Classification of Infant Cries Using Acoustic Features

F.K. Morgan-Curtis and D.M. Wilkes

KEYWORDS. cries, features, Matlab, NICU, pain, classification

BRIEF. Identifying an acoustic feature combination that can accurately differentiate between infant cries.

ABSTRACT. The prevalence of nurse burnout and the lack of a standardized pain management system in the Neonatal Intensive Care Unit (NICU) indicate the necessity of a quantitative approach to classifying infant cries. In this work, we have found distinct acoustic features of infant cries and show how these features can be used to discriminate between two differing types of cries. Forty-seven infant cries were recorded and processed, eliciting 127 unique features. Twenty-five of the recorded cries resulted from a heelstick (pain stimulus) while the remaining 22 resulted from the cap used for a preERP being placed on the infant's head (agitating stimulus). The 127 features included transition parameters (timing of phonation), power spectral density (PSD; the power distribution in correlation to frequency), and interval length Probability Density Functions (PDF; the length of phonation). Several feature sets were created using statistical and classification analyses. A multi-step testing process, including 50/50 comparison and cross-validation, was then used to determine the optimal feature combination. The optimal feature combination had accuracy percentages above 70%. These results could possibly contribute to future NICU diagnostic technology.

INTRODUCTION.

When an infant is first born, it is admitted into at least the first level of the Neonatal Intensive Care Unit (NICU). It is crucial for NICUs to have the highest standards of care. However, NICUs can be a stressful, painful, and traumatic experience for many infants, where nurse burnout and a lack of standardized pain management procedures put the health of newborns at risk. Nurse burnout often results in detachment and emotional exhaustion, which can lead to the misclassification of pain levels and overall comfort levels of their patients in the NICU [1]-[2]. Additional difficulties are caused by nurses in the NICU having to use qualitative pain scales to quantify infant pain levels. Studies attempting to find the best methods of pain treatment for infants and neonates rely on these qualitative scales, potentially missing many signs of discomfort which leads to less accurate protocols to manage pain [3]-[7]. An infant's cry is its only means of communication; though largely undetectable to the human ear, each cry conveys a significant amount of information that can be used for diagnostic purposes. Developing quantitative approaches to classify pain levels beyond the current qualitative approaches could result in more accurate diagnoses and pain protocols, and potentially reduce the prevalence of nurse burnout.

Research done into finding a quantitative method of infant pain has included using wavelets and statistical models and tests, such as Krippendorf's alpha [8]-[9], though none of these methods have produced usable methods. This project was modeled after fairly successful research using acoustic features to diagnose depression [10]. The purpose of this project was to find a set of features from the cries of infants that would be able to accurately classify a baby's cry as being induced by a heelstick (pain stimulus) or by a preERP cap being placed on the baby's head (non-pain stimulus). Classifying these two cries is the first step in creating technology that will be able to classify hundreds of different cries, allowing a more detailed assessment of the state an infant is in and allowing the research necessary for better pain management protocols in the NICU. Current acoustic diagnostic processes require infants be brought to specialists who record the infant cry, create a spectrogram from that cry (Figure 1.), and then manually find certain acoustic features that they correlate to different issues with the child [11], a process too long for the emergency setting of the NICU. The overall goal of this research is to be able to identify what level of pain an infant in the NICU in a timely manner to ensure they are getting the necessary treatment and not being under-(or over-) medicated. The obvious first step would be to establish which features best identify the difference between a cry that is not pain versus one that is, creating a basis upon which further research may be built.

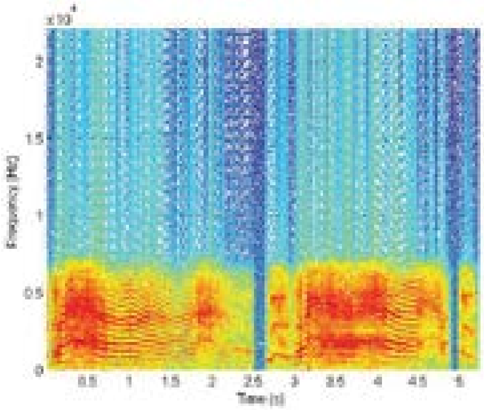


Figure 1. Example of a spectrogram that could be used to diagnose an infant by their cry using the current method of qualitative analysis. The x-axis presents time in seconds, the y-axis presents the frequency in Hertz, and the color represents the altitude of the cries.

MATERIALS AND METHODS.

For this study, the term "feature set" corresponded to the sets derived from the individual feature testing performed and the term "feature combination" corresponded to the combinations of the acoustic features created from each of the feature sets using the 50/50 comparison test (Figure 2.). The testing process used to find the most accurate feature combination follows chronologically the subheadings of this section in the order presented and is also presented in the flow chart in Figure 2.

Database Creation.

Twenty-five heelstick-induced cries and 22 preERP-induced cries were recorded and acoustic features were extracted from these recordings prior to this study. Since there is no common agreement on the most effective features to use for this classification process, a large range of features were used. In total, 127 acoustic features ranging from phonation timing to spectrum-based were extracted using Matlab, their efficacy was analyzed, and they were assigned numbers.

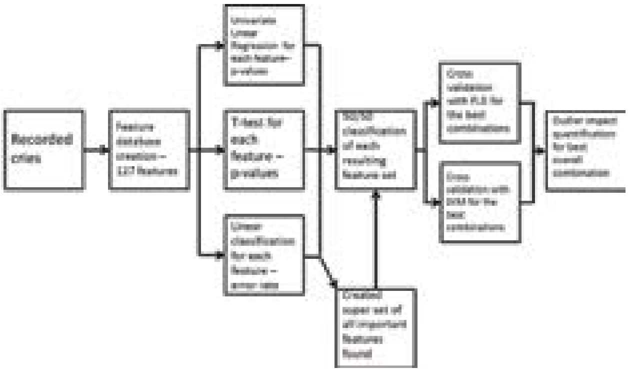

Recorded cries
Feature database creation – 127 features
Univariate Linear Regression for each feature – p-values
T-test for each feature – p-values
Linear classification for each feature – error rate
50/50 classification of each resulting feature set
Created super set of all important features found
Cross validation with PLS for the best combinations
Cross validation with SVM for the best combinations
Outlier impact quantification for best overall combination

Figure 2. A graphic of the methodology followed in this study.

Creating Feature Sets.

Using an unpaired t-test, differences in individual features were examined between the preERP and heelstick groups. The top ten features with the lowest p-value comprised a single set. Additionally, a univariate linear regression was performed prior to this study on each feature to examine the relationship between features. A second set was created from the top ten features with the lowest p-values as determined by a univariate linear regression test. Error rates from individual feature classifications were calculated with a linear classifier using commercial software, Matlab(*Mathworks*, Natick, MA, United States). These error rates correlated to the accuracy the individual feature had in classifying the heelstick data and preERP data. A set was created from the top ten features with the lowest error rates. A final set was created by taking all optimal features, determined in previous tests, without any redundancies.

Creating Feature Combinations using 50/50 Comparison Tests.

To understand the effect of total number of features on the error rate, combinations for each feature set were generated by selecting K features, where K = 1 to 10, and the efficacy of each combination produced was tested using a 50/50 cross-validation comparison where 50% of all of the data was used as training to predict the 50% of the data used as testing (the remaining 50% of the data). The classification error rates were found for every feature combination, where the classification error rate is 1 minus the sum of all the instances where the cry was correctly identified divided by the total number of cries. The combination with the lowest classification error rate was identified for each feature set.

Testing Optimal Feature Combinations Using Cross Validation.

The top feature combinations for each set (for this study, 4 combinations were tested) were subjected to a cross validation test in Matlab. The cross validation test randomly selected 70% of all of the data as training to test the remaining 30% of the data. The cross validation was repeated for 300 trials and the overall accuracy rate for each combination was calculated by averaging the accuracy rate for each of the 300 trials. The cross validation test used a Fisher's Linear Discriminant (FLD) classifier, allowing a more accurate separation and classification of the data. The FLD classifier projects feature set data to a surface, or line in a two dimensional case, that has been optimized for the best separation between classes. The feature combination set with the highest values overall was labeled as the optimal feature set.

Identifying Optimal Feature Combinations with Non-linear Classifiers.

Each feature combination tested using the cross validation with FLD was additionally run through the cross validation code that instead had a nonlinear classifier in Matlab. This verified that a different combination using a nonlinear classifier did not produce higher accuracy rates than the most accurate combination determined by the linear classifiers. Using a nonlinear classifier called a support vector machine (SVM) with a radial basis function (RBF), the nonlinear classifier accuracy rates were compared to results from the cross validation with FLD. SVM with RBF uses a hill and valley method of classification that is able to classify clustered data where classes cannot be separated linearly. The feature combination with the highest values overall was labeled as the optimal feature combination.

RESULTS.

Optimal Feature Set.

Four feature sets were created using the top ten optimal features from each efficacy test (the lowest p-values from the unpaired t-test and the univariate linear regression, the lowest individual linear classification error, and a super set of all features found in the other three sets). The average individual classification error rate of the lowest individual error set was 29.8 ± 1.838%.

Creating Feature Combinations.

Each set was tested for all values of K, the total number of features used in a combination. All four feature sets were run through a 50/50 comparison test. As the value of K grew, the classification error was reduced until K = 9 where the classification error began to increase again. Even though the error rate continues to decrease after K=5, a cap was put at K=5 for feasibility purposes (i.e., to avoid over-modeling the data). The most accurate feature combination came from the super set using a total of 5 features, including 4, 17, 24, 36, and 49 which all are acoustic features pertaining to different timing aspects of phonation, and had an error rate of 8.51%, as seen in Table 1.

Table 1. Resulting Accuracy/Error Percentages
Accuracy and error percentages resulting from the various tests in the methodology.

Combination Method	50/50 Error	50/50 Accuracy	CV FLD Accuracy	CV SVM Accuracy
T-test	17.02%	82.98%	58.0%	56.14%
Linear Regression	10.64%	89.36%	70.86%	66.71%
Lowest Error	14.89%	85.11%	56.29%	64.90%
Super Set	8.51%	91.49%	72.86%	56.24%

Testing Optimal Feature Combinations Using Cross Validation.

Cross validation was used to determine and ensure the efficacy of the best feature combinations for all four feature sets created by the 50/50 comparison test along with testing the accuracy and capabilities of the classifier used. The most accurate combination of 5 features for this test was found to be from the super set, which had an accuracy rate of 72.86%, as seen in Table 1.

Identifying Optimal Feature Combinations with Non-linear Classifiers.

The same optimal combinations (as determined by the 50/50 comparison test) that were run through the cross validation test with FLD were then run through a modification of the cross validation test using the non-linear classification SVM with RBF algorithm. The optimal feature combination using this approach was the feature combination from the univariate linear regression feature set, which had an accuracy rate of 66.71%, as shown in Table 1. Even though the univariate feature combination's accuracy rate was higher than the super set's for this test, the super set's accuracy rate from the FLD cross validation test was higher than that of the univariate's for the SVM with RBF cross validation test, so the super set combination was the most accurate feature combination.

DISCUSSION.

Using a multistep process, the optimal feature combination was found from 127 acoustic features for classification between two classes of infant cries. Features that were distinguishable between heelstick and preERP cries, were unique among all features, and had the lowest individual classification error were found and used to generate sets of 10 features. Linear and nonlinear classification methods were performed and compared for all feature sets. The optimal, five-feature combination was identified.

Optimal Feature Set.

The most effective approach to generating feature sets was the super set approach, which combined the feature sets of the three identifying methods. The next most effective method for finding optimal features would be from the univariate linear regression approach. The third most effective method was using the linear classifier approach. The least effective feature identifying method was the t-test.

Creating Feature Combinations.

The optimal feature combination each feature set produced had high accuracy rates, all higher than 80%. The most optimal feature set was the super set, producing a combination that had an error rate of only 8.51%. Other than timing features being the ones predominantly apt at classification, no trends regarding the reason the super set worked the best were noted.

Testing Optimal Feature Combinations Using Cross Validation.

The cross validation test helped to identify which feature set combinations were the most reliable and could consistently provide accurate classifications and to ensure the classifier was working optimally with the data. The t-test feature combination and the lowest error rate feature combination produced accuracy rates less than 60%, as seen in Table 1. The super set feature combination and the linear regression feature combination produced higher results, with all values over 75%, with the super set feature combination having the highest values overall. These results were used in addition to the others to identify the optimal set, especially since the accuracy rates only differed by 15% in addition to possible errors in the classification.

Identifying Optimal Feature Combinations with Non-linear Classifiers.

SVM can help to classify data that is clustered or set in "islands" (one data set is surrounded by the other). This SVM with RBF was incorporated into the previously used cross validation (substituting for the FLD classifier). Increased accuracy rates for the lowest individual classification error combination using the SVM non-linear approach indicate that feature data for that specific combination is best classified non-linearly, as shown in Table 1. Despite the increased accuracy rates for the linear regression and lowest error rate feature sets, FLD classification using the super set feature combination had the best overall accuracy. Data is typically best separated by a linear classifier, but nonlinear classifiers help to ensure optimal combinations.

CONCLUSION.

The goal of this project was to identify a feature set that could accurately and effectively classify a heelstick cry and a preERP cry. The results gathered over the course of this project will contribute to the possible creation of a technology that can differentiate between different baby cries and help Neonatal Intensive Care Unit (NICU) medical staff identify what an infant wants or needs. This technology would also help improve current studies that are attempting to standardize and find better ways to manage pain in the NICU and also in pediatrics generally [3]-[7].

The best feature combination to differentiate between a heelstick-induced cry and a preERP-induced cry, as determined by the results found in this project, included features 4, 17, 24, 36, 49, which are all features pertaining to varying aspects of the timing of phonation, and was obtained from the super set of features. According to a study looking at Krippendorf's alpha trends in different infant cries, this might be because the initial cry caused by the heelstick is one of startle or surprise rather than of pain itself, contaminating the actual pain cries and making it more difficult to identify the cries of pain [8]. Future studies should break the cry into segments, rather than using the entirety of the cry, which could allow for increased accuracy. In addition, an exhaustive search could be completed to find the best feature set by running all possible combinations through the cross validation code. However, this approach is computationally intensive. In addition, this feature database needs to be continued through further research.

ACKNOWLEDGMENTS.

The author would like to thank the School for Science and Math at Vanderbilt for opportunity to conduct this research. The author would also like to thank the Vanderbilt University School of Engineering for providing to tools, databases, and software to conduct this research.

REFERENCES .

1. C. Maslach, S. E. Jackson, The measurement of experienced bur out. *Journal of Occupational Behaviour*. 2, 99-113 (April 1981).

2. L. LaGasse, A. R. Neal, B. M. Lester, Assessment of infant cry: acoustic cry analysis and parental perception. *Mental Retardation and Developmental Disabilities Research Review Journal*. 11, 83-93 (2005).

3. S. Suraseranivongse et al, A comparison of postoperative pain scales in neonates. *British Journal of Anaesthesia*. 97, 540-544 (June 2006).

4. M. Campbell et al, Trial of Repeated Analgesia with Kangaroo Mother Care (TRAKC Trial). *BMC Pediatrics*. 13 (November 2013).

5. N. McIntosh, Pain in the newborn, a possible new starting point. *European Journal of Pediatrics*. 156, 173-177 (February 1997).

6. S. A. Furdon, V. C. Pfeil, K. Snow, Operationalizing Donna Wong's principle of atraumatic care: pain management protocol in the NICU. *Pediatric Nursing*. 24, 336-342 (August 1998).

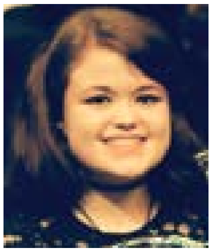
7. American Academy of Pediatrics, Committee on Fetus and Newborn and Section on Surgery, Section on Anesthesiology and Pain Medicine, Canadian Paediatric Society, Fetus and Newborn Committee, Prevention and management of pain in the neonate: an update. *Pediatrics*. 118 (November 2006).

8. T. Etz, H. Reetz, C. Wegener, F. Bahlmann, Infant cry reliability: acoustic homogeneity of spontaneous cries and pain-induced cries. *Speech Communication*. 58, 91-100 (March 2014).

9. J. Saraswathy, M. Hariharan, Thiyagar Nadarajaw, Wan Khairunizam, Sazali Yaacob, Optimal selection of mother wavelet for accurate infant cry classification. *Australian physics and engineering science in medicine*, 37, 439-456 (April 2014).

10. D. J. France, R. G. Shiavi, S. Silverman, M. Silverman, D. M. Wilkes, Acoustical properties of speech as indicators of depression and suicidal risk. *IEEE Transactions on Biomedical Engineering*. 47, 829-837 (July 2000).

11. B. Reggiannini, S. J. Sheinkopf, H. F. Silverman, X. Li, B. M.Lesterb, A flexible analysis tool for the quantitative acoustic assessment of infant cry. *Journal of Speech, Language, and Hearing Research*. 56, 1416-1428 (October 2013).



Fea Morgan-Curtis is a senior at Hume Fogg Academic High School in Nashville, Tennessee; she attended the School for Science and Math at Vanderbilt.