GENERALIZED PERCEPTUAL FEATURES FOR VOCALIZATION ANALYSIS ACROSS MULTIPLE SPECIES

Patrick J. Clemins, Marek B. Trawicki, Kuntoro Adi, Jidong Tao, and Michael T. Johnson

Marquette University
Department of Electrical and Computer Engineering
Speech and Signal Processing Lab
P.O. Box 1881
Milwaukee, WI USA 53201-1881
{patrick.clemins, marek.trawicki, kuntoro.adi, jidong.tao, mike.johnson}@marquette.edu

ABSTRACT

This paper introduces the Greenwood Function Cepstral Coefficient (GFCC) and Generalized Perceptual Linear Prediction (GPLP) feature extraction models for the analysis of animal vocalizations across arbitrary species. These features are generalizations of the well-known Mel-Frequency Cepstral Coefficient (MFCC) and Perceptual Linear Prediction (PLP) approaches, tailored to take optimal advantage of available knowledge of each species' auditory frequency range and/or audiogram data. Illustrative results are presented comparing use of the GFCC and GPLP features versus MFCC features over the same frequency ranges.

1. INTRODUCTION

Mel-Frequency Cepstral Coefficients (MFCCs) [1] and Perceptual Linear Predication (PLP) coefficients [2] are well-established feature representations for human speech analysis and recognition tasks. Each of them benefit from inclusion of frequency-domain perceptual models of the human auditory system. The MFCC approach does this by warping the linear frequency axis to match the Mel-scale cochlear frequency map, while the PLP method adds to this the use of critical band filters, equal-loudness curve amplitude transformation, and cube-root power to intensity transformation. Since these speech processing feature extraction models have been shown to be relatively robust to noise and appropriate for various types of classification tasks including speech recognition, speaker identification, and word spotting, they would likely be good choices for many bioacoustic signal analysis tasks if they could be generalized to easily adjust according to perceptual models of an arbitrary species.

The two main species-specific components of these feature extraction models are frequency warping and the equal loudness curve. To generalize the frequency warping component across arbitrary species, we build on the work of Greenwood [3], who showed that many land and aquatic mammals have a logarithmic cochlear-frequency map, and developed an equation fitting this map. The equal loudness curve component is taken directly from species-specific audiogram measurements, which are available for a wide variety of species.

The second and third sections of this paper will discuss the details of the GFCC and GPLP models, respectively. Section 4 will provide experimental validation through call- and song-type classification and speaker identification experiments on African Elephant and Ortolan Bunting vocalizations, with discussion and conclusions in Section 5.

2. THE GREENWOOD FUNCTION CEPSTRAL COEFFICIENT (GFCC) MODEL

Greenwood [3] found that many mammals perceived frequency on a logarithmic scale along the cochlea. He modeled this relationship with an equation of the form

$$A\left(10^{ax}-k\right), \tag{1}$$

where a, A, and k are species-specific constants and x is the cochlea position. This equation can be used to define a frequency warping through the following equations for real frequency, f, and perceived frequency, f_p:

$$F_p(f) = (1/a)\log_{10}(f/A+k) \tag{2}$$

$$F_p^{-1}(f_p) = A\left(10^{af_p}-k\right). \tag{3}$$

The constants a, A, and k can be found directly by fitting the equation to frequency–cochlear position data, if available. For many species, however, this information has never been measured.

LePage [4] showed that a value of 0.88 for k is optimal in the evolutionary sense. This value results from a trade-off between high-frequency resolution and linearity of the logarithmic map. Assuming this value for k, the other two constants can be solved for given the approximate hearing

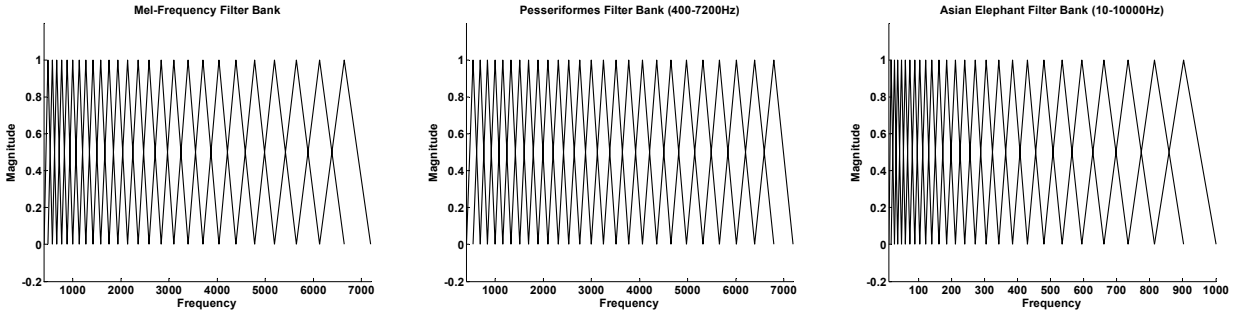

Figure 1: Filter Bank Comparison between Mel-Scale and Greenwood Scale for Passeriformes and Asian Elephants

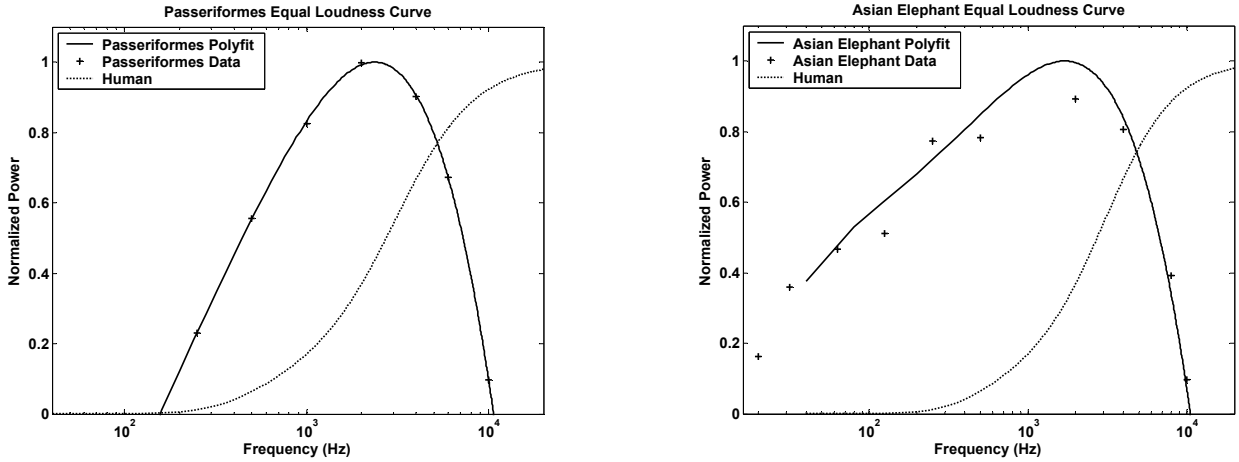

Figure 2: Passeriformes and Asian Elephant Equal Loudness Curves (Solid Line) Compared to Human (Dotted Line)

range ($f_{\min} - f_{\max}$) of the species under study. By setting $F_p(f_{\min}) = 0$ and $F_p(f_{\max}) = 1$, the following equations for a and A are derived

$$A = \frac{f_{\min}}{1-k} \tag{4}$$

$$a = \log_{10}\left(\frac{f_{\max}}{A} + k\right). \tag{5}$$

Thus, using the species specific values for $f_{\min}$ and $f_{\max}$ and an assumed value of k=0.88, a frequency warping function can be constructed. This warping can be used to compute cepstral coefficients in the same way as MFCCs, typically through filterbank windows and a Discrete Cosine Tranform (DCT). Figure 1 shows GFCC filter positioning for African Elephant [5, 6] and Passeriformes (songbird) [7] species compared to the Mel-Frequency scale.

The Mel-Frequency scale employed by the MFCC model is actually a specific implementation of the Greenwood equation. The Bark frequency scale employed by the PLP model does not fit this same form but can be closely approximated by the Mel-scale. Therefore, the above method is an appropriate generalization for both the MFCC and GPLP frequency warping models.

3. GENERALIZED PERCEPTUAL LINEAR PREDICTION (GPLP) COEFFICIENTS

In addition to frequency warping, the PLP model uses an equal loudness curve to model the range of human hearing. Audiograms, widely available for many species, are empirical hearing range curves that can be used directly to estimate equal loudness curves. By substituting a species-specific frequency warping and equal loudness curve, a generalized perceptual linear prediction (gPLP) model can be constructed that takes the perceptual abilities of the species into account during the feature extraction process. The gPLP model is presented in depth in [8, 9].

To create an equal loudness curve for an arbitrary species [10], the audiogram data, A, is first inverted and compared to the hearing threshold, T, using

$$E[f] = -(A[f] - T), \tag{6}$$

where T is usually set to 60dB for land animals and 120dB for aquatic animals [11]. A continuous approximation of the equal loudness curve, $\hat{E}(\log(f))$, is constructed with a 4th-order polynomial fit using the logarithm of frequency. The equal loudness curve is further constrained to have a minimum of zero to prevent negative equal loudness

weights. The Audiograms and equal loudness curves for Asian Elephants [5, 6] and Passeriformes [12]are shown in Figure 2, with the human equal loudness curve superimposed.

Using the previously described frequency-warping method and the above construction for equal loudness curve, GPLP coefficients can be obtained using standard methods [2]. Appropriate features for analysis are typically obtained by computing cepstral coefficients directly from the GPLP linear prediction coefficients.

4. ILLUSTRATIVE EXPERIMENTS

4.1.1. Ortolan Bunting

Norwegian Ortolan Bunting vocalization data was collected from County Hedmark, Norway in May of 2001 and 2002 [13]. Although the birds covered an area of approximately 500 km^2 on twenty-five sites, males were only recorded on eleven of those sites. A team of one to three research members who recognized and labeled the individual male buntings visited the sites. Overall, the entire sample population in 2001 and 2002 contains 150 males, 115 of which are color-ringed for individual identification. Because there are no known acoustic differences between the ringed and non-ringed males, all data was grouped together for these experiments.

Ortolan Buntings communicate with each other through fundamental acoustical units called syllables, analogous to phonetic units in human speech. To produce a song, the syllables are joined in sequence, creating multiple song-types (e.g., *ab*, *cb*, *huf*) and many specific song-type variants (e.g., *aaaabb*, *ccccbbb*, *hhhuff*). In this data set, there are a 63 song-types with 234 distinct variants [13].

Song-type classification and speaker identification experiments [14] were performed on the Ortolan Bunting dataset. MFCCs, GFCCs, and GPLP-derived CCs. Song-type classification experiments used a standard 39 element feature vector consisting of the cepstral coefficients and log energy along with delta and delta-delta coefficients. Speaker identification experiments used only the original 12 element cepstral coefficient vectors. Frequency warping for this species was done with f_{min} = 400 Hz and f_{max} = 7200 Hz, k=0.88, and 26 filterbanks spaced across that range. Equal loudness curves were computed from the audiograms of the Snow Bunting, since Ortolan Bunting measurements are not available. The vocalizations were Hamming windowed with frame and step sizes of 5 ms and 2.5 ms. Classification models for both experiments were 15-state left-to-right HMMs with each state containing a single diagonal-covariance Gaussian. The Baum-Welch Expectation Maximization (EM) algorithm was used to estimate the model parameters, and the Viterbi algorithm was employed for classification. Experiments were run using leave-one-out cross-validation across the data set. HTK software version 3.2.1 from Cambridge University was used to implement the HMMs [15].

Speaker independent song-type classification experiments were performed across the 5 most common song types using 50 exemplars of each song-type, each containing multiple song-type variants. Results are shown in Table 1 below. It can be seen that the GPLP and GFCC slightly out-performed the MFCC feature set, by 0.4% and 0.8%, respectively.

MFCC	GFCC	GPLP
97.6%	98.4%	98.0%

Table 1: Ortolan Song-Type Classification

Song-type dependent speaker identification experiments were performed using 25 exemplars of the most frequent song-type *ab* for each of the 6 vocalizing buntings. Table 2 shows the results for each of the feature sets. Results again indicate that the GFCC and GPLP features slightly outperformed the MFCCs, each by 1.4%.

MFCC	GFCC	GPLP
93.3%	94.7%	94.7%

Table 2: Ortolan Speaker Identification Accuracies

4.1.2. Loxodonta Africana

Animal behavior researchers at Disney's Animal Kingdom™ in Orlando, FL collected the African Elephant (*Loxodonta Africana*) data used in this experiment [16]. Each elephant involved in the data collection project was fitted with a custom designed collar. The collars contained a microphone and an RF radio that broadcast audio to the elephant barn, where it was recorded on DAT tapes. The audio was passed through an anti-aliasing filter and stored on computers at a sampling rate of 7518 Hz.

There were 7 elephants involved in the project: one male and 6 females. Based on social dynamics and breeding requirements, the elephants were released into one of three naturalistic yards each day. The two most common configurations in the main yard were all six females together and one male with four females. Along with the audio recordings, time synchronized video was also recorded. In this way, the researchers were able to label each vocalization with behavior information.

Speaker identification experiments were performed on the *Loxodonta Africana* dataset. As with the Ortolan Bunting experiments, the recognition accuracies were compared across MFCC, GFCC, and GPLP features. Frequency warping for the African Elephants was done with f_{min} = 10 Hz, f_{max} = 10000 Hz, and k=0.88, with 30 filterbanks spaced across the range of 10 – 150 Hz to emphasize the infrasonic vocal range of the vocalizations. Equal loudness curves were computed from the audiograms of the Asian Elephant, since African Elephant

measurements are not available. The vocalizations were all Hamming window with frame and step sizes of 300 ms and 100 ms. Classification models were 5 state left-to-right HMMs with a single diagonal-covariance Gaussian per state. Results were obtained through leave-one-out cross-validation.

Call-type dependent speaker identification experiments were performed using the entire *Loxodonta Africana* dataset. There were a total of six elephants (Bala, Fiki, Mackie, Moyo, Robin, and Thandi) with 20, 30, 14, 17, 34, and 28 rumble exemplars per elephant. Table 3 shows the results for each of the feature extraction models.

MFCC	GFCC	GPLP
81.12%	86.01%	86.01%

Table 3: Elephant Speaker Identification Accuracies

Results show that the GFCC and GPLP each outperform the MFCC features by roughly 5%.

5. CONCLUSIONS

New feature extraction models have been introduced for application to analysis and classification tasks of animal vocalizations. The GFCC, a frequency-warped cepstral coefficient using the Greenwood Function, and the GPLP, a PLP-based model utilizing both Greenwood Function warping and audiogram data, are applicable to a wide variety of species and bioacoustic applications. Results have been shown for two different species, the African Elephant and the Ortolan Bunting, with accuracies indicating performance improvement over MFCCs when used with identical frequency ranges.

6. ACKNOWLEDGEMENTS

This material is based on work supported by the National Science Foundation under Grant No. IIS-0326395.

7. REFERENCES

[1] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, pp. 357-366, 1980.

[2] H. Hermansky, "Perceptual linear predictive (PLP) analysis for speech recognition," *The Journal of the Acoustical Society of America*, vol. 87, pp. 1738-1752, 1990.

[3] D. D. Greenwood, "Critical bandwidth and the frequency coordinates of the basilar membrane," *The Journal of the Acoustical Society of America*, vol. 33, pp. 1344-1356, 1961.

[4] E. L. LePage, "The mammalian cochlear map is optimally warped," *The Journal of the Acoustical Society of America*, vol. 114, pp. 896-906, 2003.

[5] G. V. Békésy, *Experiments in Hearing*. New York: McGraw-Hill, 1960.

[6] R. S. Heffner and H. E. Heffner, "Hearing in the elephant (*Elephas maximus*): Absolute sensitivity, frequency discrimination, and sound localization," *Journal of Comparative and Physiological Psychology*, vol. 96, pp. 926-944, 1982.

[7] E. P. Edwards, "Hearing ranges of four species of birds," *The Auk*, vol. 60, pp. 239-241, 1943.

[8] P. J. Clemins and M. T. Johnson, "Generalized perceptual linear predicition (gPLP) features for animal vocalization analysis," *The Journal of the Acoustical Society of America*, in review.

[9] P. J. Clemins, "Automatic Classification of Animal Vocalizations," in *Electrical and Computer Engineering*. Milwaukee, WI: Marquette University, 2005.

[10] I. Anderson, "How penguins p-p-pick up each other," *New Scientist*, vol. 118, pp. 27, 1988.

[11] D. R. Ketten, "A summary of audiometric and anatomical data and its implications fro underwater acoustic impacts," *NOAA Technical Memorandum*, 1998.

[12] R. J. Dooling, "Avian Hearing and the Avoidance of Wind Turbines," National Renewable Energy Laboratory, Golden, CO NREL/TP-500-30844, June 2002.

[13] T. S. Osiejuk, K. Ratynska, J. P. Cygan, and D. Svein, "Song structure and repertoire variation in ortolan bunting (Emberiza hortulana L.) from isolated Norwegian population," *Annales Zoologici Fennici*, vol. 40, pp. 3-16, 2003.

[14] M. B. Trawicki and M. T. Johnson, "Automatic Song-Type Classification and Speaker Identification of Norwegian Ortolan Bunting (Emberiza Hortulana)," presented at IEEE International Conference on Machine Learning in Signal Processing (MLSP), Mystic, Connecticut, September, 2005.

[15] Cambridge University Engineering Department, *Hidden Markov Model Toolkit (HTK) Version 3.2.1 User's Guide*. Cambridge, MA, 2002.

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