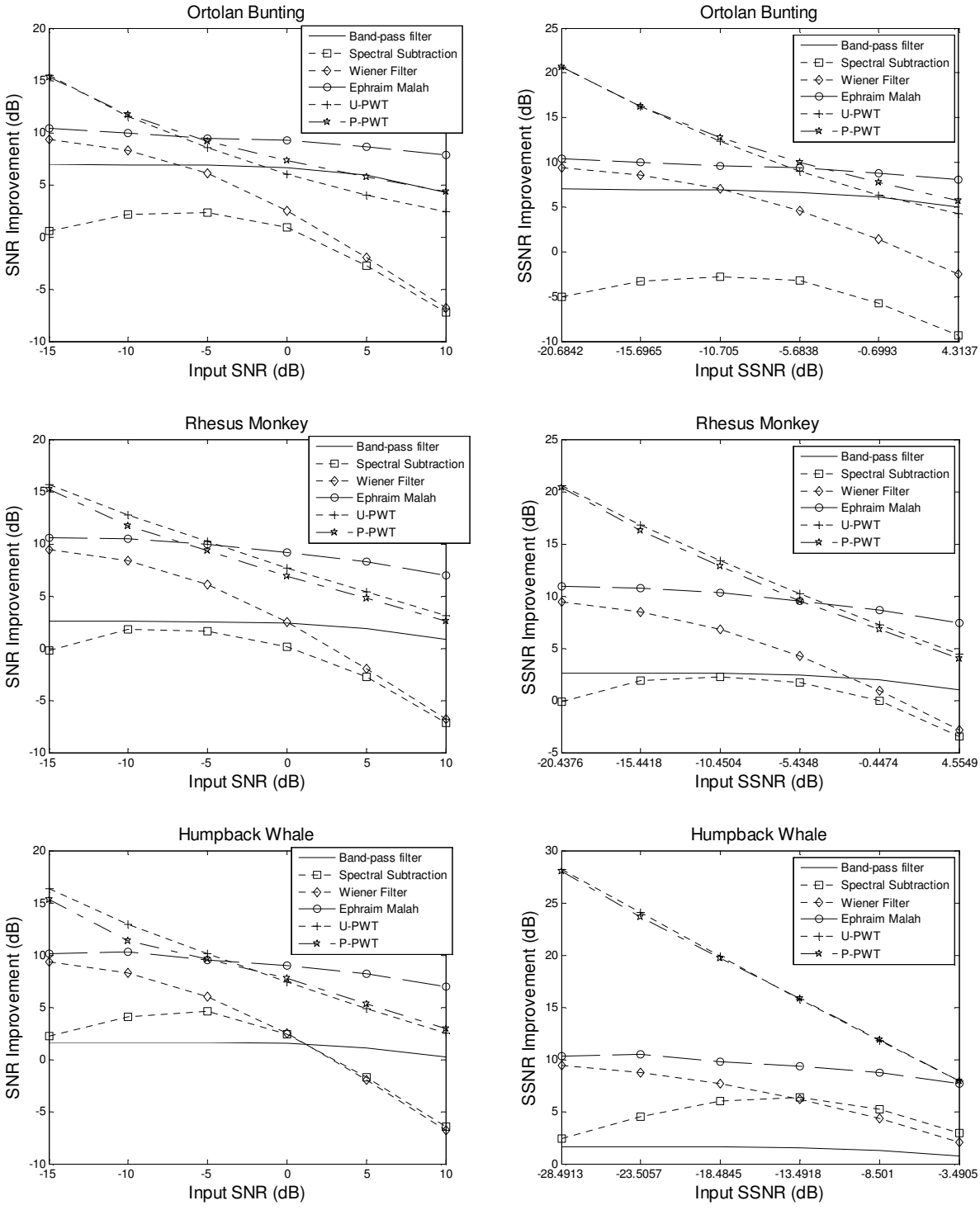
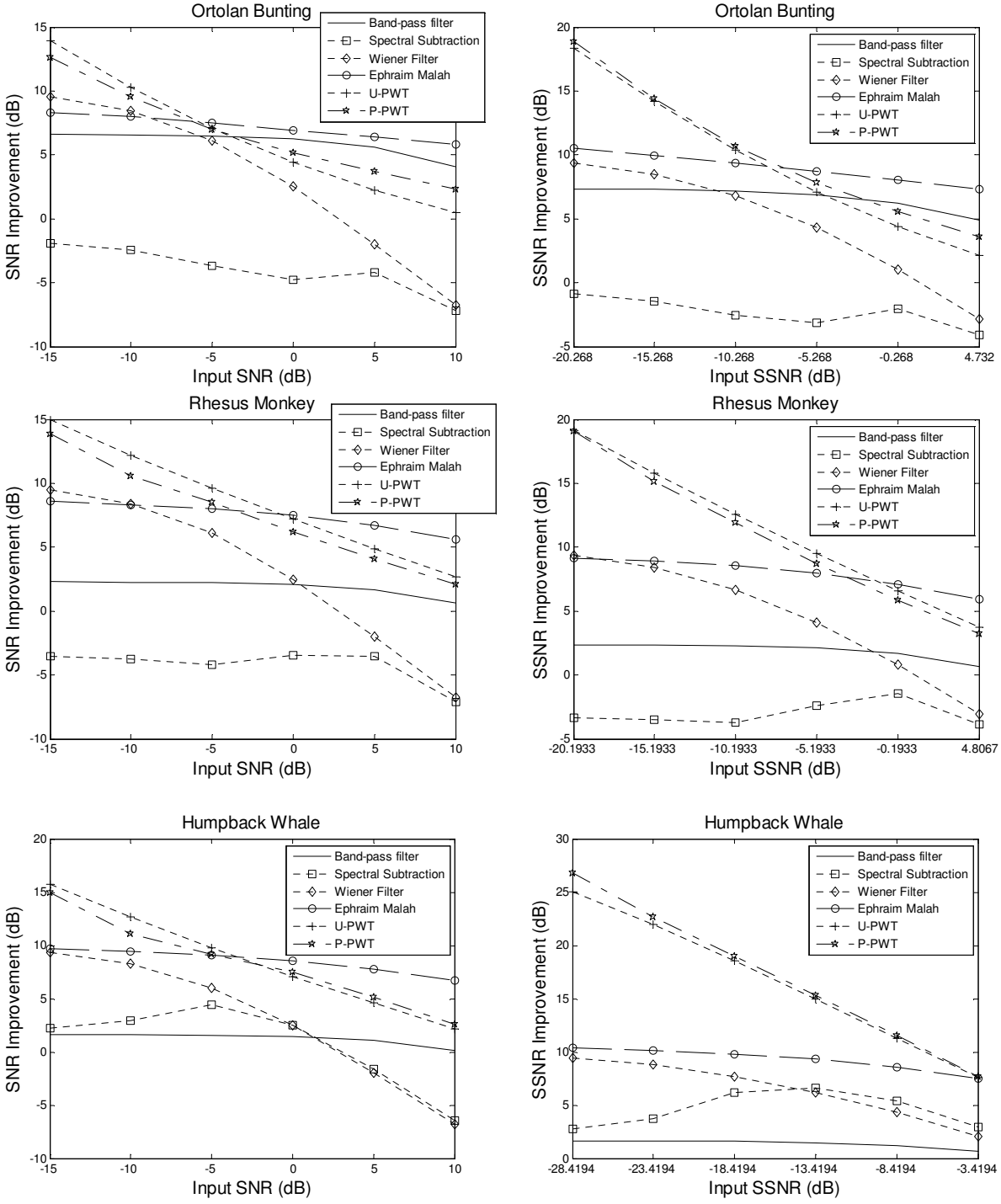

Ortolan Bunting
SNR Improvement (dB)
Input SNR (dB)
Band-pass filter
Spectral Subtraction
Wiener Filter
Ephraim Malah
U-PWT
P-PWT
Ortolan Bunting
SSNR Improvement (dB)
Input SSNR (dB)
-20.6842
-15.6965
-10.705
-5.6838
-0.6993
4.3137
Rhesus Monkey
SNR Improvement (dB)
Input SNR (dB)
Rhesus Monkey
SSNR Improvement (dB)
Input SSNR (dB)
-20.4376
-15.4418
-10.4504
-5.4348
-0.4474
4.5549
Humpback Whale
SNR Improvement (dB)
Input SNR (dB)
Humpback Whale
SSNR Improvement (dB)
Input SSNR (dB)
-28.4913
-23.5057
-18.4845
-13.4918
-8.501
-3.4905

FIG. 8


Ortolan Bunting
SNR Improvement (dB)
Input SNR (dB)
Band-pass filter
Spectral Subtraction
Wiener Filter
Ephraim Malah
U-PWT
P-PWT
Ortolan Bunting
SSNR Improvement (dB)
Input SSNR (dB)
-20.268
-15.268
-10.268
-5.268
-0.268
4.732
Rhesus Monkey
SNR Improvement (dB)
Input SNR (dB)
Rhesus Monkey
SSNR Improvement (dB)
Input SSNR (dB)
-20.1933
-15.1933
-10.1933
-5.1933
-0.1933
4.8067
Humpback Whale
SNR Improvement (dB)
Input SNR (dB)
Humpback Whale
SSNR Improvement (dB)
Input SSNR (dB)
-28.4194
-23.4194
-18.4194
-13.4194
-8.4194
-3.4194