International Journal of Electronics, Communication & Soft Computing Science and Engineering
ISSN: 2277-9477, Volume 2, Issue 4

Classification of Bird Species

Mosamee Gund **Aarti Bang**

***Abstract* — this paper is related to the development of signal processing techniques for automatic recognition of bird species. Bird sounds are divided by their function into songs and calls which are further divided into hierarchical levels of phrases, syllables and elements. It is shown that syllable is suitable unit for recognition of bird species. Diversity within different types of syllables birds are able to produce is large. Automatic recognition system for bird species used in this paper consists of segmentation of syllables, feature generation, classifier design.**

***Key Words* — Bird sounds, species recognition, audio classification, pattern recognition, feature extraction, Dynamic Time Warping.**

I. INTRODUCTION

Since prehistoric times, people have interacted with birds. They have long been utilized as a source of food. After the invention of agriculture, they were often seen as pests competing for crop resources. The relationship has continued to evolve ever since.

As humanity and technology spreads across the face of the Earth, interactions, both negative and positive, between birds and people grow. In recent years, public sentiment towards birds has changed from something to be killed for fun, food. Now birds are considered to be deserving of protection. Because birds come and go as they please, and cannot (generally) be kept out by fences, scientists and engineers seek automated ways to determine their presence. Birds, by and large, are a garrulous lot, so microphones and audio processing equipment could possibly provide this capability.

Birds are critical to ecosystem functioning, so techniques to make avian monitoring more efficient and accurate will greatly benefit science and conservation efforts. Birds are particularly abundant and diverse consisting of journalists and specialists as well as migrants and local breeders. They are important consumers; they eat fruit, grains, nectar, and insect. As such they contribute to a variety of important ecosystem functions. They play the important role in controlling insect's population; they are important plant dispersal agent and pollinator. Since bird plays such varied roles in ecosystem functions, they are vulnerable to both human induced habitat change and global climate change and as a result May species are declining.

Acoustic communication in birds is rich and in one of the most directs ways for humans to detect them. Birds sound called as calls are species specific acoustic signature that readily announces their presence. Techniques like mist netting, point counts and transect count are used for surveying birds. The most significant drawback of these methods is the reliance on highly trained professional for making identification.

In recent years, classifying bird species based on recorded vocalization is affined by manual inspection of spectrographs by experts. The fact is manual inspection of sound spectrographs yield correct judgment has encouraged research into automatic classification of bird species.

II. PREVIOUS WORK

Only few studies have been done on automatic recognition of bird species and efficient parameterization of bird sounds. In (Anderson, Dave & Margoliash 1996, Kogan & Margoliash 1998) dynamic time warping and hidden Markov models were used for automatic recognition of songs of Zebra Finches (*Taeniopygia guttata*) and Indigo Puntings (*Passerina cyanea*) [1]. In these studies syllables were represented by spectrograms. Comparison of spectrograms is computationally demanding and, in the case of field recordings, they often include also environmental information that is not relevant for recognition of bird species.

In (McIlraith & Card 1997) tested recognition of songs of six species common in Manitoba, Canada [2]. In this work songs were represented with spectral and temporal parameters of the song. Dimensionalities of the feature space were reduced by selecting features for classification by means of their discriminative ability. Dynamic Time Warping was used for classification of the songs.

III. SEGMENTATION OF BIRD SONG

Bird vocalization is usually considered to be composed of calls and songs. Calls are most commonly brief isolated sounds which are usually associated with a specific communicative function, e.g., they may represent a warning for an approaching predator. Songs are more complicated patterns of vocalization which are most commonly associated with territorial singing of male birds and mating. Bird vocalizations are often divided into hierarchical levels of phrases, syllables, and elements [3]. For example, the levels of a song of the Common Chaffinch (*Fringilla coelebs*) are illustrated in Fig. 1. A phrase is a series of syllables that occur in a particular pattern. Usually, syllables in a phrase are similar to each other, but sometimes they can be also different as in the last frame of the song presented in Fig. 1. Syllables are constructed of elements. In simple cases, syllables are equal to elements, but complex syllables may be constructed from several elements. Separation of elements is often difficult and can be ambiguous. Call sounds are usually composed of only one syllable and the phrase level cannot be detected. The phrase level is also commonly missing in songs of certain species.

In this paper, we call the smallest unit a syllable. A syllable is basically a sound that a bird produces with a single blow of air from the lungs. This is also somewhat inaccurate definition as many birds are capable of complicated circular breathing cycles during singing [4]. The rate of events in bird vocalization may also be so high that the separation of individual syllables is difficult to perform in a natural environment due to reverberation.

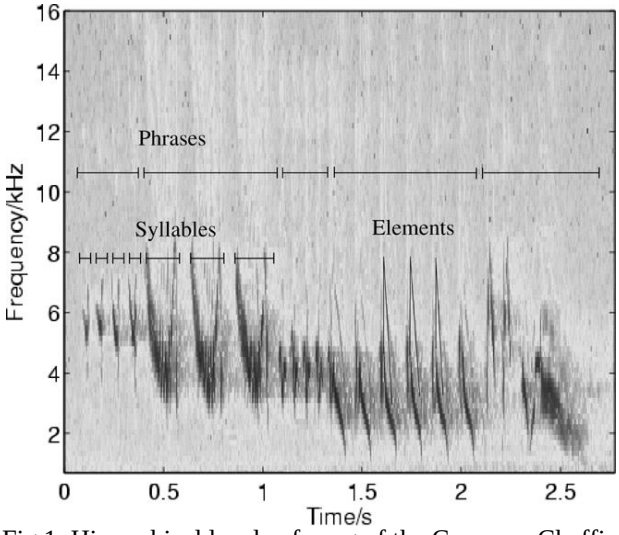

Fig.1. Hierarchical levels of song of the Common Chaffinch.

The segmentation of a recording to individual syllables is performed using an iterative time-domain algorithm [20]. First, a smooth energy envelope of the signal is computed and the global minimum energy is selected as the initial background noise level estimate $\boldsymbol{N}_{\mathrm{dB}}$. Initial threshold $\boldsymbol{T}_{\mathrm{dB}}$ is set to the half of the initial noise level, which is set to the lowest signal envelope energy. Noise level and threshold are updated using the following algorithm until convergence so that the noise level is sufficiently stable.

I. Algorithm:

1) Find syllable candidates, i.e., regions that are above syllable threshold $\boldsymbol{T}_{\mathrm{dB}}$.
2) Update $\boldsymbol{N}_{\mathrm{dB}}$ from gaps between syllable candidates.
3) Update the threshold, e.g. $\boldsymbol{T}_{\mathrm{dB}}=\boldsymbol{N}_{\mathrm{dB}}/2$, and return to step 1.

IV. FEATURES

The objective in pattern recognition or classification is to classify objects (patterns) into number of categories (classes) (Theodoridis & Koutroumbas 1998). In this work syllables extracted from songs and calls of birds are used as patterns. Classification is done based on the features, which are calculated from the syllables to be classified or recognized. Features constitute a feature vector, which is a representation of the syllable. Features are generated in three phases. First is simply calculation of features of patterns (raw data), which is followed by the removal of outliers, clearly erroneous syllables[14]. In data normalization feature values are adjusted to the same dynamic range so that each feature has equal significance to the classification result. Classifier could be also trained with normalized data, but this may require more training data. Training of the classifier could also take more time with normalized data.

Classification is done based on the features, which are calculated from the syllables to be classified or recognized. Features constitute a feature vector, which is a representation of the syllable. Features are generated in three phases. First is simply calculation of features of patterns (raw data), which is followed by the removal of outliers, clearly erroneous syllables. In data normalization feature values are adjusted to the same dynamic range so that each feature has equal significance to the classification result. Classifier could be also trained with normalized data, but this may require more training data. Training of the classifier could also take more time with normalized data [10].

Most of the features are calculated on frame basis. This is common in audio and speech analysis, because the amount and variability of data is reduced. First, syllables are divided into overlapping frames. Features are calculated from windowed frames, which results feature trajectories of the syllable. Mean and variance values of trajectories are calculated, thus each basic feature results in two actual features. Final feature vector include mean and variance values of frame based features plus parameters calculated from the entire syllable. In acoustical feature two features are there,

- **Spectral Features**
- **Temporal Features**

A. Spectral Features:

Frequency range is calculated from the entire syllable. All other spectral features are calculated on the frame basis and they provide short time spectral properties of the syllable. Frame size of 256 samples with 50% overlap is used. Fourier transform is applied to signal frames that are windowed with Hanning window.

1. Spectral centroid (SC)

Spectral centroid is center point of spectrum and in terms of human perception it is often associated with the brightness of the sound. Brighter sound is related to the higher centroid. Spectral centroid for signal frame is calculated as:

$$SC = \frac{\sum_{n=0}^{M} n\left|X(n)\right|^2}{\sum_{n=0}^{M}\left|X(n)\right|^2} \qquad (1)$$

Where X is discrete Fourier transform (DFT) of signal frame and M is half of the size of DFT.

2. Signal bandwidth (BW)

Signal bandwidth is defined as a width of the frequency band of signal frame around center point of spectrum. The bandwidth is calculated as:

$$BW = \sqrt{\frac{\sum_{n=0}^{M} (n - SC)^2 |x(n)|}{\sum_{n=0}^{M} |x(n)|^2}} \quad (2)$$

The bandwidth of syllable is calculated as average of bandwidth of DFT frames of syllable.

3. *Spectral roll off frequency (SRF)*

Spectral roll off frequency is the point below which certain amount of power spectral distribution resides. Feature is related to "skewness" of spectral shape. The measure can Distinguish sounds with different frequency ranges. Spectral roll off frequency for a DFT
Frame is defined as:

$$SRF = \max(K\sum_{n=0}^{K}|X(n)\text{^}2| < TH\sum_{n=0}^{M}|X(n)|\text{^}2) \quad (3)$$

Where *TH* is the threshold between 0 and 1, here we use a commonly used value 0.95.

4. *Delta spectrum magnitude (spectral flux) (SF)*

Delta spectrum magnitude measures difference in spectral shape. It is defined as the 2-norm of difference vector of two adjacent frame spectral amplitudes. It gives a higher value for syllables with a higher between-frame difference. Formula for delta spectrum magnitude calculations is given as:

$$DSM_i = \sum_{n=0}^{M} \left\| X_i(n) - |X_{i+1}(n)| \right\| \quad (4)$$

5. *Spectral flatness (SFM)*

Spectral flatness measures the tonality of a sound. It gives a low value for noisy sounds and a high value for voiced sounds. Measure can discriminate voiced sounds from unvoiced also if they occupy same frequency range. Spectral flatness is the ratio of geometric to arithmetic mean (Markel & Gray 1976) of signal spectrum and it is given in dB scale as:

$$SFM = 10\log_{10}\frac{Gm}{Am} \quad (5)$$

6. *Frequency range (range1, range2)*

Frequency range gives low and high limit value of the frequency range that a syllable occupies. Frequency range is calculated for the whole syllable. The frequency range and the length of the syllable together define boundaries of the syllable. Frequency range is calculated by means of normalized power spectrum of the syllable. Low and high limits are respectively the lowest and highest frequency bin whose power spectrum value in dB scale is above a threshold. The threshold value used here is -40dB.

B. *Temporal features:*

In addition to the features described below, the temporal duration of the syllable (T) is also used as the feature of the syllable. The zero-crossing rate (ZCR) and short time signal energy are calculated on frame basis. The size of a frame is 256 samples and adjacent frames overlap 50% as it was also for the spectral features. Frames are windowed with rectangular window.

1. *Zero-crossing rate (ZCR)*

Zero-crossing rate (ZCR) is number of time domain zero-crossings in processing frame. A zero-crossing occurs when adjacent samples have different signs. ZCR is closely related to spectral centroid as they both measure construction of spectral shape of frame. It is defined for the frame as:

$$\dot{ZCR} = \sum_{n=0}^{M-1} |\text{sgn}(x(n)) - \text{sgn}(x(n+1))| \quad (6)$$

Where x is time domain signal frame and M is the size of the frame. Signum function sgn is defined as:

$$\text{sgn}(x(n)) = \begin{cases} 1, & x(n) \geq 0 \\ -1, & x(n) < 0 \end{cases} \quad (7)$$

2. *Short time signal energy (EN)*

Maximum energy of the energy trajectory for the syllable is normalized to 0dB without normalization energy depends on the recording gain and other recording conditions and would not assign much information on the energy content of the syllable. Normalized energy is able to discriminate syllables with different within-syllable energy content. It is defined for the frame as:

$$E(m) = \sum_{i=1}^{N} 20\log_{10}|xi[n]|^2 \quad (8)$$

V. CLASSIFICATION METHODS AND MODELS

The role of a classifier is to decide which the best possible class for the test pattern is. This is done by comparing similarity between test pattern and model or target patterns of Classes. Classifier does the decision based on the similarity or distance measure between test pattern and model patterns. Suitable distance measure depends on the problem and selected classification scheme. Simplest distance measure is minimum length measure in which Euclidean distance between feature vectors of test pattern and model patterns of classes is calculated.

The recognition of individual syllables is based on the nearest neighbor classifier. This method involves no training process. All samples in the training data are used as such for representing the classes. In the classification phase, the test syllable is compared against all syllables of the training data,

and the class label is determined by the training data sample which has the largest similarity/smallest dissimilarity to the test syllable.

There are different classifiers,

1. Gaussian mixture model
2. Hidden Marko model
3. Support vector machine
4. k-Nearest-Neighbor
5. Dynamic Time Warping

A. Dynamic Time Warping

Syllables have typically different durations. Dynamic time warping (DTW) algorithm can be used for comparing variable length sequences [26]. Its basic idea is to warp the time axes of two sequences nonlinearly so that the maximum fitting between the sequence elements is attained. The computation can be done in a two-dimensional trellis. Here, the word element refers to the generic element of the feature vector sequence.

In the following, two syllables are represented by the trajectory models A and B. The elements of the sequences are frame-based feature vectors and the sequence lengths are denoted by L_A and L_B. The distance between the sequence elements $A(i)$ and $B(j)$ is denoted by $d(i, j)$, and the cumulative distance at trellis coordinate (i, j) is denoted by $g(i, j)$. First, the trellis is initializing

$$g(0, j) = \begin{cases} 0 & j=0 \\ \infty & j=1....L_B \end{cases}$$

$$g(i,0) = \begin{cases} 0 & i=0 \\ \infty & i=1....L_A \end{cases} \quad (9)$$

Cumulative distances are then computed using dynamic programming as follows:

$$g(i, j) = \min \begin{cases} g(i, j-1)+d(i, j) \\ g(i-1, j-1)+\omega_d d(i, j) \\ g(i-1, j)+d(i, j) \end{cases} \quad (10)$$

Where index goes from 1 to L_A and index j from 1 to L_B. Parameter ω_d is the weight of the diagonal movement in the trellis. DTW distance is define to be

$$D(A, B) = \frac{g(L_A, L_B)}{(L_A + L_B)} \quad (11)$$

Here, the cumulative distance is divided by the sum of the lengths of the sequences, but other choices are also possible. In order to use DTW, the distance measure must first be defined for the sequence elements. In the sinusoidal modeling, there are two parameters per sequence element: amplitude and the frequency. We can now consider these two parameters separately and have a two-dimensional vector or use the amplitude information to weigh the importance of the frequency information. These two approaches were compared.

The recognition system based on DTW consists of the templates which are the reference sequences of the classes. Each sequence consists of feature vectors, so each template is a point trajectory in the feature space. In order to expand the point trajectory representation into probability distributions, there are two alternatives. The templates can be divided into segments, and each segment is represented by some probability distribution of the feature vectors. The information about the temporal order of the feature vectors inside each segment will then be lost, but the order of the segments can be maintained by forming chains of segments. This is essentially the concept of Hidden Marko Model (HMM) [27]. Another alternative to bring the basic DTW algorithm into probability domain is to model the distribution of the DTW distances between the training data and the template. This is equal to adding the probability density function to each element of the DTW template and then computing the probability of the data sequence given the chain of probability density function by means of the Viterbi algorithm [28]. The difference between these two alternatives is that in the HMM the temporal resolution is usually relatively small, i.e., the number of the states in the HMM is smaller than the number of the elements in the typical DTW template.

The benefit of the HMM approach is that different states can have different probability density function and thus the changing variance of different parts of the sequence can be taken into account in the model. When only the distribution of the cumulative distance is modeled, the same probability density function is applied to all parts of the sequence. These alternatives are illustrated in Fig. 2.

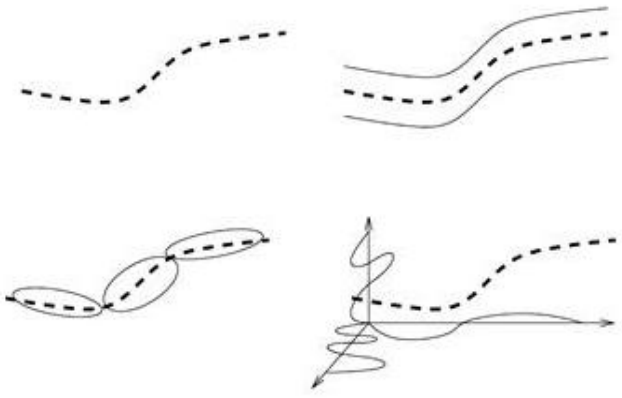

Fig. 2. Examples of trajectory models.

The motivation for the use of DTW was the desire to compute the distances between syllables with varying lengths.

CONCLUSION

In this work focus has been in species that produce regularly sounds that response not tonal or harmonic in structure. The long term objective in this research is to envelop methodology for a system that is capable to recognize majority of common Finnish bird pieces in field conditions.

The sounds of birds are produced mainly by the unique organ called syrinx. Diversity Within structure of syrinx of different species is large, which evoke large number of different sounds birds can produce. Bird sounds can be divided by function into songs and calls. Songs are more spontaneous than calls and mostly produced by males during the breeding season. Call sounds are produced by both sexes throughout the year and they occur in some particular context with certain function. The DTW method was used for classification.

ACKNOWLEDGMENT

I would like to thank my co-author and advisor Mrs. Aarti Bang, Electronics and Telecommunication Engineering Department. Her guidance this paper work is carried out and their constant interest, encouragement and proper direction during completion of this work.

I would also like to thank all the staff members of Electronics and Telecommunication Engineering Department for their co-operation and support.

REFERENCES

[1] Theodoridis, S. & Koutroumbas, K., *Pattern Recognition*, vol.1, Academic Press, San Diego, California, USA,1998.

[2] A. L. McIlraith and H. C. Card, "Birdsong recognition using backpropagation and multivariate statistics," *IEEE Trans. Signal Process*, vol. 45, no. 11, pp. 2740–2748, Nov. 1997.

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

[4] A. S. King and J. McLelland, Eds., "Larynx and Trachea," in *Form and Function in Birds*. New York: Academic, 1989, vol. 4, pp. 69–103.

[5] R. K. Potter, G. A. Kopp, and H. C. Green, *Visible Speech*. New York: Van Nostrand, 1947.

[6] B.-S. Shieh, "Song structure and microgeographic variation in a population of the Grey-cheeked Fulvetta (Alcippe morrisonia) at Shoushan nature park, southern Taiwan," *Zool. Stud.*, vol. 43, no. 1, pp. 132–141, 2004.

[7] P. J. Christie, D. J. Mennill, and L. M. Ratcliffe, "Chickadee song structure is individually distinctive over long-broadcast distances," *Behavior*, vol. 141, no. 1, pp. 101–124, 2004.

[8] O. Tchernichovski, F. Nottebohm, C. E. Ho, B. Pesaran, and P. P. Mitra, "A procedure for an automated measurement of song similarity," *Animal Beh.*, vol. 59, pp. 1167–1176, 2000.

[9] P. Galeotti and G. Pavan, "Individual recognition of male Tawny owls (Strix aluco) using spectrograms of their territorial calls," *Ethology, Ecology, Evol.*, vol. 3, no. 2, pp. 113–126, 1991.

[10] K. Ito, K. Mori, and S. Iwasaki, "Application of dynamic programming matching to classification of budgerigar contact calls," *J. Acoust. Soc. Amer.*, vol. 100, no. 6, pp. 3947–3956, Dec. 1996.

[11] C. Rogers, "High resolution analysis of bird sounds," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process*, 1995, pp. 3011–3014.

[12] A. Härmä and M. Juntunen, "A method for parameterization of tim evarying sounds," *IEEE Signal Process. Lett.*, vol. 9, no. 5, pp. 151–153, May 2002.

[13] J. A. Kogan and D. Margoliash, "Automated recognition of bird song elements from continuous recordings using dynamic time warping and hidden Markov models: a comparative study," *J. Acoust. Soc. Amer.*, vol. 103, no. 4, pp. 2185–2196, Apr. 1998.

[14] S. E. Anderson, A. S. Dave, and D. Margoliash, "Template-based automatic recognition of birdsong syllables from continuous recordings," *J. Acoust. Soc. Amer.*, vol. 100, no. 2, pp. 1209–1219, Aug. 1996.

[15] C. Kwan, G. Mei, X. Zhao, Z. Ren, R. Xu, V. Stanford, C. Rochet, J. Aube, and K. C. Ho, "Bird classification algorithms: Theory and experimental results," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Montreal, QC, Canada, May 2004, pp. 289–292.

[16] A. Härmä, "Automatic recognition of bird species based on sinusoidal modeling of syllables," in *IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP)*, Hong Kong, China, Apr. 2003, pp. 545–548.

[17] A. Härmä and P. Somervuo, "Classification of the harmonic structure in bird vocalization," in *IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP)*, Montreal, QC, Canada, May 2004, pp. 701–704.

[18] P. Somervuo and A. Härmä, "Bird song recognition based on syllable pair histograms," in *IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP)*, Montreal, QC, Canada, May 2004, pp. 825–828.

[19] W. H. Thorpe, *Bird Song*. Cambridge, U.K.: Cambridge Univ. Press, 1961.

[20] S. Fagerlund, "Automatic Recognition of Bird Species by Their Sounds," M.S. thesis, Helsinki Univ. Technol., Espoo, Finland, 2004.

[21] D. Li, I. K. Sethi, N. Dimitrova, and T. McGee, "Classification of general audio data for content-based retrieval," *Pattern Recognition Lett.*, vol. 22, pp. 533–544, 2001.

[22] B. George and M. J. T. Smith, "Speech analysis/synthesis and modification using an analysis-by-synthesis/overlap-add sinusoidal model," *IEEE Trans. Speech Audio Process.*, vol. 5, no. 5, pp. 389–406, Sep. 1997.

[23] B. P. Bogert, M. J. R. Healy, and J. W. Tukey, "The frequency analysis of time series for echoes: cepstrum, pseudo-auto covariance, cross-cestrum and saphe cracking," in *Proc. Symp. Time Series Analysis*, Istanbul, Turkey, Jun. 5–9, 1963, pp. 209–243.

[24] Nelson, D. A., '"he importance of invariant and distinctive features in species recognition of bird song", *Condor* 91, 120–130, 1989.

[25] C. R. Jankowski, Jr., H.-D. H. Vo, and R. P. Lippman, "A comparison of signal processing front ends for automatic word recognition," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 4, pp. 286–292, Jul. 1995.

[26] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP–26, no. 1, pp. 43–49, Feb. 1978.

[27] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989.

[28] A. J. Viterbi, "Error bounds for convolutional codes and an asymptotically optimal decoding algorithm," *IEEE Trans. Inform. Theory*, vol. IT-13, no. 2, pp. 260–269, Apr. 1967.

[29] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via theEMalgorithm," *J. Roy. Statist. Soc., Series B*, vol. 39, no. 1, pp. 1–38, 1977.

[30] C. M. Bishop, *Neural Networks for Pattern Recognition*. Oxford, U.K.: Oxford Univ. Press, 1995.

[31] Y. Linde, A. Buzo, and R. Gray, "An algorithm for vector quantizer design," *IEEE Trans. Commun.*, vol. C-28, no. 1, pp. 84–95, Jan. 1980.
[32] R. Gray, "Vector quantization," *IEEE Acoust., Speech, Signal Process.Mag.*, vol. 1, no. 2, pp. 4–29, Feb 1984.
[33] T. Kohonen, *Self-Organizing Maps*. Berlin, Germany: Springer,1995.
[34] B.H. Juang and L. R. Rabiner, "The segmental K-means algorithm for estimating parameters of hidden Markov models," *IEEE Trans. Acoust.,Speech, Signal Process.*, vol. 38, no. 9, pp. 1639–1641, Sep. 1990.
[35] P. Somervuo, "Speech recognition using temporally connected kernels in mixture density hidden Markov models," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Istanbul, Turkey, Jun. 5–9, 2000, pp. 3434–3437.
[36] Panu Somervuo, Aki Härmä, and Seppo Fagerlund, "Parametric representations of bird sounds for automatic species recognition", in *IEEE Trans*. on audio, speech, and language processing, vol. 14, no. 6,Nov -2006.

AUTHOR'S PROFILE

Mosamee Gund
Received the degree of B.E in Electronics and Telecommunication Engineering from Terna Engineering College, Osmanabad, Maharashtra in 2006 and pursuing M.E. in Electronics & Telecommunication Engineering from VIIT College of Engineering, Pune. With having 4 years Industrial experience.

Aarti Bang
M. E. in Electronics is assistant professor in the department of Electronics and Telecommunication Engineering at VIIT college of Engineering. With over 20 years of experience, she has published papers in national and international conferences. Her main areas of interest include Signal Processing.