NEAREST CENTROID ERROR CLUSTERING FOR RADIAL/ELLIPTICAL BASIS FUNCTION NEURAL NETWORKS IN TIMBRE CLASSIFICATION

Tae Hong Park
Tulane University
Music Department
102 Dixon Hall
New Orleans, LA 70118
USA

Perry Cook
Princeton University
Computer Science Department
and Music Department
Princeton, NJ 08544
USA

ABSTRACT

This paper presents a neural network approach for classification of musical instrument sounds through Radial and Elliptical Basic Functions. In particular, we discuss a novel automatic network fine-tuning method called Nearest Centroid Error Clustering (NCC) which determines a robust number of centroids for improved system performance. 829 monophonic sound examples from the string, brass, and woodwind families were used. A number of different performance techniques, dynamics, and pitches were utilized in training and testing the system resulting in 71% correct individual instrument classification (12 classes) and 88% correct instrument family (3 classes) classification.

1. INTRODUCTION

Examples of Radial Basis Functions can be readily found in pattern classification applications such speech recognition and prediction [14, 3], phoneme recognition [1], and face recognition [7]. However, they have not been sufficiently explored for automatic timbre recognition research. Considering that there exists only one study with RBFNs [6] and no studies of EBFNs that we know of in machine-based timbre classification, this paper may provide some insights on the prospect and possibilities for RBFN/EBFNs in automatic timbre classification. This paper does not elaborate on feature extraction algorithms or explain RBFN/EBFNs in depth (details can be found in [16]) but rather focuses on the NCC method which automatically fine-tunes the network by spawning additional finer centroids to improve performance of the system.

2. SYSTEM OVERVIEW

The architecture of the system is built around a bottom-up model with a front-end feature extraction module and back-end neural network training and classification module. A sampling frequency of 22.05 kHz and 2 second excerpts with attack and steady-state portions were used for each of the 829 monophonic samples (86% Siedlaczek Library [2], 14% personal collection). The 12 features that were used for the 12 instruments (elec. bass 30, violin 105, cello 102, viola 75, clarinet 100, flute 99, oboe 55, bassoon 35, French horn 56, trumpet 82, tuba 32 examples) included spectral shimmer, spectral jitter, spectral spread, spectral centroid, LPC noise, inharmonicity, attack time, harmonic slope, harmonic expansion/contraction, spectral flux shift, temporal centroid, and zero-crossing rate (see [16] for details). Various performance articulations were present in the majority of the samples including pizzicato, spiccato, sordino, long/sustained/short, detaché, espressivo, vibrato/non-vibrato, pianissimo, piano, mezzo-forte, forte, and fortissimo with pitches ranging between 1~3 octaves.

3. RBFN/EBFN OVERVIEW

3.1. RBFN/EBFN Characteristics

The basic structure of a RBFN/EBFN system is shown in figure 1. Some of the main attributes of a RBFN/EBFN system are the location of the weights found at the output of the basis functions and the characteristic single hidden layer.

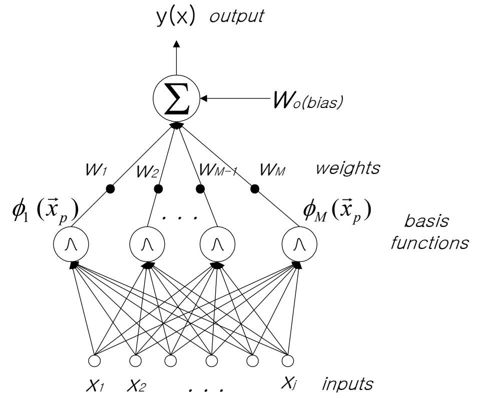

Figure 1. Basic RBF/EBF Network

Exploiting the configuration of activation functions and weights, RBF/EBF networks can take non-linear input spaces and output linear activation outputs, effectively modeling complex patterns which Multi-Layered Perceptrons (MLP) can only achieve through multiple hidden layers [11]. Each basis function consists of a unique centroid, spread, and particular activation function (Gaussian type was used in this paper). The objective in the training phase is to adjust the weights and basis function parameters to reduce the error between the known network outputs and the actual computed outputs. This is determined via gradient descent and back-propagation (see [16] for details).

3.2. Basis Functions

The RBF basis function seen below is computed via Euclidian distance r where p is the sample number, μ_i is the mean for cluster i, and N is the input dimension.

$$r_{Euclidian} = \left\| x_p - \mu_i \right\| = \sqrt{\sum_{n=1}^{N} (x_{np} - \mu_{nj})^2} \qquad (1)$$

$$\phi_i(x_p) = e^{-r_{Euclidian}^2 / 2\sigma^2} \qquad (2)$$

The difference between RBF and EBF is in the distance computation. That is, for EBFs the Mahalanobis distance is computed:

$$r_{Mahalanobis} = \sqrt{(x_p - \mu_i)^T \Sigma_i^{-1} (x_p - \mu_i)} \qquad (3)$$

Figure 2 illustrates a 2-dimensional space with radial and elliptical activation functions.

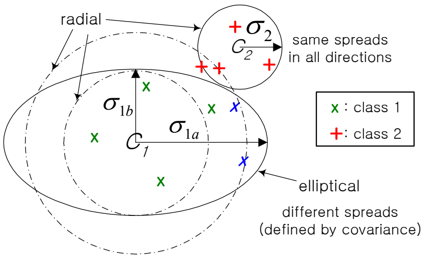

Figure 2. EBF/RBF clustering patterns

3.3. Network Initialization

For network initialization we used k-means to compute the initial basis function parameters and $\frac{\partial E}{\partial w} = 0$ to solve for initial weights w with respect to the total error squared E yielding *(A* is activation output, d is known output):

$$w = (A^T A)^{-1} A^T d \qquad (4)$$

4. NEAREST CENTROID ERROR CLUSTERING

The performance of any classification system depends largely on a robust fine-tuning algorithm. In this study the fine-tuning stage was designed through a novel method called "NCCÓ[16]. Figure 3 shows an example of misclassified data in a 3-class synthesized system after training (without NCC) at 94%. We note that the larger centroids encompass more area and at the same time are less "preciseÓ and rougher than the smaller centroids which tend to evolve around class boundaries. It is also observable that errors occur mostly between class boundaries that are either overlapping or close together. By exploiting this tendency for multi-class systems and placing additional more localized, smaller and essentially "finerÓ centroids at those problematic areas, it may perhaps be possible to improve the performance of the network since these new finer centroids will have more flexibility in modeling more intricate pattern spaces due to their limited activation range.

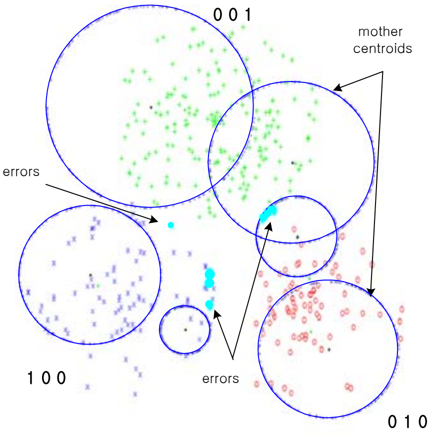

Figure 3. Initial training with a 6 centroid RBFN

The NCC algorithm is built on this idea and automatically determines the locations of "problematic areasÓand adds new centroids in those areas during the training phase. This is achieved using information from "motherÓcentroids (original 6 centroids in this example) and spawning new smaller "childrenÓ centroids by determining the nearest misclassified sample to a mother centroid (see figure 4). The mother centroids' spreads are initially inherited by the children centroids and used as a guide as they are already "roughly tuned.ÓAlthough just blindly increasing the number of centroids is an option for possibly improving performance, there is no consideration of error feedback.

The NCC method can be summarized as follows i) selection of closest error pattern (p) to a mother centroid, ii) inheritance of the mother centroid's spread (σ_j) by error sample (p), iii) finding any existing siblings (misclassified samples) encompassed by the inherited spread satisfying the general case hyperellipsoid (5), iv) computing new centroids and spreads for children via arithmetic mean (6) of its members (if no members are found spread is scaled according (7)), v) repeating the process until all error patterns are analyzed vi) reinitializing weights using (4).

$$\left(\frac{(q_m^{(1)} - q_p^{(1)})}{\sigma^{(1)}} \right)^2 + \left(\frac{(q_m^{(2)} - q_p^{(2)})}{\sigma^{(2)}} \right)^2 + \ldots + \left(\frac{(q_m^{(N)} - q_p^{(N)})}{\sigma^{(N)}} \right)^2 \leq 1 \qquad (5)$$

$$\boldsymbol{\mu}_p = \frac{1}{M} \sum_{m=1}^{M} \mathbf{e}_m, m \in \{members\ \ of\ \ new\ \ child\ \ centroid\ \ p\} \qquad (6)$$

$$\sigma_p = \alpha \cdot \sigma_j^{mother}, \text{ where } 0 < \alpha < 1 \qquad (7)$$

Figure 4 depicts this idea – the new child centroid has one "siblingÓand the original 6 centroid system becomes a 10 centroid system with 100% correct classification in figure 5. This additional fine-tuning method could be applied a number of times until a desired performance is achieved. However, due to the characteristics of the algorithm, there will usually be an increase in the number of total centroids, resulting in over-fitting issues.

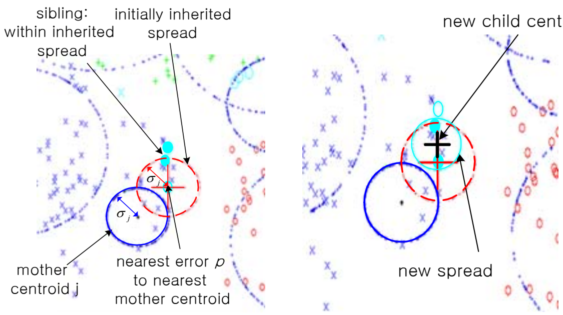
sibling: within inherited spread
initially inherited spread
new child centroid
σ_j
mother centroid j
nearest error p to nearest mother centroid
new spread

Figure 4. Spawning a new child centroid

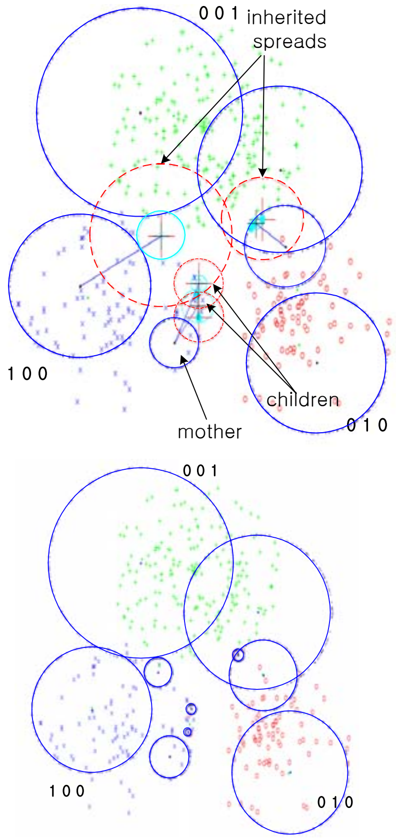

Figure 5. Spawning of centroids and final result (100%)

A partial solution to over-fitting the pattern space was to include the option to exclude “single-memberÓchildren centroids in the retraining process as these tended to address only very localized and specific error patterns. This methodology lessened the overall increase of centroids (generally desirable) by allowing only those new centroids to survive that had at least two error patterns members (one sibling) associated with it.

5. CLASSIFICATION RESULTS

The networks were trained using 80% of the total 829 samples and cross-validation performance was assessed using 20% of the remaining samples. Each new training/classification session was subjected to a random pattern shuffling scheme.

5.1. Family and Individual Instrument Classification

The best performance for family recognition was approximately 88% for RBFNs (figure 6) and 85% for EBFNs. For individual instrument classification network performance for RBFNs and EBFNs were 71% and 67% respectively. The French horn was the main cause for performance degradation with a 32% success rate as shown in figure 7.

	Strgs.	Wwinds	Brasses	%
Strgs.	**278**	10	4	**96**
Wwinds	24	**261**	4	**90**
Brasses	22	24	**202**	**81**

Figure 6. Confusion matrix for instrument family

	eb.	vln.	vcl.	vla.	clar.	flute	obo.	bsn.	horn	trp.	trb.	tuba	%
eb.	9				1								*90*
vln.		83	4	9	4	3	1			1			*79*
vcl.	1	8	75	10	1		2			1	4		*68*
vla.		14	16	44				1					*59*
clar.		4	1		82	11	1			1			*82*
flute		7	3	2	9	73	4		1				*74*
obo.		4			9	4	38						*69*
bsn.			3		1	1		29			1		*85*
horn	1	3	5		1	6	10	3	18	2	2	5	*32*
trp.		1	3	4	2	1				67			*86*
trb.		4	5	1	1	5	1	1	4	7	53		*65*
tuba											3	29	*91*

Figure 7. Confusion matrix for individual instruments

6. DISCUSSION

The initial studies with RBFN/EBFN when used with NCC present a possible approach to automatic timbre recognition with substantial increased performance – in some cases up by 25% when trained with NCC [16]. Generally speaking more centroids increase performance but at the same time decreases generality while larger centroids are most suitable for clustering larger pattern spaces and smaller centroids helpful in modeling finer pattern spaces. Furthermore, the location of each centroid is critical in obtaining satisfactory results which is extremely difficult to determine without fine-tuning and is often subject to guess-work during the network training phase. Hence, the number, choice, and location of a centroid is essential in improving RBF/EBF network performance which the NCC method achieves automatically by increasing the number of centroids and taking into consideration the characteristics of a pattern space as well as the roughly tuned centroids already in the system.

Although other neural network-based systems reported higher success rates in the 90-percentile range such as one cited by Herrera-Boyer [12], the number of instruments (4) and examples (240) used to evaluate their system seems less-than-ideal. Another ANN study [6] reported seemingly impressive results with 94~100% accuracy for individual instruments but only 40 samples and 10 classes were employed. Yet another study with neural networks reported 97% accuracy for classifying bass trombone, trombone, English horn, and contra bassoon [13]. However, pitch information was provided to the system and training and cross-validation patterns came from the same stereo audio file – one channel for training the other for cross-validation. On the other hand other types of systems such as k-NN based models [9, 10, 15, 8] reported 50.3% (1338/23), 68% (1300/23), 70% (1023/15), and 80% (1498/30) (samples/classes) respectively which are more akin to the rates obtained with this neural network model (pitch information was provided in [8]).

Although the results have not yet been thoroughly analyzed, it can be clearly noted that the system does not excel with French horn timbres (32%) although on the average it performed better for brass instruments than strings and woodwinds. This trend has not been observed in other reports. The deficiency may perhaps lay in the use of a number of new features (LPC noise, harmonic slope, harmonic expansion/contraction, spectral flux shift [16]) and a different sound library previously not used by other researchers. However, it can also be observed from the confusion matrix that the majority of errors for the French horn (76%) occur outside the brass family which is not necessarily an undesirable result as it is generally more difficult to differentiate "within-familyÓ instruments from each other than "cross-familyÓ instruments. Finally, this neural network system performs similarly or outperforms human counterparts in comparable testing environments reported in [17] (46%~67%), [5] (72%), and [4] (85%), for 27, 6, 4 instruments respectively.

7. ACKNOWLEDGEMENTS

Tae Hong Park would like to thank Paul Lansky for his insights, valuable help, and support with this research.

8. REFERENCES

[1] Berthold, M. R. "A Time Delay Radial Basis Function Network for Phoneme RecognitionÓ, *Proceedings of the IEEE International Conference on Neural Networks, vol. 7*, 1994.

[2] Best Service – Sounds & More. Hanauer Stra§e 91a, 80993 München, Germany.

[3] Birgmeier, M. "Nonlinear Prediction of Speech Signals Using Radial Basis Function NetworksÓ, *EUSIPCO*, vol. 1, 1996.

[4] Brown, J.C., Houix, O., McAdams, S. "Feature Dependence in the Automatic Identification of Musical Woodwind InstrumentsÓ, *Journal of the Acoustical Society of America*, 2001.

[5] Campbell, W. C., Heller, J. J. ÔÔThe Contribution of the Legato Transient to Instrument Identification,'' *Proceedings of the Research Symposium on the Psychology and Acoustics of Music*, 1978.

[6] Cemgil, A. T. and Gürgen, F. "Classification of Musical Instrument Sounds using Neural NetworksÓ, *Proc. of SIU97*, 1997.

[7] Er, M., Wu S., Lu J., Toh H. "Face Recognition with Radial Basis Neural NetworksÓ, *IEEE Transactions on Neural Networks*, Vol. 13, No. 3, 2002.

[8] Eronen, A., Klapuri, A. "Musical Instrument Recognition using Cepstral Coefficients and Temporal FeaturesÓ, *Proceedings of the ICASSP*, 2000.

[9] Fujinaga, I. "Machine Recognition of Timbre using Steady-State Tone of Acoustical Musical InstrumentsÓ, *Proceedings of the ICMC*, 1998.

[10] Fujinaga, I., MacMillan, K. "Realtime Recognition of Orchestral InstrumentsÓ, *Proceedings of the ICMC*, 2000.

[11] Haykin, S. *Neural Networks: A Comprehensive Foundation*, Macmillan, 1994.

[12] Herrera-Boyer, P., Amatriain X., Batlle E., Serra X. "Towards Instrument Segmentation for Music Content Description: a Critical Review of Instrument Classification TechniquesÓ, *International Symposium on Music Information Retrieval*, 2000.

[13] Kostek, B. "Soft Computing in Acoustics: Applications of Neural Networks, Fuzzy Logic and Rough Sets to Musical AcousticsÓ, Physica-Verlag, 1999.

[14] Mak, M.W., Allen W. G., Sexton G. "Speaker Identification using Radial Basis FunctionsÓ, *The 3rd IEEE Int. Conf. on Artificial Neural Networks*, 1993.

[15] Martin, K. D., Kim, Y. E. "Musical Instrument Identification: A pattern-recognition approachÓ, *Proceedings of the 136th meeting of the Acoustical Society of America,* 1998.

[16] Park, T. H. *"Towards Automatic Musical Instrument Recognition",* Princeton University, Music Department, Ph.D. Dissertation, 2004.