CLASSIFICATION OF INFANT CRY VOCALIZATIONS USING ARTIFICIAL NEURAL NETWORKS (ANNs)

M. Petroni[1], A.S. Malowany[1], C.C. Johnston[2], B.J. Stevens[3]

[1] Department of Electrical Engineering and Center for Intelligent Machines (CIM), McGill University, 3480 University Street, Montréal, Québec, Canada, H3A 2A7, e-mail: marco@cim.mcgill.ca
[2] School of Nursing, McGill University, 3506 University Street, Montréal, Québec, Canada, H3A 2A7
[3] Faculty of Nursing, University of Toronto, 50 St. George Street, Toronto, Ontario, Canada, M5A 1A1

ABSTRACT

The analysis of infant cry vocalizations has been the focus of a number of efforts over the past thirty years. Since the infant cry is one of the only means that an infant has for communicating with its care-giving environment, it is thought that information regarding the state of an infant, such as hunger or pain, can be determined from an infant's cry. To date, research groups have determined that adult listeners can differentiate between different types of cries auditorily, and at least one group has attempted to automate this classification process. This paper presents the results of another attempt at automating the discrimination process, this time using artificial neural networks (ANNs). The input data consists of successive frames of one of two parametric representations generated from the first second of a cry following the application of either an anger, fear, or pain stimulus. From tests conducted to date, it is determined that ANNs are a useful tool for cry classification and merit further study in this domain.

1. INTRODUCTION

Most parents learn to distinguish between the different types of cries of their infant, and in so doing, can determine whether their child is angry, hungry, uncomfortable, or in pain. Once the state of their child is determined, the parent can then take the necessary steps to tend to the child's needs. In a clinical setting, however, an abnormal cry can be an indicator of genetic or pathological problems and in the latter cases, the rapid identification of infants who are said to be "at risk" can lead to a faster, and, hopefully, more successful treatment which will enable the development of these infants to proceed along a normal path.

The analysis of the infant cry has been the subject of a number of research efforts over the past thirty years. Although the art of so-called diagnostic listening dates back to the days of ancient Greece and Hippocrates, this art was essentially ignored until in the mid 1800's [1]. At that time, Charles Darwin treated the topic of diagnostic listening in reference to infant crying and screaming quite comprehensively using a series of photographs and drawings to illustrate various expressions of emotion [2].

Almost a century after Darwin first penned his observations, perhaps the most comprehensive treatment of infant crying was started by the research group led by Olé Wasz-Höckert [3]. For over 30 years, this Scandinavian research group has determined that a number of different cries, uttered when an infant is hungry or in pain, can be classified auditorily, and that the cries of healthy babies can be distinguished from those that have pathological conditions, central nervous system disorders, or genetic disorders both auditorily and visually through the use of spectrograms [4].

These and other studies have also attempted to identify the distinguishing features in the cry signal which would eventually enable the classification to be done automatically by a computer system, so as to be useful in a clinical setting [5, 6, 7]. The additional information derived from a more detailed analysis of this readily available signal and from parameters derived from this signal, using superior analysis and classification techniques better suited to this class of signal, could potentially be useful for eventual diagnosis of pathology or for identifying the potential of an infant at risk.

Although a number of groups have attempted to identify the distinguishing characteristics between different types of cries, few have gone on to use these features to automate the classification process. Recently, one group has attempted to identify an infant's so-called "level-of-distress", an indicator intended to mimic adult perceptions of the aversiveness of a specific cry, using hidden Markov models (HMMs) that are based on "cry phonemes" [8]. However, this work has not attempted to specifically identify infant state, pathological, or genetic disorders from the cry. If other groups have attempted automating the classification or discrimination process between different types of cries, their results have not been documented in the literature.

The method presented in this paper uses the robustness, flexibility, and generalization properties of artificial neural networks to perform classification of infant pain cries, a domain where neural networks have not been used in the past. In this first stage of tests, different parameters derived from the input signal serve as inputs into a given neural network, with different network architectures and learning paradigms being investigated as well, in order to determine the combination of input parameters and network architectures which are best suited for the correct classification of different cry types. It is hoped that just as a parent can learn over time to differentiate between the different types of cries uttered by their infant, an artificial neural network will also be able "learn" to differentiate between different types of cries.

0-7803-2431-5/95 $4.00

2. METHODS

2.1. Data Set

Two hundred and thirty cry episodes were recorded at the Montreal Children's Hospital from sixteen healthy 2-6 month old infants who had no history of perinatal or postnatal complications. The cry episodes were the result of one of three stimulus situations: *pain* or *distress* from routine immunization; *fear* or *startle* from a jack-in-the-box; and *anger* or *frustration* from a head restraint. The recordings in this data set were made on a Sony TCM-500DEV cassette recorder with an omni-directional Senheiser MKE 2 microphone placed 10 cm from the infant's mouth. The audio signals were then low-pass filtered to 8000 Hz and sampled at 16 kHz using a 12-bit analogue-to-digital converter prior to their transfer on a SPARC 10+ for subsequent analysis.

Since important events in the cry signal are thought to occur in the first second of the utterance following the onset of the cry after the stimulus event [7], the first second of cry utterances lasting at least 0.75 seconds after the cry onset were used for the subsequent feature extraction data set to be used in the classification experiments. Of the 238 recordings in this data set, 195 had vocalizations with durations that satisfied this criterion. The other 37 recordings were discarded from the study.

2.2. Parametric Representations

Using the research done on the parametric representation of speech signals for the purposes of speech recognition as a starting point for similar representations of infant cry vocalizations for classification purposes, the following two feature sets were extracted from the signal; 10 mel-cepstrum coefficients [9], and 19 filter-band energies per input frame of cry utterance data, with the filters spaced according to the "Bark" or "mel" scale [10].

The utterances were segmented into successive 256-sample (16 ms) frames with successive frames overlapping by 50%. Consequently, for a 1 second portion of the cry utterance, 125 frames of feature vectors would be generated. If a given utterance lasted less than 1 second, the last vector of values was repeated from the end of the utterance to fill the remaining empty vectors. These two parametric representations were then normalized and scaled to values between ± 1 in order to decrease their dynamic range, and then used as inputs into the different neural network architectures for training and testing.

3. ANN ARCHITECTURES

Four neural network architectures were investigated in this study. First, simple feed-forward neural networks, using both full and tessellated connections between adjacent layers [11], were trained and tested using the static input patterns described above. In order to determine if time information would be useful for the purpose of classification, recurrent neural networks were used [12]. For this set of tests, both the size and overlap of the input data frame were varied in order to determine the "granularity" of the input parameters which would yield the best results for this application. Note that larger frames with less overlap between the frames give a *coarser* time representation, whereas smaller input frames with more overlap between the frames yields a *finer* time representation.

Time-delay neural networks (TDNNs) [13] were also tested in order to determine if the work done to model the dynamic nature of speech using this architecture could be useful for the classification of infant cry signals. In this set of experiments, a number of parameters in the neural network were varied, such as the input delay length size, and the hidden layer configuration as well.

The fourth neural network architecture trained and tested was a cascade correlation neural network [14]. This paradigm "grows" a hidden layer based on the network error of a previous training session. This particular architecture was selected in order to determine if this application could benefit from the use of a method that continues to add hidden layer nodes until the overall error of the network drops below a predefined value.

In order to train and test the above paradigms, three different public domain neural network software simulators were used. Version 6.0a of the Aspirin/Migraines package [15], developed at the MITRE Corporation, was used to train feed-forward networks, with both full and tessellated connections, and recurrent networks. Version 3.1 of Xerion, a package developed at the University of Toronto [16], was used to simulate feed-forward networks with full connections and cascade correlation nets. The other package used was version 3.2 of the Stuttgart Neural Network Simulator, which was developed at the University of Stuttgart in Germany [17]. This software was used to create and simulate both time-delay and cascade correlation neural networks.

4. EXPERIMENTAL RESULTS

Since a comprehensive and complete presentation of all the results obtained for all the tests performed, including a fair treatment of error analysis, would be quite prohibitive for inclusion into this paper, only the best results for the two different parameter sets used on the four different architectures are presented in tables 1 and 2, which are explained below. All the artificial neural networks were trained to distinguish between anger, fear, and pain, so that every neural network trained had three outputs, one used for each of the three stimulus events in the data set. Due to the large input frame size for most of the neural network architectures and of the small number of input data files available, the strategy of 10-fold cross validation was used to partition the data into ten mutually exclusive sample sets, nine of which were used to train the network and the remaining set of files used as the test set [18]. This process was repeated until each of the sets were

used as the test set at least once, in order to perform an error rate estimation which is as close as possible to an unbiased estimator of the true classification rate.

4.1. Mel-Cepstrum Coefficients Results

The feed-forward and cascade correlation networks using the mel-cepstrum data set as inputs, all had 1375 input nodes, corresponding to 125 vectors of 11 mel-cepstrum coefficient elements. The number of input frames corresponds to a 1 second segment of a cry vocalization and the 11 element vector corresponds to 10 mel-cepstrum coefficients augmented by the total energy of the frame.

For the feed-forward network, both for full and tessellated connections between the adjacent layers, both the hidden layer number and size were varied. In table 1, the column labeled FF reports the results for a fully connected network having 45 hidden nodes for this input data set. Column FT displays the results of a feed-forward neural network with tessellated connections grouping 25 mel-cepstrum vectors ([25 × 11]), with an overlap of 10 vectors in subsequent groupings making for a hidden layer consisting of 23 nodes.

The RNN column of table 1 reports on the best results obtained for a recurrent neural network. This network had an input frame size of 75 mel-cepstrum vectors, with an overlap of 66% between subsequent input frames, three delay units on each of the input and output nodes, and containing 36 hidden layer nodes.

The time-delay neural network which gave the best result for this input set is reported under column TDNN in table 1. This ANN had a delay length of 105 vectors, representing a rather large delay length, considering that the total input delay length is 125 vectors. The hidden layer width of this TDNN consisted of 5 nodes.

Column CC corresponds to the best results for a cascade correlation network. After training was completed for this ANN, 69 hidden layer units had been created.

4.2. Mel Filter-Band Energies Results

For the mel filter-band energy parameter representation of the signal, the 19 filter-band energies per window were augmented by an additional value containing the total energy for that window, so that in all, 125 vectors, each comprised of 20 mel filter-band energy values, were used, corresponding to 2500 input nodes.

In tests involving feed-forward networks, both for full and tessellated connections between adjacent layers in the network, both the number of hidden layers and number of nodes in the layers were varied, as was the case for the mel-cepstrum coefficients.

In table 2, the column labeled FF reports the best results for a fully connected network. This network has 25 hidden nodes. Column FT displays the results of a feed-forward neural network with tessellated connections grouping 20 mel filter-band energy vectors ([20 × 20]), with an overlap of 15 vectors in subsequent groupings making for a hidden layer consisting of 22 nodes.

Mel-Cepstrum Coefficients

ANN	FF	FT	RNN	TDNN	CC
Anger	74.1%	70.7%	85.2%	70.3%	29.6%
Fear	12.5%	12.5%	12.5%	0.00%	0.00%
Pain	90.4%	84.0%	74.4%	64.8%	49.6 %
Overall	79.4%	74.7%	72.3%	61.0%	40.0 %

Table 1: Correct Classification Rates Versus ANN Architecture for Mel-Cepstrum Inputs

Mel Filter-Band Energy Coefficients

ANN	FF	FT	RNN	TDNN	CC
Anger	85.2%	85.2%	48.1%	55.5 %	18.5 %
Fear	12.5%	0.00%	6.25%	0.00%	0.00 %
Pain	83.2%	75.2%	88.8%	72.0%	84.0%
Overall	77.9%	71.8%	70.7%	61.5%	59.0%

Table 2: Correct Classification Rates Versus ANN Architecture for Mel Filter-Band Energy Inputs

The RNN column of table 2 reports on the best results obtained for a recurrent neural network. This network had an input frame size of 75 mel filter-band energy vectors, with an overlap of 66% between subsequent input frames, three delay units on each of the input and output nodes, and containing 18 hidden layer nodes.

Column TDNN reports one the time-delay neural network which gave the best results for this input set. This ANN had a delay length of 105, given a total delay length of 125 vectors, and a hidden layer width of 10 nodes.

Column CC corresponds to the best results for a cascade correlation network. After training was completed for this ANN, only 4 hidden layer units had been created.

5. DISCUSSION

From the results obtained from both parametric representations displayed in table 1 and 2, the following can be stated. For both cases, feed-forward neural network architectures yield the highest correct classification rates with fully-connected networks giving slightly better results than with what could be achieved using tessellated connections. Insofar as time information is concerned, it would seem that the recurrent neural networks perform better than time-delay neural networks. This would imply that the "stricter" encoding of sequential and time-dependent information inherent in the time-delay neural network architecture is not very useful for cry discrimination. This observation is intuitive if one thinks of the type of information contained in both cry signals and in speech signals.

Since speech is defined in terms of phonemes, with a specific sequence of acoustic events denoting a specific phoneme, then time-delay neural networks are an effective architecture for capturing this information in an input frame

of parameters. For neonates, however, vocal tract shape is affected by a number of physiological or psychological effects, which may not be under the direct volitional control of the infant. Consequently, the *occurrence* of specific acoustic events in cries of the same class would seem to be more important than the *sequence* in which these events occur.

That being said, it is understandable that recurrent networks with the large time granularity fares better than the time-delay neural network, since the former encodes time information on a more general level than the sequential information encoded by a TDNN. As well, since it would seem that the occurrence of acoustic events is more relevant than the sequence with which these events occur, feed-forward neural networks yield better results than their time-dependent counter parts, with the full-connection of nodes in adjacent layers, which are capable of computing more complex relations between the inputs than with sparser connections, thus yielding better results.

Also, it would appear the the best of the cascade correlation results does not yield impressive results, and hence this application would not seem to benefit from a paradigm which can grow its own hidden layer.

Lastly comparing the results between the two parametric representations, it can be observed that the results for both feed-forward networks yield comparable results, with the mel-cepstrum achieving slightly better results and using a smaller input pattern size. However, the mel filter-band energy values would seem to give better results for TDNNs and for cascade correlation nets than the mel-cepstrum inputs.

6. CONCLUSION

This paper has presented the development and application of artificial neural networks for the classification and discrimination between three types of infant cries; anger, fear, and pain. This study represents the first attempt at the application of ANNs in the domain of infant cry classification. While the results are somewhat preliminary in nature, they do indicate that ANNs are indeed useful for this application, with fully connected feed-forward networks giving the best results, and cascade correlation networks giving the worst results. Further tests should focus on using different features and different neural network paradigms. As well, future work will expand the study to include premature infants in this study.

7. REFERENCES

[1] H. L. Golub and M. J. Corwin. A physioacoustic model of the infant cry. In *Infant Crying: Theoretical and Research Perspectives*, chapter 3. Plenum Press, New York, New York, 1985.

[2] C. Darwin. *The Expression of the Emotions in Man and Animals*. J. Murray, London, 1872.

[3] O. Wasz-Höckert, J. Lind, V. Vuorenkoski, T. Partanen, and E. Valanne. *The Infant Cry: A Spectrographic and Auditory Analysis*. Spastics International Medical Publications, Lavenham, U.K., 1968.

[4] O. Wasz-Höckert, K. Michelsson, and J. Lind. Twenty-five years of scandinavian cry research. In *Infant Crying: Theoretical and Research Perspectives*, chapter 4. Plenum Press, New York, New York, 1985.

[5] F. Benini, C. C. Johnston, D. Faucher, and J. V. Aranda. Topical anesthesia during circumcision in newborn infants. *Journal of the American Medical Association*, 270(7):850–853, August 1993.

[6] B. F. Fuller. Acoustic discrimination of three types of infant cries. *Nursing Research*, 40(3):336–340, May-June 1991.

[7] C. C. Johnston and D. O'Shaughnessy. Acoustical attributes of infant pain cries: discriminating features. In R. Dubner, G. F. Gebhart, and M. R. Bonds, editors, *Proceedings of the Vth World Congress on Pain*, pages 336–340, Hamburg, Germany, August 1988.

[8] Q. Xie, R. K. Ward, and C. A. Laszlo. Determining normal infants' level-of-distress from cry sounds. In *Proc. of the 1993 Canadian Conf. on Elec. and Comp. Eng.*, pages 1094–1096, Vancouver, B.C., September 14–17 1993.

[9] S. B. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans. ASSP*, 28(4):357–366, August 1980.

[10] E. Zwicker. Subdivision of the audible frequency range into critical bands (frequenzgruppen). *J. Acoust.*, 33(2):248, February 1961.

[11] Judith E. Dayhoff. *Neural Network Architectures: An Introduction*. Van Nostrand Reinhold, New York, NY, 1990.

[12] R. L. Watrous and L. Shastri. Learning phonetic features using connectionist networks: An experiment in speech recognition. In *Proc. IEEE First Int. Conf. on Neural Nets*, volume 4, pages 619–627, San Diego, CA, June 1987. IEEE, IEEE.

[13] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. Lang. Phoneme recognition using time-delay neural networks. Technical Report TR-1-0006, ATR Interpreting Telephony Research Laboratories, October 1987.

[14] S. E. Fahlman and C. Lebiere. The cascade-correlation learning architecture. Technical Report CMU-CS-90-100, School of Computer Science, Carnegie Mellon University, August 1991.

[15] R. R. Leighton. *The Aspirin/MIGRAINES Neural Network Software: User's Manual Release 6.0*. The MITRE Corporation, October 1992.

[16] D. van Camp. *A Users Guide for the Xerion Neural Network Simulator Version 3.1*. Department of Computer Science, University of Toronto, Toronto, Canada, May 1993.

[17] A. Zell et al. *Stuttgart Neural Network Simulator User's Manual, Version 3.2*. University of Stuttgart, Institute for Parallel and Distributed High Performance Systems, Stuttgart, Germany, 1994.

[18] S. M. Weiss and C. A. Kulikowski. *Computer Systems That Learn: Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning, and Expert Systems*. Morgan Kaufmann Publishers Inc., San Mateo, CA, 1991.