Generalized perceptual linear prediction features for animal vocalization analysis

Patrick J. Clemins[a] and Michael T. Johnson
Speech and Signal Processing Laboratory, Marquette University, P.O. Box 1881, Milwaukee, Wisconsin 53233-1881

(Received 30 June 2005; revised 31 March 2006; accepted 18 April 2006)

A new feature extraction model, generalized perceptual linear prediction (gPLP), is developed to calculate a set of perceptually relevant features for digital signal analysis of animal vocalizations. The gPLP model is a generalized adaptation of the perceptual linear prediction model, popular in human speech processing, which incorporates perceptual information such as frequency warping and equal loudness normalization into the feature extraction process. Since such perceptual information is available for a number of animal species, this new approach integrates that information into a generalized model to extract perceptually relevant features for a particular species. To illustrate, qualitative and quantitative comparisons are made between the species-specific model, generalized perceptual linear prediction (gPLP), and the original PLP model using a set of vocalizations collected from captive African elephants (*Loxodonta africana*) and wild beluga whales (*Delphinapterus leucas*). The models that incorporate perceptional information outperform the original human-based models in both visualization and classification tasks. © *2006 Acoustical Society of America.* [DOI: 10.1121/1.2203596]

PACS number(s): 43.80.Lb, 43.66.Gf [WWA] Pages: 527–534

I. INTRODUCTION

One of the primary tasks when analyzing animal vocalizations is determining and measuring acoustically relevant features. Currently, many features used in bioacoustic analysis are based on the entire vocalization, often extracted by hand from spectrogram plots (Fristrup and Watkins, 1992; Leong *et al.*, 2002; Owren *et al.*, 1997; Riede and Zuberbühler, 2003; Sjare and Smith, 1986). Some of the features commonly used for analysis include duration, fundamental frequency measures, amplitude information, and spectral information such as Fourier transform coefficients. These traditional features are unable to capture temporally fine details of vocalizations because each feature has only one value for the entire vocalization. In addition, these features are often susceptible to researcher bias because the features are determined interactively. An alternative to this feature extraction paradigm is to divide signals into frames and extract features automatically on a frame basis. This generates a feature matrix for each vocalization that captures information about how the vocalization changes over time. Another limitation of traditional features, either global or frame based, is that they typically do not use information about the perceptual abilities of the species under study explicitly in the feature extraction process.

The generalized perceptual linear prediction (gPLP) model introduced here is a frame-based feature extraction model that uses perceptual information about the species under study to calculate features that are relevant to that species. The gPLP model is applicable to different species by incorporating experimental data from available perceptual tests. Furthermore, the gPLP model can significantly decrease the time spent analyzing vocalizations and generates features with finer temporal resolution that are largely uncorrelated and not subject to researcher bias.

The gPLP feature extraction model generates features based on the source filter model of speech production. Although this model was originally developed for human speech processing, it has been shown to be applicable to the vocalizations of terrestrial mammals for the purposes of describing vocal production mechanisms (Fitch, 2003). The source excitation, modeled as a pulse train for voiced sound or white noise for unvoiced sound, is produced by physiology such as the glottis in land mammals, the tympaniform membrane in birds, or air sacs in marine animals. This excitation then propagates through a filter consisting of the vocal tract and nasal cavity in terrestrial animals or the body cavity and melon in marine animals.

The gPLP model presented here is designed to suppress excitation information and quantify the vocal tract filter characteristics of the vocalizations. Excitation information includes the fundamental frequency contour, while vocal tract characteristics are represented by formant information. Vocal tract features carry the majority of the information in human speech, but there are a number of languages in which the fundamental frequency contour discriminates between units of speech with similar vocal tract characteristics. There is reason to believe that excitation information is also important to the discrimination of animal vocalizations. In fact, many studies have used fundamental frequency measures in order to classify vocalizations (Buck and Tyack, 1993; Darden *et al.*, 2003). Excitation information such as fundamental frequency measures can be added to the gPLP feature vector to include excitation information.

[a]Electronic-mail: patrick.clemins@marquette.edu

The gPLP feature extraction model generates features in the discrete cepstral domain. The discrete cepstral domain is defined as

$$c[n] = F^{-1}\{\log[F(s[n])]\}, \quad (1)$$

where F is the discrete Fourier transform and $s[n]$ is the original sampled time domain signal. This domain is preferred for speech processing systems because the general shape of the spectrum is accurately described by the first few cepstral coefficients, yielding an efficient signal representation. The cepstral domain is particularly appropriate for source filter model analysis because the logarithm operation effectively separates the excitation from the vocal tract filter (Deller *et al.*, 1993, p. 355). Finally, because cepstral values tend to be relatively uncorrelated with each other because of their orthonormal set of basis functions (Deller *et al.*, 1993, p. 377), the coefficients are good for statistical analysis methods.

The following section of this paper will describe the gPLP model in detail. Examples of the use of the gPLP model in vocalization analysis follow. Visualization, vocalization classification, and statistical testing tasks will be presented.

II. METHODS

A. Generalized perceptual linear prediction (gPLP)

The gPLP model is based on the perceptual linear prediction (PLP) model developed by Hermansky (1990). The goal of the original PLP model is to describe the psychophysics of human hearing more accurately in the feature extraction process. The gPLP model incorporates frequency warping to account for nonlinear frequency perception along the basilar membrane, critical bandwidth analysis to model frequency masking, equal-loudness normalization using audiogram information, and intensity-loudness power normalization. A block diagram of the gPLP method is shown in Fig. 1. The gPLP model includes the same components as the PLP, but incorporates experimentally acquired perceptual information as shown in Fig. 1 to tailor the feature extraction process to the species under study. The components designated by dotted boxes indicate where species-specific perceptual information is incorporated into the model. The various components of the model are discussed in detail in the following sections.

1. Preprocessing

The vocalization is first filtered using a preemphasis filter of the form

$$s'[n] = s[n] - ks[n-1], \quad (2)$$

where k is typically chosen to be between 0.95 and 0.99. This preemphasis filter gives greater weight to higher frequencies to emphasize the higher frequency formants and reduce spectral tilt (Deller *et al.*, 1993, p. 330). It also reduces the dynamic range of the spectrum so that the spectrum is more easily approximated by the autoregressive modeling component.

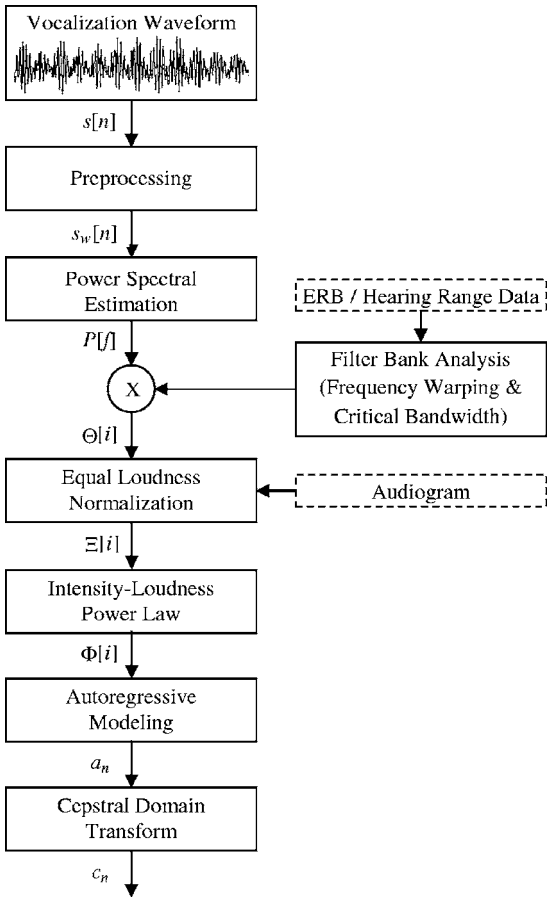

FIG. 1. PLP feature extraction block diagram. This original wave form is filtered and windowed in the preprocessing component. The power spectrum is then estimated for each frame of the vocalization. The power spectrum is convolved with a number of filters to generate filterbank energies which effectively smoothes and down samples the power spectrum. The filterbank energies are multiplied by the equal-loudness curve and cube-root compressed to account for the physiology of the ear. The down-sampled, normalized power spectrum is modeled by a set of autoregressive coefficients which are then converted to cepstral coefficients to take advantage of the cepstral domain.

The vocalization is then broken into frames and windowed using the Hamming window function (Oppenheim *et al.* 1999, p. 465). The frame size is usually chosen to include several fundamental frequency peaks, which is typically about 30 ms for human speech but may vary for other species' vocalizations. The vocalization is broken into frames so that the spectral estimation can be performed on quasistationary segments of the signal to ensure the precision of the spectral estimation. More information about the effects of windowing can be found in Oppenheim *et al.* (1999, p. 465).

2. Power spectral estimation

Once the signal is divided into windowed frames, the power spectrum is estimated. The discrete fast Fourier transform is used to estimate the power spectrum in this work, but other spectral estimation methods could also be used (Stoica and Moses, 1997). The discrete-time power spectrum $P[f]$ is estimated using

$$P[f] \approx abs\{F(s_w[n])\}^2, \tag{3}$$

where F is the discrete Fourier transform and $s_w[n]$ is the wth windowed frame of the signal.

3. Filter bank analysis

The next few components of the gPLP model transform the power spectrum to take into account various psychoacoustic phenomena. The filter bank analysis component accounts for two such phenomena, frequency masking and the nonlinear mapping between cochlear position and frequency sensitivity. Greenwood (1961) found that the cochlear-frequency map could be described logarithmically in many animal species with the equation

$$f = A(10^{ax} - k), \tag{4}$$

where f is frequency in Hz, x is the position on the basilar membrane that perceives that frequency, and A, a, and k are species-specific constants. Functions to convert between real frequency f and perceived frequency f_p can be created by replacing the basilar membrane position variable with perceived linear frequency as follows:

$$F_p(f) = (1/a)\log_{10}(f/A + k), \text{ and} \tag{5}$$

$$F_p^{-1}(f_p) = A(10^{af_p} - k), \tag{6}$$

where $F_p(f)$ converts from real frequency to perceived frequency and $F_p^{-1}(f_p)$ converts from perceived frequency to real frequency. The Mel-frequency scale, commonly used in speech processing, is a specific implementation of this warping function, using constant values of $A=700, a=1/2595 = 3.85\times10^{-4}$, and $k=1$. Values of A, a, and k can be determined for various species by fitting Eq. (6) to frequency-position data (Greenwood, 1990).

In cases where frequency-position data is not available, there are two other ways to acquire values for the constants. The first and most accurate method is to use equal-rectangular bandwidth (ERB) data (Zwicker and Terhardt, 1980). If the ERB data is fit by an equation of the form

$$\text{ERB} = \alpha(\beta f + \delta), \tag{7}$$

then the appropriate values of A, a, and k can be determined using the equations

$$A = \frac{1}{\beta},$$

$$a = \alpha\beta \log(e), \text{ and}$$

$$k = \delta. \tag{8}$$

where e is Euler's constant, the natural logarithm base. These equations are derived by taking the integral of the reciprocal of Eq. (7). The derivation of these equations is in the Appendix.

An alternative method for determining appropriate values for the constants requires an estimate of the hearing range of the species ($f_{\min}$ and $f_{\max}$). LePage (2003) noted that most mammals have a value of k near 0.88 and showed that this value is an optimal value when the tradeoffs between high frequency resolution, loss of low frequency resolution, minimization of map nonuniformity, and map smoothness are considered (LePage did not include non-mammalian species in the analysis, therefore using 0.88 as the value for k for those species may not be appropriate). Using the assumption that $k=0.88$, values for A and a can be determined using the equations

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

$$a = \log_{10}\left(\frac{f_{\max}}{A} + k\right). \tag{9}$$

If this method is used, the lower bound of the filter bank must be greater than $f_{\min}$, otherwise negative values of f_p result.

The second psychoacoustic phenomenon the filter bank takes into account is frequency masking. The original PLP model (Hermansky, 1990) constructed the filter bank using filters, Ψ_i, shaped like the critical band masking filters described by Fletcher (1940). These exponential-shaped masking filters are based on human sound perception and are computationally complex. Because of this complexity, the gPLP model implemented in this work uses triangular-shaped filters to approximate the critical band masking curve. Triangular-shaped filters can be described by the equation

$$\Psi_i[f] = 1 - \left|\left(\frac{2}{f_H - f_L}\right)f - \left(\frac{f_H + f_L}{f_H - f_L}\right)\right|, \tag{10}$$

where f_L and f_H are the low and high cutoff frequencies of each filter. This approximation is common in human speech processing feature extraction models (Davis and Mermelstein, 1980). Another reason for using a simple filter shape is that there is little data on the auditory filter shapes of animals other than humans, so more complex filter shapes are not necessarily more accurate.

The number of filters contained in the filter bank should be determined so that the bandwidth of each filter approximates the critical bandwidth of each species. However, because of the limitations on the resolution of the Fourier spectral estimate, this is not always possible. The lower frequency filters in the filter bank can become very narrow due to the Greenwood frequency warping. If too many filters are specified for the filter bank, the low frequency filters become narrow enough that they do not contain any points, or frequency bins, of the spectral estimate. The maximum number of filters the filter bank can contain before some filters contain no spectral points is a function of window size and the range of the filter bank (Clemins *et al.*, 2005).

As an example of the incorporation of perceptual information into filter bank design, the filter bank for the Indian elephant, is shown in Fig. 2. Perceptual data from Heffner and Heffner (1982) is used to determine the Greenwood equation constants. The equal loudness curve, discussed below, is applied to the filter bank in the figure which results in the variable height of the individual filters. Using the filter bank, filter energies $\Theta[i]$ are calculated with

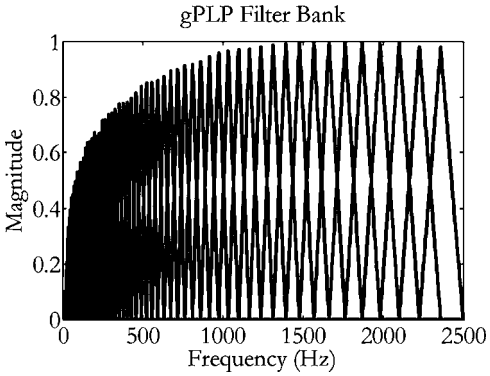

FIG. 2. Perceptual filterbank for an Indian elephant. The filters are logarithmically spaced according to the Greenwood cochlear map function. The constants A, a, and k are computed assuming the optimal $k=0.88$ for mammals as calculated by LePage (2003) and the approximate range of hearing for an Indian elephant (10–10 000 Hz). The equal loudness curve has been applied to the filter bank magnitudes to show its effect. Perceptual data is from Heffner and Heffner (1982).

$$\Theta[i] = \sum_{f=f_L}^{f_H} P[f]\Psi_i[f], \tag{11}$$

where $P[f]$ is the power spectrum, and f_L and f_H are the low and high cutoff frequencies of each filter $\Psi_i[f]$.

4. Equal loudness normalization

Once the filter bank energies are calculated, an equal-loudness curve is used to normalize the filter bank energies. Hermansky (1990) originally used a filter transfer function based on human sensitivity at about the 40-dB level adopted from Makhoul and Cosell (1976). For other species, an equal-loudness curve $E[f]$ can be approximated from the audiogram $A[f]$ of a species, which is much more widely available, using

$$E[f] = T - A[f], \tag{12}$$

where T is 60 dB for species which acquire sound through the air, and 120 dB for species that acquire sound in water. The different values are a result of the different propagation properties of sound waves and different reference dB pressures in the different mediums (Ketten, 1998). A polynomial curve is then fitted to $E[\log(f)]$ in order to interpolate for the frequency values sampled in $\Theta[i]$. A fourth-order curve has been found to adequately model most equal-loudness curves when $\log(f)$ is used. The constraint that $E(f)$ is always positive is maintained by setting all negative values of $E(f)$ to zero. The equal loudness curve is applied by multiplying the filter bank energies by the fitted curve using the equation

$$\Xi[i] = \Theta[i]E(f_i), \tag{13}$$

where $\Xi[i]$ are the equal loudness normalized filter bank energies and f_i is the center frequency of the ith filter. The multiplication of the filter bank energies, in linear units, by the equal loudness curve, in decibel units, results in filter bank energies in arbitrary units. The resulting energy scale is relative to the perceptual abilities of the species at that frequency.

5. Intensity-loudness power law

The last psychoacoustic related operation is the application of the intensity-loudness power law

$$\Phi[i] = \Xi[i]^{1/3}, \tag{14}$$

where $\Phi[i]$ are the power law and equal loudness normalized filter bank energies. Stevens (1957) found this cube root relationship between the intensity of sound and its perceived loudness in humans. Although this exact relationship may not hold for other species, it is likely that the structural similarities between species yield a comparable correspondence between power and loudness. This relationship may also be different for marine species because of the differences in the propagation of sound through air and water. Regardless of the appropriate power coefficient, this operation is beneficial from a mathematical modeling sense because it reduces the spectrum's dynamic range to make the normalized filter bank energies $\Phi[i]$ more easily modeled by a low-order autoregressive all-pole model.

6. Autoregressive modeling

The last two components of the gPLP model transform the filter bank energies into more mathematically robust features. First, $\Phi[i]$ is approximated by an all-pole model using the autocorrelation method and the Yule-Walker equations as specified in Makhoul (1975). A fifth-order model has been shown to be adequate to model the first two formants of human speech and suppress interspeaker details of the auditory spectrum (Hermansky, 1990). The appropriate order of the LP analysis for other species is dependent on the number of harmonics present in the vocalization, the relative complexity of the power spectrum, and the task being performed.

7. Cepstral domain transform

The autoregressive coefficients a_n from the LP analysis can be transformed directly into equivalent cepstral coefficients c_n using a recursive formula (Deller *et al.*, 1993, p. 376). The primary reason to transform autoregressive coefficients into the cepstral domain is that Euclidean distance is perceptually meaningful in the cepstral domain (Deller *et al.*, 1993, p. 377), whereas a more complex distortion measure such as Itakura distance must be used for autoregressive coefficients to maintain consistency (Itakura, 1975). Cepstral coefficients are generally less correlated with each other than autoregressive coefficients because they are based on an orthonormal set of functions (Deller *et al.*, 1993, p. 377).

B. Greenwood frequency cepstral coefficients (GFCCs)

As an alternative to gPLP it is possible to apply similar techniques to the Mel frequency cepstral coefficients (MFCC) feature extraction model. The MFCC feature extraction model was made popular by Davis and Mermelstein (1980) and has been the most commonly used feature extraction method in human speech processing for many years. While the MFCC model is still widely used because of its computational efficiency, PLP is sometimes preferred because of its robustness and more accurate modeling of the

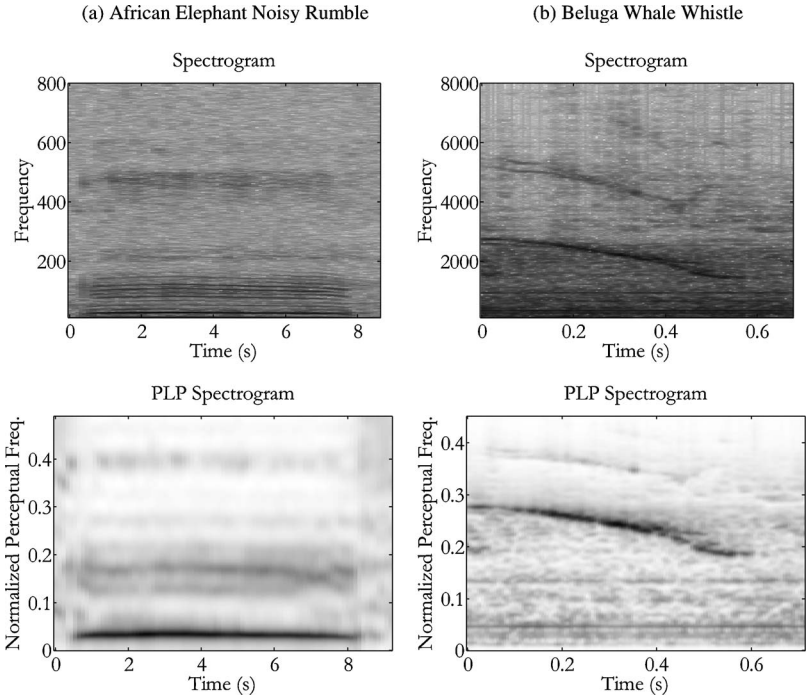

FIG. 3. Perceptual spectrograms. The top plots are traditional FFT-based spectrograms, while the bottom plots are perceptual spectrograms created using gPLP features. The left plots are of an African elephant's noisy rumple, and the right plots are of a beluga whale's whistle. Notice how the perceptual spectrogram enhances the peaks and valleys of the spectrum and warps the frequency axis according to the Greenwood cochlear map function.

human auditory system (Milner, 2002). The application of the Greenwood warping function to the MFCC model results in Greenwood Frequency Cepstral Coefficients (GFCCs) (Clemins *et al.*, 2006, p. 162). Although the GFCC model does not contain all of the psychophysical components of the gPLP model, the same filter bank design details presented here for gPLP are used to design the filter bank in the GFCC model. The main difference between the gPLP model and the GFCC model is the method used for calculating the cepstral coefficients. gPLP uses linear predictive coding (LPC)-derived cepstral coefficients, while GFCC calculates the cepstral coefficients directly from the filter bank energies using a discrete cosine transform. For more details on the GFCC feature extraction model, see Clemins *et al.* (2006).

III. EXAMPLES

The features generated by the gPLP model outlined above can be used to perform many types of analyses on animal vocalizations. Three typical types of analysis are described below utilizing gPLP features.

A. Visualization with perceptual spectrograms

Spectrograms have become an important analysis and visualization tool in the field of bioacoustics. They are useful for many different species and help researchers determine the differences between vocalizations. However, spectrograms do not incorporate any information about the perceptual abilities of the animal and are sometimes dominated by fundamental frequency content rather than spectral shaping. gPLP features can be used to generate perceptual spectrograms, incorporating information about the animal's perceptual abilities into the spectrogram. Because of the incorporation of perceptual data, perceptual spectrograms more closely represent the sound as the animal would hear it.

Figure 3 shows perceptual spectrograms of an African elephant's noisy rumble and a beluga whale's down whistle along with traditional FFT-based spectrograms. These two species were chosen to show that the gPLP model can be used to analyze vocalizations which include formants (elephant rumble) as well as vocalizations with harmonics (beluga down whistle). The perceptual spectrograms are plots of the linear prediction spectrum of each frame of the signal generated directly from LPC coefficients instead of transforming the coefficients into the cepstral domain (Deller *et al.*, 1993, p. 336).

For the perceptual spectrograms, a frame size of 300 ms with 100 ms step size was used to calculate 18 autoregressive coefficients from a filter bank of 50 filters. The perceptual data for the African elephant was taken from Heffner and Heffner (1982), while the data for the beluga whale was acquired from Ketten (1998) and Scheifele (2003). All of the plots, spectrograms, and perceptual spectrograms, are normalized so that pure white represents the absence of spectral energy and pure black represents the peak spectral energy of the vocalization.

In the two examples, notice the frequency warping that occurs in each perceptual spectrogram. The logarithmic warping as dictated by the Greenwood cochlear map function causes the lower frequencies to make up a larger portion of the perceptual spectrogram's horizontal axis. This warping makes small changes in the low frequency components of the vocalization more visible in the perceptual spectrogram. This effect can be seen in the whistle vocalization by examining the dynamics of the first (lowest) harmonic.

In a spectrogram, the excitation signal, which consists of the fundamental frequency and its harmonics, typically masks the response of the vocal tract filter in the spectrum. In contrast to this, the gPLP method enhances the spectral envelope's peaks and valleys and smoothes out the harmonics of the fundamental frequency. This can be seen best in the rumble vocalization's perceptual spectrogram in Fig. 3(a)

where the fundamental frequency harmonics are no longer present. These harmonics, the dark horizontal lines spaced about 20 Hz apart in the fast Fourier transform (FFT)-based spectrogram, distract from the formants (at 40, 100, and 220 Hz) which are much easier to see in the perceptual spectrogram along with their relative magnitudes.

The whistle example in Fig. 3(b) shows how the gPLP extraction model can track signals with quickly changing spectral characteristics. Although the whistle's harmonics move throughout the vocalization, the perceptual spectrogram tracks these changes as well as the FFT-based spectrogram. As with FFT-based spectrogram analysis, smaller window sizes can be used to better track faster moving spectral dynamics.

The gPLP spectrograms improve the contrast between the vocalization energy and the background noise in the spectrogram, making the vocalization easier to visualize. Both vocalizations have a much lighter background in the gPLP spectrogram when compared to the FFT-based spectrogram. However, the darkness of the vocalization energy stays the same, enhancing the contrast between the background noise and the vocalization energy. The gPLP spectrograms also enhance the visualization of narrow-band noise as seen in the beluga whale whistle in two places near 0.05 perceptual Hz and 0.14 perceptual Hz. On the other hand, the noise sources are not as dark as the vocalization energy because of the equal loudness curve applied to the filter bank energies.

B. Classification

Features obtained from gPLP analysis can also be effectively used in various machine learning classification systems. Since gPLP coefficients are perceptually relevant and largely uncorrelated they are a good choice for these tasks. To demonstrate the effectiveness of features generated from the gPLP model in a classification system, an example speaker identification task is performed on a set of vocalizations. This task is appropriate since the spectral characteristics of the rumble continuously change during the vocalization. The data set consists of 143 rumbles from five different African elephants, one male and four females. For more information on the data collection procedure, see Leong *et al.* (2002).

The classification model used for this experiment is a hidden Markov model (HMM). A HMM is a statistical classification model that can represent both the temporal and spectral characteristics of a signal. For more information about the HMM, refer to Clemins *et al.* (2005) and Rabiner (1989). Three state HMMs were used to model the rumble of each elephant and an additional three state HMM was used to model the silence before and after each rumble.

Table I shows the speaker identification accuracies for both MFCC and PLP features as various psychophysical signal processing methods are applied to the feature extraction process. Eighteen coefficients were extracted from 50 filter bank energies using an 18th-order autoregressive model for all trials. The total energy in each frame was also included in the feature vector. The vocalizations were framed using a window size of 300 ms and a window step size of 100 ms. These parameters choices were chosen empirically based on the perceptual information about the species.

TABLE I. Speaker identification accuracies. This table shows the effect of the various psychophysical signal processing components of the gPLP model on the classification accuracy of a speaker identification task.

	Filter bank range	
	10–3000 Hz (%)	10–500 Hz (%)
MFCC	46.9	72.7
PLP with Mel warp, human EQL	49.0	76.2
PLP with Mel warp, elephant EQL	49.0	75.5
PLP with Greenwood warp, human EQL	63.6	74.1
PLP with Greenwood warp, elephant EQL	68.5	81.8

Two different filter bank ranges are used: 10–3000 Hz and 10–500 Hz, to show the effect of limiting the filter bank range for each set of parameters. It is expected that since the range of most of the vocal energy of an elephant rumble is contained in the 10–500 Hz range, the use of that range for the filter bank should result in higher accuracies because it filters out noise in the other frequencies. This hypothesis is verified by the experimental results.

The rows of Table I represent various trials of the experiment with different feature extraction parameters. The first two rows show the change in accuracy when the cepstral coefficients are derived using autoregressive coefficients (gPLP) as opposed to a direct discrete cosine transform (GFCC) of the filter bank energies. Although the use of autoregressive coefficients results in slightly higher accuracies, the significance of the improvement is marginal.

The third row shows the accuracies when the human equal loudness curve is replaced by the derived African elephant equal loudness curve using audiogram data from Heffner and Heffner (1982). In this trial, the Mel-frequency scale was used to place the filters in the filter bank. The incorporation of the elephant equal loudness curve, when used with the Mel-frequency scale, does little to improve accuracy and in one case, decreases classification accuracy.

The fourth row shows the effect of using the Greenwood warping function instead of the Mel-frequency scale. The Greenwood constants were calculated using the optimal $k=0.88$ as suggested by LePage (2003) and the approximate hearing range for the African elephant, 10–10 000 Hz. The human equal loudness curve is used in this trial. While the use of the Greenwood warping function greatly improves accuracy for the larger filter bank range, it does little to improve accuracies for the smaller filter bank range. This suggests that the Greenwood warp helps to focus the analysis on the perceptually important parts of the vocalization when too large of a filter bank range is chosen.

The bottom row combines both the African elephant equal loudness curve and the Greenwood warping function as derived for the African elephant hearing range. When all available species-specific data is incorporated in to the feature extraction process, the classification accuracies improve significantly over the trials in which parameters based on human perception are used. While the Greenwood warping

TABLE II. Results of MANOVA analysis. MANOVA results Wilk's Λ statistic. Each row represents a different experimental setup.

	MANOVA results
All frames	$F_{95,13426}=142.8, P<0.001$
Middle frame	$F_{95,143}=5.81, P<0.001$
Average of all frames	$F_{95,143}=7.09, P<0.001$

function had the most effect when the larger filter bank range was used, the biggest increase in accuracy for the smaller filter bank range occurred when the African elephant equal loudness curve was incorporated into the feature extraction process. It is interesting to note that when used with the Mel-frequency warping scale, the species-specific equal loudness curve decreased the accuracy. However, when the appropriate warping function for that species is used, the species-specific equal loudness curve improved the classification accuracy.

C. Statistical tests

Features generated by the gPLP model can also be used as dependent variables in various statistical tests such as multivariate analysis of variance (MANOVA) or the multivariate *t* test. Since the cepstral coefficients generated by the gPLP model are orthogonal and relatively uncorrelated with each other, techniques such as principle components analysis (PCA) and linear discriminant analysis (LDA) are not necessary as preprocessing steps. To demonstrate the effectiveness of gPLP features in a statistical analysis scenario, a speaker identification experiment using MANOVA is presented.

The main issue with using frame-based features, such as those derived using the gPLP feature extraction model, with statistical tests is that frame-based features generate a feature matrix for each data example instead of a feature vector. Although repeated measures statistical tests might at first seem like an appropriate solution for handling the multiple feature vectors for each vocalization, the vocalizations are the result of a time-varying vocal production system. Therefore, the assumption that the system is unchanging, required for repeated measures tests, is invalidated. Three different methods for overcoming this issue are presented. Each method's advantages and disadvantages are also discussed. Other approaches are discussed in Clemins (2005).

The MANOVA analysis is performed on the same African elephant speaker identification data set used for the classification example. As in the classification example, 18 gPLP coefficients were extracted from each frame of the vocalizations using 50 filters spaced between 10 and 500 Hz along with the energy in each frame.

The first MANOVA analysis uses all of the frames of data from each vocalization in the analysis. The second analysis uses only the feature vector from the middle frame for each vocalization. Finally, the third analysis uses the average feature values across the entire vocalization.

Table II shows the results for the three different trials of the MANOVA analysis. The trial using all of the frames of data had the highest *F* value. The two trials that use one frame of data for each vocalization had substantially lower *F* values. It is interesting to note that using the average value of each feature of all frames in each vocalization resulted in a slightly higher *F* value as compared to using the features from the middle frame of each vocalization. This suggests that there is additional information in other parts of the vocalization besides the middle that could help separate the vocalizations by speaker.

Each of these methods for determining the variables to use has its own advantages and disadvantages. In the first method, the number of observations is much larger than the actual number of vocalizations because each vocalization generates a number of data points, one for each frame of the vocalization. However, because the spectral characteristics of the vocalizations vary over time, this first method more completely quantifies each vocalization. The last two methods have the advantage that they give more reasonable (i.e., lower) *F* values in the analysis because there is only one observation for each vocalization. On the other hand, it is difficult to determine which frame of the vocalization should be used to quantify the vocalization because the spectral characteristics of the vocalization can change dramatically. Therefore, for highly dynamic vocalizations, it might be better to use all of the observed frames instead of picking one frame for analysis as long as the higher *F* values are noted.

IV. CONCLUSIONS

The gPLP model generates perceptually meaningful features for animal vocalizations by incorporating psychophysical information about each species' sound perception. Physically, gPLP coefficients represent the shape of the vocal tract filter during vocalization production. gPLP coefficients are relatively uncorrelated and perceptually meaningful in a Euclidean space. They are also efficient in that a small number of coefficients can model a vocalization frame accurately. These features can be utilized for various types of animal vocalization analyses including visualization, classification, and statistical tests.

gPLP spectrograms are shown to enhance the spectral peaks and suppress broadband background noise. For the speaker identification task, the perceptual information included in the gPLP feature extraction model improves classification accuracy. Finally, the MANOVA analysis shows that the elephants produce significantly different vocalizations, which is consistent with the speaker identification task.

The features generated by the gPLP model can augment or replace traditional frequency-based features. gPLP coefficients can be added to a feature vector of traditional features before a statistical analysis and because they are relatively uncorrelated with each other, they can be added before or after principal component analysis (PCA) or a related technique. Finally, gPLP coefficients have no interpretive bias and decrease analysis time because they can be automatically extracted from the vocalization. Because of its efficiency and adaptability to various species' perceptual abilities, the gPLP model for feature extraction is an innovative and valuable addition to current tools available for bioacoustic signal analysis.

ACKNOWLEDGMENTS

The authors would like to thank the staff of the Wildlife Tracking Center and the Elephant Team at Disney's Animal Kingdom™ for the collection and organization of the acoustic data used in this research.

APPENDIX

The constant values A, a, and k for the frequency warping function, Eq. (5), can be derived from an ERB function of the form in Eq. (7) by taking the integral of the inverse as follows (Zwicker and Terhardt, 1980).

$$f_p = \int \frac{1}{\alpha(\beta f + 1)}, \tag{A1}$$

$$f_p = \frac{1}{\alpha} \int \frac{1}{\beta f + 1}, \tag{A2}$$

$$f_p = \frac{1}{\alpha\beta} \ln(\beta f + 1) + C. \tag{A3}$$

The integration constant C is then set to 0 in order to meet the constraint that $f_p = 0$ when $f = 0$. The base of the logarithm is then changed to 10 in order to match the base in Eq. (9).

$$f_p = \frac{1}{\alpha\beta \log(e)} \log(\beta f + 1). \tag{A4}$$

The equation is in the same form as Eq. (5) and the constant values can be read directly as

$$A = \frac{1}{\beta},$$

$$a = \alpha\beta \log(e),$$

$$k = 1. \tag{A5}$$

Buck, J. R., and Tyack, P. L. (**1993**). "A quantitative measure of similarity for *tursiops truncatus* signature whistles," J. Acoust. Soc. Am. **94**(5), 2497–2506.

Clemins, P. J. (2005). Automatic Classification of Animal Vocalizations. Ph.D. dissertation, Marquette University, Milwaukee, WI.

Clemins, P. J., Johnson, M. T., Leong, K. M., and Savage, A. (**2005**). "Automatic classification and speaker identification of African elephant (*Loxodonta africana*) vocalizations," J. Acoust. Soc. Am. **117**(2), 956–963.

Clemins, P. J., Trawicki, M., Adi, K., Tao, J., and Johnson, M. T. (**2006**). "Generalized perceptual feature for vocalization analysis across multiple species," *Proceedings of ICASSP*, Toulouse, France, May 14–19, 2006, in press.

Darden, S., Dabelsteen, T., and Pedersen, S. B. (**2003**). "A potential tool for swift fox (Vulpes velox) conservation: Individuality of long-range barking sequences," J. Mammal. **84**(4), 1417–1427.

Davis, S. B., and Mermelstein, P. (**1980**). "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," IEEE Trans. Acoust., Speech, Signal Process. **28**(4), 357–366.

Deller, J. R., Proakis, J. G., and Hansen, J. H. L. (**1993**). *Discrete-Time Processing of Speech Signals*, Macmillan Publishing Company, New York.

Fitch, W. T. (**2003**). "Mammalian vocal production: Themes and variation," *Proceedings of First International Conference on Acoustic Communication by Animals*, University of Maryland, College Park, MD, July 27–30, 2003, pp. 81–82.

Fletcher, H. (**1940**). "Auditory patterns," Rev. Mod. Phys. **12**, 47–65.

Fristrup, K. M., and Watkins, W. A. (1992). Characterizing Acoustic Features of Marine Animal Sounds, Technical Report WHOI-92-04 Woods Hole Oceanographic (Woods Hole, MA, Institution.)

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

Greenwood, D. D. (**1990**). "A cochlear frequency-position function for several species–29 years later," J. Acoust. Soc. Am. **87**(6), 2592–2605.

Heffner, R. S., and Heffner, H. E. (**1982**). "Hearing in the elephant (*Elephas maximus*): Absolute sensitivity, frequency discrimination, and sound localization," J. Comp. Physiol. Psychol. **96**(6), 926–944.

Hermansky, H. (**1990**). "Perceptual linear predictive (PLP) analysis for speech recognition," J. Acoust. Soc. Am. **87**(4), 1738–1752.

Itakura, F. (**1975**). "Minimum prediction residual principle applied to speech recognition," IEEE Trans. Acoust., Speech, Signal Process. **23**(1), 67–72.

Ketten, D. R. (1998). "A summary of audiometric and anatomical data and its implications for underwater acoustic impacts," *NOAA Technical Memorandum*.

Leong, K. M., Ortolani, A., Burks, K. D., Mellen, J. D., and Savage, A. (**2002**). "Quantifying acoustic and temporal characteristics of vocalizations of a group of captive African elephants (*Loxodonta africana*)," Bioacoustics 13(3), 213–231.

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

Makhoul, J. (**1975**). "Spectral linear prediction: properties and application," IEEE Trans. Acoust., Speech, Signal Process. **23**, 283–296.

Makhoul, J., and Cosell, L. (**1976**). "LPCW: An LPC vocoder with linear predictive spectral warping," *Proceedings of 1976 International Conference on Acoustics, Speech, and Signal Processing*, Philadelphia, pp. 466–469.

Milner, B. (**2002**). "A comparison of front-end configurations for robust speech recognition," *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, May 13–17. Vol. **1**, pp. 797–800.

Oppenheim, A. V., Schafer, R. W., and Buck, J. R. (**1999**). *Discrete-Time Signal Processing 2nd ed.* Prentice-Hall, Upper Saddle River, NJ.

Owren, M. J., Seyfarth, R. M., and Cheney, D. L. (**1997**). "The acoustic features of vowel-like grunt calls in chacma baboons (*Papio cyncephalus ursinus*): implications for production processes and functions," J. Acoust. Soc. Am. **101**(5), 2951–2963.

Rabiner, L. R. (**1989**). "Tutorial on hidden Markov models and selected applications in speech recognition," Proc. IEEE **77**, 257–286.

Riede, T., and Zuberbühler, K. (**2003**). "The relationship between acoustic structure and semantic information in Diana monkey alarm vocalizations," J. Acoust. Soc. Am. **114**(2), 1132–1142.

Scheifele, P. M. (2003). Investigation into the response of the auditory and acoustic communication systems in the beluga whale (Delphinapterus leucas) of the St. Lawrence River estuary to noise, using vocal classification, Ph.D. dissertation, University of Connecticut, Hartford, CT.

Sjare, B. L., and Smith, T. G. (**1986**). "The vocal repertoire of white whales, *Delphinapterus leucas*, summering the Cunningham Inlet, Northwest Territories," Can. J. Zool. **64**, 407–415.

Stevens, S. S. (**1957**). "On the psychophysical law," Psychol. Rev. **64**, 153–181.

Stoica, P., and Moses, R. L. (**1997**). *Introduction to Spectral Analysis* (Prentice-Hall, Englewood Cliffs, NJ).

Zwicker, E., and Terhardt, E. (**1980**). "Analytical expressions for critical-band rate and critical bandwidth as a function of frequency," J. Acoust. Soc. Am. **68**(5), 1523–1525.