International Journal of Technical Research and Applications e-ISSN: 2320-8163,
www.ijtra.com Volume 4, Issue 3 (May-June, 2016), PP. 463-467

BIRD SOUND RECOGNITION FOR SPECIES IDENTIFICATION

[1]Uddhav Bhosle, [2]Aishwarya Raman
[1,2] Computer Engineering Department, Thadomal Shahani Engineering College, Mumbai University
[1]bhosleuddhav@gmail.com
[2]aishwarya.raman1795@gmail.com

Abstract— **This paper deals with recognizing the species of the bird from the sound signal recorded from the bird's environment. First, the actual part of the signal input that corresponds to the bird sound is selected and the remaining part of the signal is assumed to be silence and removed. The resulting signal contains the bird sound from start to end along with some noise from the surroundings. These noises are eliminated and the features of the actual sound are extracted by using Fast Fourier Transforms(FFT). The features are used to form a model that is compared with already existing models for various bird sounds. The model that matches with the highest probability is chosen and the bird species is determined from it.**

Index terms- **bird; bird species recognition; sound recognition; end point detection; silence removal; noise removal; FFT; HMM;**

I. INTRODUCTION

In many a scenarios a bird cannot be seen, or cannot be identified by people. Bird sounds have a specific quality of being distinct amongst most species. Bird vocalization is a good example of a class of natural sounds where we can expect to find a vocabulary and other structural elements. Bird song can be often seen as an organized sequence of brief sounds from a species specific vocabulary. Those brief sounds are usually called elements or syllables [1].

Bird species recognition has been vital to study in conservation, migration patterns and related research. It has been predicted that wide-scale application of automatic pattern recognition techniques to bird vocalization research could have similar effects as the introduction of the sonogram earlier [1]. In many practical cases, there are limited possibilities to select templates or prototype sounds by hand. Therefore, it is highly preferable to use methods which allow the use of large amount of training data and provide automatic ways to estimate the parameters of the classifier. Kogan et al. compared the manually selected template method to a traditional speech recognition system based on mel-frequency cepstrum coefficients (MFCC) and hidden Markov models (HMMs). In their study, HMM was found more robust to changes in background noise [2].

The working hypothesis was that it would be possible to recognize bird species directly from syllables (or elements), which are building blocks of bird song [1]. Typically, the duration of a syllable ranges from few to few hundred milliseconds. If this were possible, the recognition of species could then be performed even in a noisy environment using brief clean periods. The alternative approach of recognizing song melodies is more challenging. The main problem is that in a typical habitat there may be several birds singing simultaneously. Also, the high regional variability in the songs of many species and imitation of the songs of other species makes it difficult to define characteristic song patterns for each species.

This paper deals with detection of syllables of bird sound in the following way. First the silence or unvoiced part of the signal is differentiated from the voiced (bird sound) part. After doing this, the part necessary for detection is achieved but it still has components of noise (unwanted data from the surrounding environment of the input source). The actual component of the data necessary for the analysis and recognition is obtained by applying the Fast Fourier Transforms on the frequency domain of the signal. This eliminates noise from the input and the data can be further analyzed. Now the features are extracted from the signal and the hidden Markov model (HMM) are used for detecting the species that resembles the signal the most (highest probability of the signal's occurrence).

II. END POINT DETECTION AND SILENCE REMOVAL

For detecting end-points and removing the silenced part of the sound this method uses Probability Density Function(PDF) of the background noise. For differentiating the voiced and silence part of a bird sound, uni-dimensional Mahalanobis Distance function which is a Linear Pattern Classifier(LPC) is used.

Bird Sound can be represented in 3 states. The states are

(i) Silence (S), where no speech is produced

(ii) Unvoiced (U), in which the vocal cords [5] are not vibrating, so the resulting sound waveform is aperiodic or random in nature

(iii) Voiced (V), in which the vocal chords are tensed and therefore vibrate periodically when air flows from the lungs, so the resulting waveform is quasi-periodic [6].

It should be clear that the segmentation of the waveform into well-defined regions of silence, unvoiced, signals is not exact; it is often difficult to distinguish a weak, unvoiced sound from silence, or weak voiced sound from unvoiced sounds or even silence. Since for most of the practical cases the unvoiced part has low energy content and thus silence (background

noise) and unvoiced part is classified together as silence/unvoiced and is distinguished from voiced part.[9]

The speech signal [13] is a slowly time varying signal [14] in the sense, that, when examined over a sufficiently short period

of time (between 5 and 100 msec), its characteristics are fairly stationary[8].

We assume that background noise present in the utterances are Gaussian [11] in nature, however a speech signal may also be contaminated with different types of noise [12]. In such cases the corresponding properties of the noise distribution function are to be used for detection purpose.

The algorithm described below is divided into two parts. First part assigns label to the samples by using a statistical properties of background noise while the second part smoothens the labeling by the physiological aspects from the speech production process. The Algorithm two passes over speech samples. In Pass I (Step 1 to 3) statistical property of background noise is used to make a sample as voiced or silence/unvoiced. In Pass II (Step 4 and 5) physiological aspects of speech production for smoothening and reduction of probabilistic errors are used in statistical marking of Pass I. Step 1: Calculate the mean and standard deviation of the first 1600 samples of the given utterance. If μ and σ are the mean and the standard deviation respectively then analytically we can write,

$$\mu = \frac{1}{1600}\sum_{i=1}^{1600} x(i) \qquad (1)$$

$$\sigma = \sqrt{\frac{1}{1600}\sum_{i=1}^{1600}(x(i) - \mu^2)^2} \qquad (2)$$

Note that background noise is characterized by this μ and σ.

Step 2: Go from 1 st sample to the last sample of the speech recording. In each sample check whether one dimensional Mahalanobis distance function i.e. |x-μ|/σ greater than 3 or not. Analytically,

$$\text{If,} \quad \frac{|x-\mu|}{\sigma} > 3 \qquad (3)$$

the sample is to be treated as voiced sample otherwise it is an silence/unvoiced. Note that the threshold reject the samples

upto 99.7% as per given by equation no. 4 in a Gaussian Distribution thus accepting only the voiced samples.

Step 3: Mark the voiced sample as 1 and unvoiced sample as

0. Divide the whole speech signal into 10 ms non-overlapping windows. Now the complete speech is represented by only zeros and ones.

Step 4: Consider there are M no. of zeros and number of ones in a window. If M ⩾ N then convert each of ones to zeros and vice versa. This method adopted here keeping in mind that a speech production system consisting of vocal chord, tongue, vocal tract etc. cannot change abruptly in a short period of time window taken here as 10 ms. N

Step 5: Collect the voiced part only according to the labeled '1' samples from the windowed array and dump it in a new array. Retrieve the voiced part of the original speech signal from labeled 1 samples.[9]

Now the recorded sound does not contain any Silence parts. The recorded sound is now processed to remove background noise.

III. BACKGROUND NOISE REMOVAL

Noise can distort and disguise the quality of the sound signal. Due to presence of background noise in a recorded bird sound, it cannot be successfully matched with the pre-recorded bird sounds present in the database. Hence background noise removal is a crucial part of audio pre-processing.

Noise removal cannot be successfully implemented in the time domain; rather, it is performed in the frequency domain using Spectral subtraction.

Spectral subtraction noise removal involves segmenting the noisy speech signal into half-overlapped time domain data buffers multiplied by a Hanning window and then transforming the result into the frequency domain using the fast Fourier transform (FFT). Subsequently, noise is removed by subtracting the average magnitude of the noise spectrum from the noisy speech spectrum and zeroing out the negative values using half-wave rectification.[8]

A continuous time signal is sampled at equally spaced time impulses tn = nTs as follows

$$X[n] = X(nT_s) \qquad (4)$$

where is the sampling period or fixed time between each sample. Each impulse value of X[n] is called sample of the discrete-time signal. The sampling period can also be represented as a fixed sampling rate:

$$f_s = \frac{1}{T_s} Hz \qquad (5)$$

The noisy signal is stored using Half-Overlapped Data Buffers. Each segment contains 256 samples of the noisy

sound. Each segment is called a data-buffer [n]. Each data buffer half-overlaps another data buffer by a total of 128

samples.

Data in the half-overlapped data buffers is multiplied with the Hanning time window to optimize the efficiency of computer processing speed. The Hanning time window is a bell curve shape. The portions of the noisy data that lie outside the Hanning time window are zeroed-out, while the portions inside are further evaluated for processing. The mathematical general expression of a Hanning time window is in the form

Otherwise

$$W[n] = \begin{cases} 0.5 - 0.5\cos\left(\frac{2\pi}{L}n\right) & 0 \le n \le L-1, \quad (6) \\ 0 & \text{Otherwise} \end{cases}$$

where L is the length or the number of samples of the Hanning time window. The data that are stored in the Hanning time window are evaluated for spectral computation, which involves computing the discrete Fourier transform (DFT) using the FFT algorithm.

The spectrum of the speech activity containing the noisy speech time frame is denoted as

$$\hat{X}[k] = S[k] + N[k], \quad (7)$$

where S[k] is the spectrum of the clean speech "Real graph" and N[k] is the spectrum of the noise.

We calculate the average of the noise magnitude spectrum for each frequency

$$\mu(k) = E\{|N[k]|\}, \quad (8)$$

where E is the average value operator. In the next subsection, we explain the role of μ (k).

The average magnitude of the noise spectrum is subtracted from the noisy speech spectrum resulting in the signal

$$\hat{S}[k] = |\hat{X}[k]| - \mu(k) \quad (9)$$

In some cases, for each frequency ω, the value of the average magnitude of the noise spectrum is larger than the magnitude of the noisy speech spectrum. This results in negative values after subtracting the average magnitude of the noise spectrum from the noisy speech spectrum. Half-wave rectification consists in replacing those negative values with zero resulting in the signal.

$$\bar{S}[k] = \begin{cases} 0 & if\ \hat{S}[k] < \mu[k] \\ \hat{S}[k] & otherwise \end{cases} \quad (10)$$

After subtraction of average noise magnitude and half wave rectification, the spectrum of sound is in frequency domain.[8] This processed sound is then compared to the pre-recorded sound stored in a database.

IV. HIDDEN MARKOV MODELS(HMM)

Hidden Markov Models (HMMs) provide a simple and effective framework for modeling time-varying spectral sequences. As a consequence, almost all present day large vocabulary continuous speech recognition (LVCSR) systems are based on HMMs [3].

HMM creates stochastic models from known utterances and compares the probability that the unknown utterance was generated by each model. HMMs area broad class of doubly stochastic models for non-stationary signals that can be inserted into other stochastic models to incorporate information from several hierarchical knowledge sources. Since we do not know how to choose the form of this model automatically but, once given a form, have efficient automatic methods of estimating its parameters, we must instead choose the form according to our knowledge of the application domain and train the parameters from known data. Thus the modeling problem is transformed into a parameter estimation problem. Such a model might have one state per syllable with probabilistic arcs between each state. Each syllable would cause (or be produced by) a transition to its corresponding state. One could then train the model for a certain sound and use the parameter sequence to generate other sequences [4].

A. *The Model*

Front-End: The purpose of the Front-End is to parametrize an Input signal (eg. audio) into a sequence of output Features. Voice samples are taken every 10-25 msec. This sample data is feed to the Front-End module for further processing. Output of Front-End is list of feature vector (MFCC-39D). This feature vector are then mapped to symbol using vector quantization [5].

Vector Quantization: It maps Feature vector to symbol. This is also knows as acoustic modeling. This symbols represent HMM state. During recognition process these symbols are matched against unknown symbols. This gives us way to map complex vector in to manageable symbol set.

HMM model creation: Depending on implementation, HMMs

are created for every basic sound unit, in our case it is a syllable. Further all HMMs are linked together to represent the sound under consideration. This linked representation is known as search space for given problem. During recognition phase this graph is searched for finding occurrence of given word.

Training: The most difficult task is to adjust the model parameter to accurately represent the sound under

consideration. In training mode large amount of data is given to HMM model. Using this, HMM adjusts its probability distribution and transition matrix. There is no global optimal algorithm for learning. Every HMM must be trained to maximize it's (local optimum) recognition power.

Recognition: unknown sound is fed to HMM and it's output probability is calculated [6].

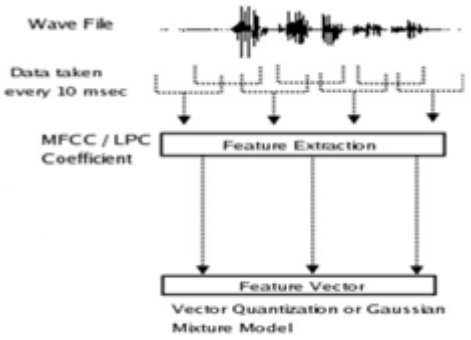

Fig. 1 Feature Vector from the input wave signal

B. *Viterbi Algorithm (Detection)*

The Viterbi algorithm is a dynamic programming algorithm for finding the most likely sequence of hidden states, called the Viterbi path – that results in a sequence of observed events, especially in the context of Markov information sources and hidden Markov models [7].

Given an observed sequence $O = \{o_1, o_2, \ldots, o_N\}$, the single best state sequence $S = \{s_1, s_2, \ldots, s_K\}$ has to be found that matches the observed state sequence with the highest probability. Hence the value of $P(q|O, \lambda)$

$$maxP(q|O,\lambda) = max\frac{P(q|O,\lambda)}{P(O)} = maxP(q,O|\lambda) \quad (11)$$

The last step follows from the fact that the probability of the observation sequence can be seen as a constant. Now we define δt(i) as the best path (the path with the highest probability) from the start into some state i at time t.

$$\delta_t(i) = max_{q_1,q_2,\dots,q_{t-1}} P(q_1 q_2 \dots \dots q_{t-1} q_t = i, o_1 o_2 \dots o_t | \lambda)$$

(12)

Notice that

$$\delta_T(i) = max_{q=q_1,q_2,\dots,q_T} P(q, O | \lambda) \quad (13)$$

Is the value we are looking for.

δt(i) can be calculated using a recursive procedure similar to the forward algorithm, but this time using a maximization over previous states instead of a summing procedure. This algorithm is known as the Viterbi algorithm. The optimal path can be found by keeping track of the argument i that maximized δt(j) in equation 13

Initialization:

$$\delta_1(i) = \pi_i b_i(o_i) \quad 1 \le i \le N \quad (14)$$

$$\varphi_1(i) = 0 \quad (15)$$

Recursion:

$$\delta_t(j) = max_{1 \le i \le N}[\delta_{t-1}(i) a_{ij}] b_j(o_t) \; 2 \le t \le T, 1 \le j \le N \quad (16)$$

$$\varphi_t(j) = \arg max_{1 \le i \le N}[\delta_{t-1}(i) a_{ij}] \quad (17)$$

Termination:

$$\tilde{P} = max_{1 \le i \le N}[\delta_T(i)] \quad (18)$$

$$\tilde{q}_T = \arg max_{1 \le i \le N}[\delta_T(i)] \quad (19)$$

Path backtracking:

$$\tilde{q}_t = \varphi_{t+1}(\tilde{q}_{t+1}) \quad t = T-1, T-2, \dots\dots, 1 \quad (20)$$

Hence the best path for the given sound is found which corresponds to a particular model that is of the bird species as the input sound. Hence the species of the bird corresponding to the input sound is determined.

V. LIMITATIONS

The described method above assumes that there is only one sound in the signal corresponding to a specific bird that has been recorded. If there are two or more sounds of birds of same or different species then the model may not give satisfactory results. Results depend on the intensities of the sounds. There must be only sounds (syllables) as inputs. Bird songs are more complex to determine and require more sophisticated models. Also bird songs vary within a species and are also imitated by other species in some cases.

If the sound to be recognized is not from any modeled specimen, then false output that is closest to a particular model may be generated. Hence models of all required species must be available for determination.

VI. FUTURE SCOPE

The model can be expanded to identifying other animal species in a similar fashion and also other sounds that need to be recognized for various purposes. Also the bird songs or other sequential sounds that are similar to sentences in a human language can be recognized using this technique. Sound signal having two or more bird inputs can be guessed by giving output of more than one bird species along with the probability that the input sound corresponds to that species. This may give the user an estimate of what the species of bird may be.

VII. CONCLUSION

Bird species recognition from sound generated by an individual specimen is studied. The end points of the sound signal and the silence/unvoiced part in the input signal are eliminated using the probability density function of background noise and uni dimensional Mahalanobis Distance function technique. The noise in the resultant signal is eliminated and the feature extraction for forming HMM was done using the FFTs. Finally the extracted symbols are fed into the HMM and the output with the highest probability is determined. This output determines the species of the bird that the input sound belongs to.

REFERENCES

[1] C. K. Catchpole and P. J. B. Slater, Bird Song: Biological Themes and Variations. Cambridge, U.K.: Cambridge Univ. Press, 1995.

[2] Parametric Representations of Bird Sounds for Automatic Species Recognition, IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 14, NO. 6, NOVEMBER 2006.

[3] [3] The Application of Hidden Markov Models in Speech Recognition, Cambridge University Engineering Department, Trumpington Street,

Cambridge, CB2 1PZ, UK, mjfg@eng.cam.ac.uk

[4] Speech Recognition Using Hidden Markov Models, D.B. Paul, The Liru:oln Laboratory Journal, Volume 3, Number 1 (l990)

[5] Willie Walker, Paul Lamere, and Philip Kwok. Sphinx-4: A Flexible Open Source Framework for Speech Recognition. SUN Microsystem, 2004.

[6] Hidden Markov Model and Speech Recognition, Department of Computer Science and Engineering Indian Institute of Technology, Bombay Mumbai.

[7] https://en.wikipedia.org/wiki/Viterbi_algorithm

[8] Noise Removal in Speech Processing Using Spectral Subtraction Marc Karam1, Hasan F. Khazaal2, Heshmat Aglan3, Cliston Cole1

[9] A New Silence Removal and Endpoint Detection Algorithm for Speech and Speaker Recognition Applications

G. Saha1, Sandipan Chakroborty2, Suman Senapati3

[9] K. Ishizaka and J.L Flanagan, "Synthesis of voiced Sounds from a Two- Mass Model of the Vocal Chords," Bell System Technical J., 50(6): 1233-1268, July-Aug., 1972.

[10] Atal, B.; Rabiner, L., "A pattern recognition approach to voiced- unvoiced-silence classification with applications to speech recognition" Acoustics, Speech, and Signal Processing [see also IEEE Transactions on Signal Processing], IEEE Transactions on , Volume: 24 , Issue: 3 , Jun 1976, Pages: 201 - 212.

[11] L. R. Rabiner and R.W. Schafer, "Digital Processing of Speech Signals", First Edition, Chapter-4, Pearson Education, Prentice-Hall.

[12] http://cslu.ece.ogi.edu/nsel/data/SpEAR_technical.html.