Hidden Markov and Gaussian mixture models for automatic call classification

Judith C. Brown[1]
and Paris Smaragdis[2]

[1] Physics Department, Wellesley College, Wellesley Massachusetts 02481 and Media Laboratory, Massachusetts Institute of Technology Cambridge Massachusetts 02139 brown@media.mit.edu
[2] Adobe Systems, Newton, Massachusetts paris@adobe.com

(Received 17 Dec 2008; published online 11 May 2009)

Automatic methods of classification of animal sounds offer many advantages including speed and consistency in processing massive quantities of data. Calculations have been carried out on a set of 75 calls of Northern Resident killer whales, previously classified perceptually (human classification) into seven call types, using, hidden Markov models (HMMs) and Gaussian mixture models (GMMs). Neither of these methods has been used previously for classification of marine mammal call types. With cepstral coefficients as features both HMMs and GMMs give over 90% agreement with the perceptual classification, with the HMM over 95% for some cases.

Outline

1. Introduction
2. Background
    1. Gaussian mixture models
    2. Hidden Markov models
3. Calculations and results
    1. GMM results
    2. HMM results
4. Conclusion
5. Acknowledgments
6. References

Introduction

The automatic classification of marine mammal sounds is very attractive as a means of assessing massive quantities of recorded data, freeing humans, and offering rigorous and consistent output. Calculations on a set of vocalizations of Northern Resident killer whales using dynamic time warping were reported recently. (Brown and Miller, 2006[2], 2007[3]). Since this method requires the time-consuming preprocessing measurement of the frequency contours, the methods of Gaussian mixture models (GMMs) and hidden Markov models (HMMs) have been explored. These calculations can be applied directly to the time-

frequency decomposition of the recorded signals and have not been used previously for the classification of call types of marine mammals.

Background

Gaussian mixture models

The GMM is a commonly used estimate of the probability density function used in statistical classification systems. GMM classifiers (Duda et al., 2001[10]) are well known for their ability to model arbitrarily complex distributions with multiple modes and are effective classifiers for many tasks.
Although GMMs have found widespread use in speech research, primarily for speaker recognition (Reynolds and Rose, 1995[20] and references therein), and have been used in other fields, for example, for musical instrument identification (Brown, 1999[1] and Brown et al., 2001[5]), there is only one report in animal bioacoustic research (Roch et al., 2007[21]).
Roch et al. (2007)[21] used GMMs to distinguish among three dolphin species obtaining 67%–75% accuracy. In this study they vary the number of mixtures from 64 to 512, in steps differing by a factor of 2. Best results were obtained with 256 mixtures using 64 cepstral coefficients as features on sounds of duration from 1 to 30 s.
The cepstrum is the Fourier transform of the log magnitude spectrum (Oppenheim and Schafer, 1975[13]); it involves two transforms which makes it computationally more intensive than fast Fourier transform (FFT) based calculations. The choice of cepstra as features has been particularly successful in characterizing the vocal tract resonances which identify individual speakers, speech, or vowels. See Rabiner and Schafer, 1978[18] and Rabiner and Huang, 1993[17] for a discussion of the use of cepstra for speech applications.

Hidden Markov models

The HMM is widely used in human speech processing and is described in tutorials by Rabiner and Juang (1986)[15] and Rabiner (1989)[14]. An excellent introuction to HMMs can be found on the website of R. D. Boyle
"Hidden Markov models," http://www.comp.leeds.ac.uk/roger/HiddenMarkovModels/html_dev/main.html. Last viewed 5/6/2009.
with an introduction for animal bioacousticians in Clemins (2005)[6]. A HMM models temporal data in as a sequence of *states*. States are usually defined as separate GMMs, and their successive usage across time is governed by a *transition matrix*. The transition matrix is learned from training data and defines the probabilities of moving from one state to another, ensuring that the data are optimally explained. Ultimately, what the HMM does is create a sequence of GMM models to explain the input data, thus being sensitive to temporal changes.
HMMs have been used far more extensively than GMMs in the field of animal bioacoustics. The principal difference in the two methods is that the HMM takes account of the temporal progression of the sound and is thus able to describe the structure of the call. The GMM treats the entire sound as a single entity with unique spectral properties which characterize each class. Since the HMM takes account of the temporal structure of the call, it uses the temporal variation of the calls as additional information to disambiguate among call types. In comparison a GMM could not distinguish a call type from itself played backwards since it does not examine the temporal structure.
Weisburn et al. (1993)[24] used a HMM for detecting bowhead whale notes with the three largest peaks in

the FFT as features. Kogan and Margoliash (1998)[11] compared the methods of HMM and dynamic time warping for automated recognition of bird song elements and found that the HMM was more robust. Mellinger and Clark (2000)[12] compared spectrogram correlation to HMMs on the task of recognizing bowhead whale calls finding that the spectrogram worked marginally better. Datta and Sturtivant (2002)[9] used HMMs to identify three different groups of dolphin whistles, finding that one group was very distinct from the other two.

More recently HMMs have been used for classification or detection of vocalizations by African elephants (Clemins et al., 2005[8]; Clemins and Johnson, 2006[7]), red deer (Reby et al., 2006[19]), and the ortolan bunting (Trawicki et al., 2005[23], Tao et al., 2008[22]). There have been no computations with HMMs or GMMs on automatic classification of call types of marine mammal sounds.

Calculations and results

The features chosen for all of the calculations were cepstral coefficients and their temporal derivatives. These were calculated with the program melcepst available with the MATLAB toolbox VOICEBOX.
M. Brooks, "VOICEBOX: Speech processing toolbox for MATLAB,"
http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html. Last viewed 5/6/2009.

The sample rate was 44 100 samples/s with each sound divided into 23 ms segments for the calculations. The HMM/GMM computations were carried out with software by Paris Smaragdis. The training set for all classifications consisted of all the sounds except the one being classified, called the "leave one out" method. Preliminary results were reported by Brown and Smaragdis (2008)[4].

GMM results

Results for the GMM calculations are given in Fig. 1 with the number of Gaussians in the probability distributions varying from 1 to 6 and the number of cepstral coefficients from 8 to 30. The calculation diverges for more than 4 Gaussians with 18 or more features due to model overfitting.

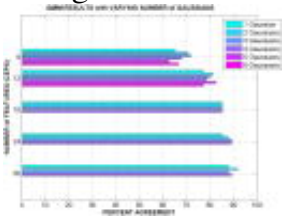

Fig 1.
Gaussian mixture model results showing the dependence on the number of features (cepstral coefficients) and the number of Gaussians in the estimate of the probability density function.
View first occurrence of Fig. 1 in article.

Agreement with the perceptual (human classification) results were over 85% with 18–30 features. The calculation is not highly sensitive to the number of Gaussians. Best results were obtained for 30 features with two Gaussians and gave 92% agreement.

HMM results

For the HMM classification, a left-to-right model was used, and there were three variable parameters rather than two. The number of Gaussians in the probability function was varied from 1 to 4 with the results consistently about 5% better for one Gaussian than two and from 3% to 10% better for two Gaussians than three. The number of states was varied from 5 to 17 and the number of features from 8 to 42 with results

given in Fig. 2. Excellent agreement with the perceptual classification is obtained over a wide range of these paramters with over 90% for from 18 to 42 features and from 9 to 17 states. Truly outstanding agreement of over 95% was obtained for 24 to 30 features and 13 to 17 states.

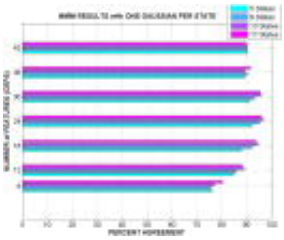

Fig 2.
Hidden Markov model results showing the dependence on the number of features (cepstral coefficients) and the number of states in the model with one Gaussian in the estimate of the probability density function.
View first occurrence of Fig. 2 in article.

Conclusion

These results demonstrate that both GMMs and HMMs are highly successful in the task of automatic classification of killer whale call types, with the performance of the HMM being truly outstanding. Even more impressive is the wide range of parameters over which the calculations agree with the perceptual classification indicating a very robust calculation and great promise for successful extension to other data sets and other species.

Acknowledgments

We are very grateful to Patrick Miller for the killer whale sounds which made this study possible.

References (23)

1. Brown, J. C. (1999). "Computer identification of musical instruments using pattern recognition with cepstral coefficients as features," J. Acoust. Soc. Am. **105**, 1933–1941JASMAN000105000003001933000001. [ISI] [MEDLINE] | first citation in article
2. Brown, J. C., and Miller, P. J. O. (2006). "Classifying killer whale vocalization using time warping," Acoust. Today **16**, 45–47. | first citation in article
3. Brown, J. C., and Miller, P. J. O. (2007). "Automatic classification of killer whale vocalizations using dynamic time warping," J. Acoust. Soc. Am. **122**, 1201–1207JASMAN000122000002001201000001. [MEDLINE] | first citation in article
4. Brown, J. C., and Smaragdis, P. (2008). "Automatic classification of vocalizations with Gaussian mixture models and hidden Markov models," J. Acoust. Soc. Am. **123**, 3345JASMAN000123000005003345000003. | first citation in article
5. Brown, J. C., Houix, O., and McAdams, S. (2001). "Feature dependence in the automatic identification of musical woodwind instruments," J. Acoust. Soc. Am. **109**, 1064–1072JASMAN000109000003001064000001. [ISI] [MEDLINE] | first citation in article
6. Clemins, P. J. (2005). "Automatic classification of animal vocalizations," Ph.D. thesis Marquette University, Milwaukee, WI. | first citation in article
7. Clemins, P. J., and Johnson, M. T. (2006). "Generalized perceptual linear prediction feature for animal vocalization analysis," J. Acoust. Soc. Am. **120**, 527–

534JASMAN000120000001000527000001. [MEDLINE] | first citation in article
8. 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**, 956–963JASMAN000117000002000956000001. [MEDLINE] | first citation in article
9. Datta, S., and Sturtivant, C. (2002). "Dolphin whistle classification for determining group identities," Signal Process. **82**, 251–258. [Inspec] | first citation in article
10. Duda, R. O., Hart, P. E., and Stork, D. G. (2001). *Pattern Classification*, 2nd ed. (Wiley, New York). | first citation in article
11. Kogan, J. A., and Margoliash, D. (1998). "Automated recognition of bird song elements from continuous recordings using dynamic time warping and hidden Markov models: A comparative study," J. Acoust. Soc. Am. **103**, 2185–2196JASMAN000103000004002185000001. [ISI] [MEDLINE] | first citation in article
12. Mellinger, D. K., and Clark, C. W. (2000). "Recognizing transient low-frequency whale sounds by spectrogram correlation," J. Acoust. Soc. Am. **107**, 3518–3529JASMAN000107000006003518000001. [ISI] [MEDLINE] | first citation in article
13. Oppenheim, A. V., and Schafer, R. W. (1975). *Digital Signal Processing* (Prentice-Hall, Inc., Englewood Cliffs). | first citation in article
14. Rabiner, L. R. (1989). "A tutorial on hidden Markov models and selected applications in speech recognition," Proc. IEEE **77**, 257–286. | first citation in article
15. Rabiner, L. R., and Juang, B. H. (1986). "An introduction to hidden Markov models," IEEE ASSP Mag. **3**, 4–16. | first citation in article
16. Rabiner, L. R., and Juang, B. H. (1993). *Fundamentals of Speech Recognition* (Prentice Hall, Englewood Cliffs). | first citation in article
17. Rabiner, L. R., and Schafer, R. W. (1978). *Digital Processing of Speech Signals* (Prentice-Hall, London). | first citation in article
18. Reby, D., Andr-Obrecht, R., Galinier, A., Farinas, J., and Gargnelutti, B. (2006). "Cepstral coefficients and hidden Markov models reveal idiosyncratic voice characteristics in red deer Cervus elaphus stage," J. Acoust. Soc. Am. **120**, 4080–4089JASMAN000120000006004080000001. [MEDLINE] | first citation in article
19. Reynolds, D. A., and Rose, R. C. (1995). "Robust text-independent speaker identification using Gaussian mixture speaker models," IEEE Trans. Speech Audio Process. **3**, 72–83. [Inspec] [ISI] | first citation in article
20. Roch, M. A., Soldevilla, M. S., Burtenshaw, J. C., Henderson, E. E., and Hildebrand, J. A. (2007). "Gaussian mixture model classification of odontocetes in the Southern California Bight and the Gulf of California," J. Acoust. Soc. Am. **121**, 1737–1748JASMAN000121000003001737000001. [MEDLINE] | first citation in article
21. Tao, J., Johnson, M. T., and Osiejuk, T. S. (2008). "Acoustic model adaptation for ortolan bunting (Emberiza hortulana L.) song-type classification," J. Acoust. Soc. Am. **123**, 1582–1590JASMAN000123000003001582000001. [MEDLINE] | first citation in article
22. Trawicki, M. B., Johnson, M. T., and Osiejuk, T. S. (2005). "Automatic song-type classification and speaker identification of Norwegian ortolan bunting Emberiza hortulana vocalizations," IEEE Workshop on Machine Learning for Signal Processing, Mystic, CT, pp. 277–282. | first citation in article
23. Weisburn, B. A., Mitchell, S. G., Clark, C. W., and Parks, T. W. (1993). "Isolating biological acoustic transient signals," IEEE ASSP Mag. **1**, 269–272. | first citation in article