Journal of Theoretical and Applied Information Technology
31st January 2023. Vol.101. No 2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

NOISE-ROBUST IN THE BABY CRY TRANSLATOR USING RECURRENT NEURAL NETWORK MODELING

MEDHANITA DEWI RENANTI[1] , AGUS BUONO[2] , KARLISA PRIANDANA[3] , SONY HARTONO WIJAYA[4]

[1]Doctoral Study Program of Department of Computer Science Bogor Agricultural University, Software Engineering Technology of College of Vocational Studies Bogor Agricultural University, Indonesia
[2,3,4]Department of Computer Science Bogor Agricultural University, Indonesia

E-mail: [1]medhanita@apps.ipb.ac.id, [2]agusbuono@apps.ipb.ac.id, [2]karlisa@apps.ipb.ac.id, [2]sony@apps.ipb.ac.id

ABSTRACT

The development of baby cry translators is still uncommon in Indonesia. Hence, this research aims to improve the Madsaz application as a noise-robust baby cry translator. Despite its success in translating the Dunstan Baby Language version of baby cry, the Madsaz application encounters a problem of decreased accuracy by up to 30% due to noise. Therefore, the objective of this research is to solve this problem using recurrent neural networks as deep learning modeling with the input of representative feature extraction. This modeling can classify and resolve noise in the dataset. This research utilizes an architecture modification of recurrent neural networks, i.e., the gated recurrent unit (GRU) and long short-term memory network (LSTM). This research also employs the Mel-Frequency Cepstrum Coefficient (MFCC) as a feature extraction method and the Dunstan Baby Language version of baby cry as the dataset. An experiment was carried out in two scenarios, namely input data without noise and input data with noise. The results show that the accuracy levels of the GRU and LSTM methods are 94% and 91%, respectively, on data without noise. On the other hand, the accuracy of data with noise is decreased by 5%, from 94% to 89%, in the GRU method, but decreased by 34%, from 91% to 57%, in the LSTM method. Hence, this finding indicates that the GRU method is more noise-robust, specifically against 5 to 20 dB of noise, compared to the LSTM method. In terms of effectiveness, the GRU is equivalent to the LSTM. However, the GRU method has more computational efficiency due to its simple network structure and fewer trained parameters, making it ideal for situations with small amounts of data and present noise.

Keywords: *Baby cry translator, GRU, LSTM; MFCC; Noise-robust; Recurrent neural network*

1. INTRODUCTION

The baby cry is a means of communication between babies and adults to meet their needs. Parents who have just had a baby will find it difficult to interpret their baby's cry [1]. On the other hand, expertise in translating a baby's cry in Indonesia is still uncommon [2]. In some cases, the mother's stress level increases in after giving birth to her first child; hence, social support is highly needed [3]. Dunstan Baby Language (DBL) is a language used to understand the cries of babies from all countries and ethnic groups [4]. DBL applies to infants up to three months of age because, after that age, babies will develop their communication skills according to their environment. Research has shown that 90% of mothers worldwide who follow DBL feel satisfied, 70% are more confident, and their stress levels are reduced [2].

Expertise in identifying the meaning of the DBL version of the baby's cry still needs to be studied. Initially, this identification is carried out by certified experts, such as experts or pediatricians, who are present when the baby is crying and conduct the identification. This technique requires a long time to identify the baby's cry and is considered less efficient. In addition, experts also provide training or seminars to prospective parents regarding DBL as guidance to identify the meaning of their baby's crying. This method is effective for parents who have attended the training, even though there are still trainees who cannot fully understand. The problem arises from the inability of parents to access the training related to DBL, while experts can only be present at certain times.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

The machine learning method can automate the DBL so that the classification process can be carried out anywhere and anytime, and the need for experts can be minimized. This method has also been widely used to identify and classify baby sounds through healthy and normal babies' cries, including ensemble learning, neural networks, and Hidden Markov Model (HMM) methods. Interpretation of babies' cries with the feature selection and classification process using the ensemble learning method after the feature selection stage. This method detects and analyzes discomfort signals automatically, which repeatedly affect 20–25% of newborns. This method creates a consistent dataset of babies' cries and selects sound features that match experimental results, which promise to improve infant monitoring in real-world settings [5].

The Dunstan Baby Language (DBL) automation has been conducted using a codebook, and Mel Frequency Cepstrum Coefficients (MFCC) have been successfully created with an accuracy of 94%. This method is used to classify the DBL version of baby cry [6]. This research develops an android-based application [7] named Madsaz 2.0, which has been downloaded in 175 countries. However, the application indicates a weakness in the presence of noise. The accuracy drops to 30% when using new data containing noise. Based on user reviews, the previously developed Madsaz application helps parents be more confident in raising their children and be more agile about the steps needed when their baby cries. Identification of the DBL version of baby cry has been carried out by combining Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) which simultaneously take roles in feature extraction and classifier methods. CNN acquires prominent features from the raw spectrogram data, whereas RNN acquires the obtained features' temporal data. In the dataset of the Dunstan Baby Language version, the model of CNN-RNN surpasses the prior technique by up to 94.97% in average classification accuracy. The promising result shows that using CNN-RNN and the five-fold cross-validation yields reliable and robust results. This method has not been subjected to noise testing [8].

Deep learning methods for classifying, detecting, predicting, and handling noise problems have gained attention because they can increase accuracy and reduce processing time despite the presence of noise. Deep neural networks (DNN) are part of deep learning, in which RNN is the most popular network and has been widely used for classification. It can be calculated using sampling delay differential equations, which simulate processes in physics, biology, and neural networks [9]. RNN is utilized in time-series and sequential data applications. Therefore, enhancing RNN models algorithmically and reducing memory access overhead are critical for achieving high performance [10]. This research examines a recurrent neural network with a vector of time-varying thresholds that can resolve linear programming problems. The suggested recurrent neural network has asymptotic stability and the capability to find suitable solutions to linear programming problems in general. The design of an analog circuit based on op-amps for implementing the recurrent neural network is explained [11].

A significant number of current studies have sought to differentiate between snoring and non-snoring. Nonetheless, developing a common reference point to define snoring is exceptionally challenging because the snoring episodes (SE)'s length, frequency, and duration vary depending on the individual. Hence, the classification algorithm of learning-based snoring is required to identify distinct snoring patterns and noise for various individuals. This research proposed a recurrent neural network (RNN)-based classification method that can identify SEs and non-snoring episodes (NSEs) by recognizing individuals' characteristics of SEs and NSEs, monitored in their routines on the recording device. Even though the suggested RNN-based classification method is trained using a limited dataset, an accuracy of 98.9% is evident [12]. This potential end-to-end (E2E) design, the RNN transducer (RNN-T), may become the alternative for automated speech recognition and replace the popular hybrid model. This research focuses on the RNN-T model development with lower GPU memory usage during training, a better initialization strategy, and advanced encoder lookahead modeling. The enhanced RNN-T model outperforms a highly qualified hybrid model with more precise recognition accuracy and reduced latency when trained with 65,000 hours of anonymized Microsoft's training data [13].

The crux of music data can be retrieved using musical tags. The current music auto-tagging methods typically include stages of preprocessing (extracting features) and machine learning. Nonetheless, most existing methods fail from either information loss or inadequate features in the

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

preprocessing stage. In contrast, the machine learning stage strongly relies on the extracted features in the preprocessing stage, resulting in the inability to utilize information. This problem can be resolved by deep recurrent neural networks (RNN) with scattering-transformed inputs. The scattering transform, which works as the first stage, recovers features from raw data while keeping more data than older approaches such as the Mel-frequency cepstral coefficient (MFCC) and the Mel-frequency spectrogram. The second stage of the algorithm utilizes a five-layer RNN with the gated recurrent units (GRU) and a sigmoid output layer. On the MagnaTagATune dataset, an experiment was carried out by measuring the area under the ROC-curve (AUC-ROC) to determine its architecture's efficiency. The experimental results indicate that the suggested strategy can improve tagging performance compared to state-of-the-art models. Furthermore, the architecture generates shorter training times and lower memory utilization [14]. The RNN method outperforms the Autoregressive Integrated Moving Average (ARIMA) model in traffic flow prediction. This research uses Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU). The application of the GRU method to traffic flow prediction is still new [15].

A feature extraction process is conducted before the classification process. MFCC feature extraction is preferred for its function to accurately extract features. The results of MFCC feature extraction become the training input of the backpropagation neural network. The training results demonstrate that it identifies 98.9% accuracy [16]. It is also proposed to use feature extraction to identify segments of features that should be shifted as diverse distortions, such as audio signal processing disrupts the audio signal [17]. The comparisons of feature extractions of Linear Prediction Cepstral Coefficients (LPCC) and Perceptual Linear Prediction Coefficients (PLP) in experiments that determine the sex of the speaker exhibit the best results of 99.37% for 16 MFCC coefficients [18]. MFCC performance outperforms LPCC feature extraction on speaker verification systems that use fixed phrases. The error rate of the MFCC method is 0% in this verification system [19]. In addition, the selection of feature extraction used in the formulation and application of voice commands using MFCC and HMM can produce 93.89% accuracy on average in the noise-free scenario and 58.1% in the noise-filled scenario [20].

The problem of clinical depression or major depressive disorder (MDD) can be solved using RNN and MFCC feature extractions. Speech can diagnose depression and forecast its severity level using RNN. A multimodal and multifeatured experiment is also used to assess the performance of the suggested method. High-level MFCC features contain depression-related information. Combining facial action units and auditory features increases classification accuracy by 20% and 10%, respectively, to 95.6% and 86% [21].

Based on the previous studies, this research proposes recurrent neural network modeling, namely LSTM and GRU, to be applied to baby cry data without noise and with 5 dB to 20 dB of noise, and feature extraction using MFCC. This method accommodates noise problems; hence, it is expected to be robust against noise. The noise is formed using a Gaussian SNR. The use of SNR is based on the results of a related study [22], which presents an algorithmic method for estimating the signal-to-noise ratio (SNR) for a linear system from the received signal's single realization. Moreover, it is assumed that a Gaussian matrix with a one-sided left correlation characterizes the linear system. The unidentified signal and noise inputs can be selected from any distribution and are considered freely and equally distributed with zero means. This linear model's ridge regression function is utilized with methods and tools developed from the theory of random matrices to provide an accurate assessment of the SNR in closed form without previous statistical data on noise or signal. The advised method is highly accurate, according to the simulation findings.

Subsequently, the research aims to model the recurrent neural network to develop a robust-noise baby cry translator. The benefit of this research is that it helps parents interpret their infant's cries and provide the right solution accordingly. Therefore, it is expected to reduce parental stress because the baby calms down more quickly. The scope of the research covers the following:

1. Classification of baby cry according to the Dunstan Baby Language version, divided into groups of babies who are hungry and tired, want to burp, experience pain (lower gas) in the stomach, and feel discomfort.
2. Baby cry detection between the ages of 0 and 3 months.

2. LITERATURE REVIEW

Journal of Theoretical and Applied Information Technology
31st January 2023. Vol.101. No 2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

2.1. Baby Cry

Baby cry can be defined as a complex action involving components of movement, facial expressions, and sounds. It is commonly perceived as a negative emotion. The duration and frequency of the baby's cries vary. The cries tend to increase until six weeks of age, then decrease gradually. In addition, babies cry more frequently at night over a 24-hour cycle [23]. Baby crying is classified as a type of speech and communication. Speech sometimes transforms its signals into another type to facilitate human communication [24] [25] [26]. Several studies have broken down sound signals into smaller parts called phonemes. Various methods are used to evaluate each piece of information in the voice signals [25] [26] [27].

2.2. Dunstan Baby Language

Dunstan Baby Language (DBL) is a language used to understand baby cry aged 0–3 months, applicable to all countries, ethnicities, and languages. There are five baby language versions of DBL, namely [2]:

1. "Neh" means hunger.
   When feeling hungry, the baby produces a "neh" sound. "Neh" is defined as the sound produced when the baby tastes while breastfeeding. Recognition of the "neh" sound by hearing the insertion of the letter N in the cry. Other signs are:
   - Moving the tongue to the roof of the mouth (taste);
   - Sucking fingers or the head;
   - Licking his lips;
   - Shaking his head left and right.
2. "Owh" signifies tiredness, which indicates a sleepy baby.
   "Owh" is the sound you make when you yawn. Other signs are:
   - The baby begins to move restlessly.
   - The baby begins to rub his eyes and scratch or pull his ears.
   - The baby begins to squirm and arch his body.
3. "Eh" means wanting to burp.
   The "eh" cry occurs when the baby's chest is working hard to let out the gas. Usually, the frequency with which the "eh" cry is pronounced is faster and shorter as the baby tries to burp. Other signs are:
   - Tightened chest.
   - Wriggling movements when placed in bed.
   - Stop drinking milk and get restless.
4. "Eairh" indicates lower gas in the baby's stomach. The "eairh" cry occurs because of lower gas in the baby's stomach that causes pain (colic). Other signs are:
   - The baby's legs twitch and are pulled to the stomach.
   - The baby's body becomes stiff.
   - The baby screams in pain.
5. "Heh" means discomfort.
   One reason for baby fussing is because they feel discomfort. This situation may be caused by a wet diaper, too hot or cold temperature, or something else. The "heh" cry is usually breathless (like exhaling air), and there is an emphasis on the letter H at the beginning of the word.

2.3. Voice Recognition of Baby Cry

Voice recognition is divided into two types, i.e., speech recognition and speaker recognition. Speech recognition is the process of identifying sounds based on spoken words. The level of sound suppression is the first parameter to be compared, which is then aligned with the accessible database format. On the other hand, speaker recognition, refers to a voice recognition system based on the person speaking [28].

The system of speech recognition includes recognizing baby crying sounds. In general, the two main modules for recognizing a baby's cries are feature extraction and feature matching. The technique of collecting data from a sound stream to represent each speaker is known as feature extraction, whereas recognizing voices by comparing the voice feature extraction with previously known voice characteristics is called feature matching [29].

2.4. Signal Transformation into Information

A signal is a measurement that varies in accordance with time, space, or other variables. It is characterized mathematically as a function of one or more independent variables. The following is an example of a function that describes two signals (equation 1 and equation 2): the first is a linear function with the independent variable t (time), and the second is a quadratic function with t [30].

$$s_1(t) = 5t$$
$$s_2(t) = 20t^2 \quad (1)$$

Another example is as follows:

$$s(x,y) = 3x + 2xy + 10y^2 \quad (2)$$

Journal of Theoretical and Applied Information Technology
31st January 2023. Vol.101. No 2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

The function displays the signals of two independent variables, x and y, which can be represented in two spatial coordinates in a plane. In some cases, the function relating time to the signal quantity is unknown or so complex that its application is impractical, as in the case of the sound signal shown in Figure 1. The signal cannot be represented as in expression (1). Generally, a segment of sound is represented with high accuracy by a waveform, which is the summation of several sine functions of different amplitude and frequency (3) and is written as follows:

$$\sum_{i=1}^{N} A_i(t) \sin[2\pi F_i(t)\, t + \theta_i(t)] \quad (3)$$

where $\{A_i(t)\}$, $\{F_i(t)$, and $\{\theta_i(t)\}$ are the set of possible sine wave amplitude, frequency, and phase for any t- time. One way to represent the content information or message of a segment of a voice signal is to measure the segment's amplitude, frequency, and phase of the [30].

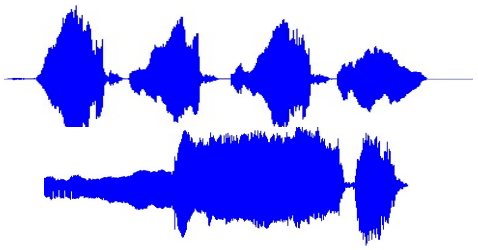

Figure 1: Example of Sound Signal

Voice signal processing is a technique of transforming voice signals into meaningful information as desired [31]. In the transformation process, some steps need to be conducted, including the digitalization of analog signals, feature extraction, and pattern recognition. The baby crying signal has the same pattern as the signal in general.

2.5. Noise

Adaptive, additive, random additive, airport, core background sound, vehicle, cross-noise, exhibition area, factory, multi-speaker sound, music, nature, office, restaurant, street, suburban train, train station, and others are some of the noise classifications. There are four primary types of noise: additive noise, interference, reverberation, and echo. They led to the advancement of acoustic signal processing, including noise reduction, sound enhancement, and echo suppression [32].

If the transmitted signals encounter an object with a characteristic impedance different from the medium, the signal will strike the object, and a reflection will occur. The reflected signal recovered by the radar consists of target echo, noise, and interference. The reflected signal recovered by the radar consists of target echo, noise, and interference. The primary goals of signal processing are [33]: suppressing all other signals except the echo of the desired target so that it can be easily detected; and gathering information about the target's behavior, such as its position, speed, and characteristics.

Signal processing has the function of decreasing these signal interferences [34], including:

- Noise, caused by electrically random particle motion, is unavoidable at all temperatures above absolute zero and is emitted by the radar receivers and, to a lesser extent, by antennas, transmission lines, and the sun as the external sources.
- Cluttera is an echo of unwanted signals from the ocean, the terrain, and the weather. It is a realistic echo signal that is typically subdued due to a distinct Doppler shift from the targeted target.
- Electronic Countermeasures (ECM) or jamming is purposeful interference intended to impede the detection of target echoes.
- Electronic interference (EMI) is accidental interference from other surrounding radars, communication systems, and nearby jammers.
- Spillover primarily takes place in the radar's continuous band and is caused by the simultaneous operation of the transmitter and receiver. This is a leak from the transmitter to the receiver.

2.6. Mel-Frequency Cepstral Coefficients

The extraction of features helps characterize the properties of voice signals. The most employed feature extraction method is Mel-Frequency Cepstrum Coefficients (MFCC) because it has been proven to have satisfactory performance and helps create human perception and compassion that take frequency into account [35]. The initial stage of MFCC is to break the amplitude value of the input signal into frames processed using a Mel-filter bank in accordance with the way humans hear [20].

In MFCC feature extraction, the Mel log energy calculation is referred to as the Discrete

Journal of Theoretical and Applied Information Technology
31st January 2023. Vol.101. No 2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Cosine Transform (DCT). Afterwards, the CNN network was used to recognize the twelve DCT main coefficients [35].

Framing is where the speech signal is blocked in the N sample frame, and the following sample is entered in an M (M > N) way. In the first N samples, the first frame is accommodated. The M sample comprises the second frame [36]. Following the addition of the first frame, the signal extends through N-M samples while the procedure is repeated until the signal becomes one or more frames (equation 4). Windowing facilitates signal discontinuity reduction at the beginning of the frame and at the end. The windowing process is adjusted to minimize the spectral distortion of the signal. The signal setting is zero at the beginning and the end of each frame.

S(a), 0 ≤a ≤ N−1.

(4)

In this equation, N refers to the sample count per frame following the marking of the window result [35]. Furthermore, the Fast Fourier Transform (FFT) aids in the conversion of the time domain in each N sample frame to the frequency domain. Through $N(X_N)$, this sample set is generated. The proposed method identifies most X_N as complex numbers [37]. The magnitude of the frequency to be determined is represented by the absolute value. The step yields the sound signal's spectrum or gram period.

Mel-frequency wrapping and its log use wrap the linear scale values in the following voice signal. Each sensitivity of the human voice has different frequencies, but not all of them are on a linear scale. The frequency of each tone in the Hz signal is measured. In the Mel-frequency scale, the logarithmic range is above 1,000 Hz, and the linear frequency range is below 1,000 Hz *[38]*.

In terms of perceived and resolved frequencies, the Mel-scale simulates human hearing. It is a unit of measurement for a tone's perceived frequency (pitch) [39] [40] [41].

DCT on MFCC Cepstrum is the final step, i.e., the output of the time log converter of Mel-spectrum. The cepstral illustration of the sound spectrum provides the most suitable representation of the sound signal's small cepstral for the selected frame examination. DCT is used to transform the Mel frequency into the time domain in accordance with the coefficient of the Mel spectrum. The first component identifies the average parameter of the input signal. This signal includes specific information about the speaker. The MFCC flow diagram can be seen in Figure 2.

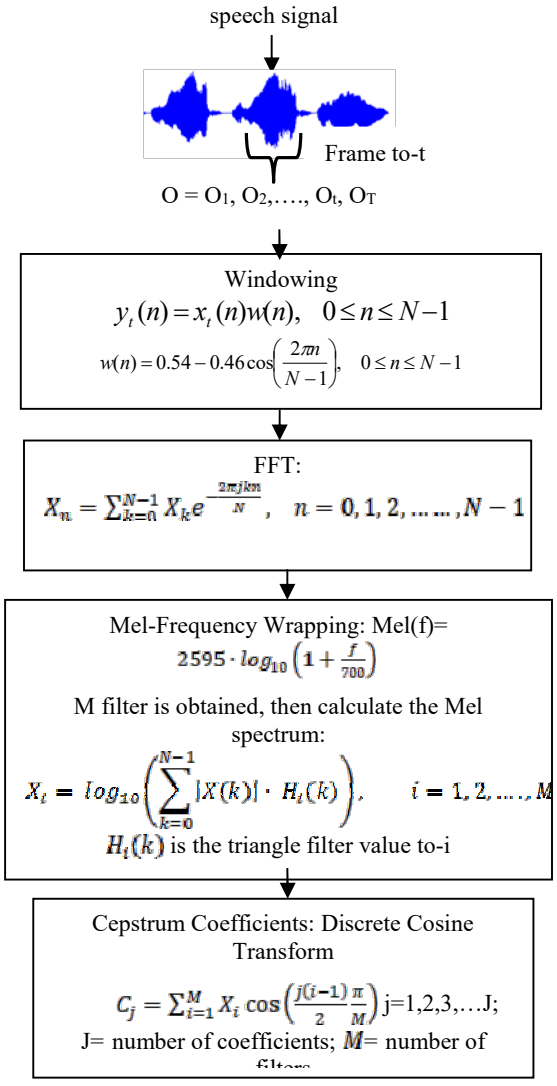

Figure: 2 The MFCC Flow Diagram [31]

2.7. Deep Learning

Artificial intelligence (AI) refers to the advancement of machines to acquire the intelligence of the human mind. In computer science, AI means the study of intelligent agents, which are devices that perceive their surroundings and take actions that maximize their chances of success in reaching a goal. In an informal context, the term "AI" is applied when machines can perform activities related to the human brain, including learning, and solving problems. One

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

significant aspect of machines is learning. Therefore, machine learning is classified as AI.

Machine learning is a field in the modern world of computing. A significant amount of research has been done to enhance machines' intelligence. Learning is a natural human behavior that has been made an essential aspect of machines. Researchers put much effort into improving the accuracy of machine-learning algorithms. Another dimension leads to deep learning, which is part of machine learning. Up to this point, deep learning applications have been explored to solve problems in the application domain and subdomain. In deep learning, the feature extraction process is combined with classification. This is to overcome the manual engineering of features, the large number of features created from the parent variable list, and the effect of human bias. Deep learning can handle a large number of layers and parameters [42].

2.8. Recurrent Neural Network

Deep learning, a subtype of machine learning, has numerous layers of neurons used with elaborate architectures or nonlinear transformations to simulate high-level data abstractions. The Recurrent Neural Network (RNN) is a deep learning architecture. RNN is a network architecture superset of feed-forward neural networks that are widely used in handling sequential data in text, audio, and video. RNN performs better sequence or time-step data processing than feed-forward neural networks. The simple recurrent network passes activation via solid edges as in the feed-forward network at each time-step T. At time T, dashed edges connect the source node J', i.e., J'(T), to the target node at the next time-step J(T+1) [43].

The sequence of sounds in a baby's cry within a certain length of time is one consideration in choosing RNN as the deep learning model. However, one of the weaknesses of RNN is the problem of vanishing gradient; it does not have long-term memory. Subsequently, to overcome the situation, the long short-term memory (LSTM) and the gated recurrent unit network (GRU) methods are proposed [44].

2.9. Long Short-Term Memory

Long Short-Term Memory (LSTM) was proposed to overcome the vanishing gradient problem in 1997 [45]. There is a memory unit in the LSTM network architecture to store previous learning, even if it has been long. The LSTM is primarily an RNN with long-term memory in the form of weights. Each LSTM cell has an internal state, a constant error carousel, an input node, and multiplicative gating (input gate, output gate, and forget gate) [43].

Each memory cell with a linear activation function has an internal state called the core. A constant error carousel is an internal state with a self-connected (recurrent) weight and allows an error to circulate across time steps without disappearing. The input node is responsible for receiving input, whereas the input gate accepts input and multiplies it by the input node. The input value can only be stored in the cell state if the input gate allows it. The output gate is used to multiply the internal state when deciding the conditions and values for the next hidden state and the current output node. Moreover, the forget gate determines whether information should be stored or forgotten.

2.10. Gated Recurrent Unit

The Gated Recurrent Unit (GRU) was also proposed to overcome the vanishing gradient. However, the architecture of the GRU is more straightforward than the LSTM [45]. The GRU does not embed an output layer like the LSTM, which results in a lower number of trained parameters. Unlike the LSTM, which has three gates for input, output, and forget, there are only two gates in the GRU, i.e., the reset and update gates. The reset gate serves to regulate the amount of past information to be forgotten or the combination of new input and the condition or value of the previous state stored in memory. In addition, the update gate plays the same role as the input and forget gates in the LSTM, i.e., determining which information needs to be stored and forgotten.

3. EXPERIMENTS

3.1. Data

The dataset consisted of 175 records taken from the Dunstan Baby Language version of baby cry. The records were classified into five classes consisting of 140 training and 35 testing data records. Hence, each class has 28 training and seven testing data records.

Journal of Theoretical and Applied Information Technology
31st January 2023. Vol.101. No 2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

among the fully connected layers. The activation

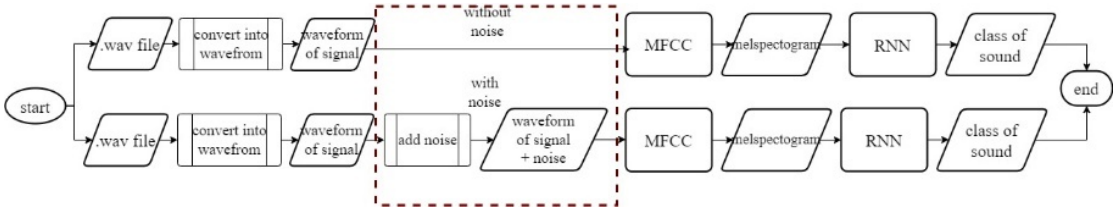

Figure 3: Research Flowchart

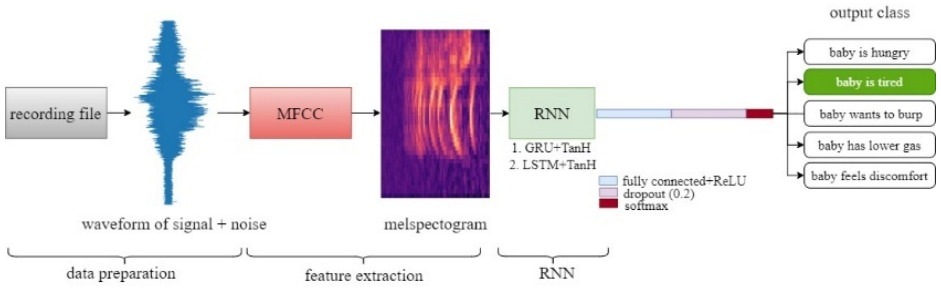

Figure 4: Architectures of RNN (LSTM and GRU) on Data with Noise

3.2. Experimental Setup

The research was conducted in two scenarios, i.e., input data without given noise and input data with given noise. Figure 3 displays the research flowchart, and Figure 4 illustrates the research flow data with given noise in the architecture of RNN, the LTSM, and GRU. Each model of RNN variances was applied in both scenarios. Figure 3 also shows that the recording data was saved in .wav format. The recording file was first converted to an audio signal waveform, and then gaussian noise was added to the signal waveform for data input in the scenario with noise. The gaussian noise observed the signal-to-noise ratio (SNR) within the 5-20 dB range so that the added random noise to the signal wave was in accordance with its amplitude.

The subsequent step was feature extraction using MFCC, which accepted input in the form of signal waves added with and without noise. MFCC generated essential sound features in a Mel-spectrogram, which was then used to train the RNN classifier. Figure 4 depicts a Mel-spectrogram with the size of 64x64 and was given noise which represented key features of the recorded baby cries as input to the LSTM and GRU models. The LSTM and GRU models had the same number of layers and parameters. The architecture started with the input layers: one layer of LSTM/GRU with 128 units, two fully connected layers, and the SoftMax function, as well as a dropout layer with 0.2 rates function used tanh and ReLu for the fully connected layer. SoftMax, in this case, was used for the activation function in determining the sound class in the last layer. This research used Adam as the optimizer, sparse categorical cross entropy as the loss function, and the evaluation metrics of precision, recall, F1 score, and accuracy. With a batch size of 64, the model was trained for 100 epochs.

4. RESULT AND DISCUSSION

Experiments were conducted in accordance with the two scenarios previously mentioned. The results showed that GRU had a higher performance than LSTM. Table 1 demonstrates the evaluation results for precision, recall, F1 score, and accuracy of the LSTM and GRU models with data without noise. Table 2 displays the evaluation results of the two models on data with noise. The graph of the GRU training and validation accuracy from 1 to 100 epochs without added noise is shown in Figure 5, whereas that with noise is shown in Figure 6. Moreover, the LSTM graph of the training and validation accuracy from 1 to 100 epochs without added noise is shown in Figure 7, whereas that with noise is shown in Figure 8.

Journal of Theoretical and Applied Information Technology
31st January 2023. Vol.101. No 2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 1: Results of LSTM and GRU on Data without Noise.

Model	Precision	Recall	F1-Score	Accuracy
GRU	0.95	0.94	0.94	0.94
LSTM	0.92	0.91	0.91	0.91

Table 2: Results of LSTM and GRU on Data with Noise.

Model	Precision	Recall	F1-Score	Accuracy
GRU	0.91	0.89	0.88	0.89
LSTM	0.41	0.57	0.46	0.57

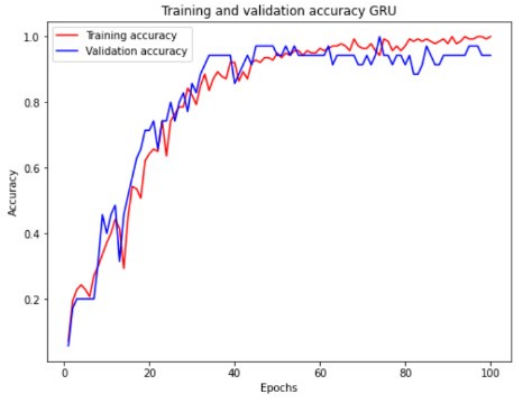

Figure 5 Graph of Training and Validation Accuracy of GRU on Data without Noise

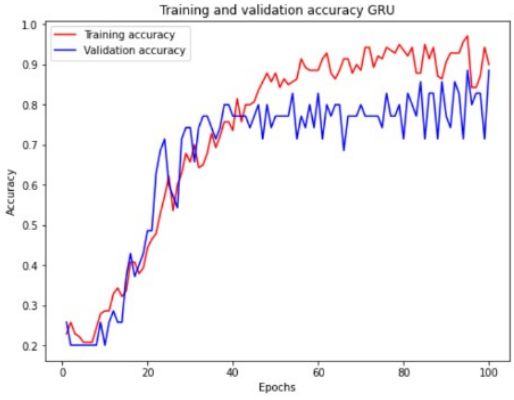

Figure 6: Graph of Training and Validation Accuracy of GRU on Data with Noise

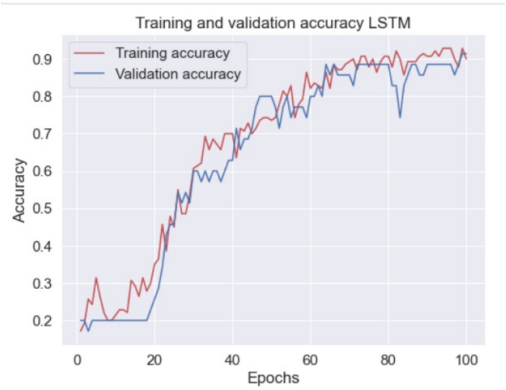

Figure 7: Graph of Training and Validation Accuracy of LSTM on Data without Noise

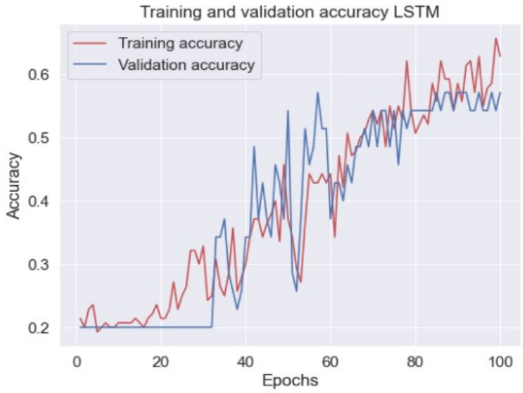

Figure 8: Graph of Training and Validation Accuracy of LSTM on Data with Noise

Table 1 and Table 2 indicate that GRU is superior to LSTM, and is also more robust against noise. The decrease in GRU accuracy is only 5% when there was noise. However, when there was noise, the accuracy of LSTM dropped by 34%. This result happens because the LSTM has a more complex architecture than the GRU. Complex architectural structures do not learn both small amounts of data and data sequences. Particularly when added to Gaussian noise, LSTM network learning has not been able to recognize prominent features and see patterns in the data. Information entered and forgotten at the input and forget gate does not work properly in the presence of noise. Most of the incoming noise cannot be forgotten by the forget gate, so the model becomes weak in generalization. In contrast, the GRU, which has a simpler network structure and a smaller number of trained parameters, is more suitable for situations with small amounts of data, when there is noise.

The results of this research support the previous studies that attempted to examine the use

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

of recurrent neural networks [14] and the novel use of the GRU method [15]. This research also confirms that the architectures of RNNs can generate shorter training times and lower memory utilization.

5. CONCLUSION

In conclusion, this research has managed to demonstrate the solution to overcome the problems of noise that decrease the accuracy of the baby cry translator application. For the baby cry data with and without noise, deep learning model of recurrent neural networks (RNN) modified as the LSTM and GRU architectures is tested. The overall results reveal that the GRU method outperforms the LSTM method. The use of baby cry data without noise in the LSTM and GRU architectures does not significantly differ in accuracy, with a 3% difference. Nonetheless, the difference becomes 32% more significant when the data is subjected to noise. Hence, it can be inferred that GRU is more robust against noise than LSTM. This finding is due to the more complex architecture of LSTM compared to GRU. Complex architectural structures do not learn both small amounts of data and small data sequences. LSTM network learning, in particular, is unable to recognize key features and observe patterns in data when combined with Gaussian noise. Information entered and forgotten at the input and forget gate does not work properly in the presence of noise. Most of the incoming noise cannot be forgotten by the forget gate, so the model becomes weak at generalizing noise. In contrast, the more straightforward network structure and a smaller number of trained parameters of GRU allow it to be more suitable for situations with small amounts of data or when noise is present.

REFERENCES:

[1] S. Kitzinger, *Understanding Your Cryring Baby*. London: Carroll and Brown Limited, 2005.

[2] A. Gunawan, "Dunstan Baby Language Indonesia," 2011. http://www.mommeworld.com/post/view/49/dunstan-baby-language-indonesia/.

[3] A. Fitriani and I. Nuryati, "Dukungan Sosial dan Tingkat Stres pada Ibu Pasca Melahirkan Anak Pertama," *J. Psikol. Malahayati*, vol. 1, no. 2, pp. 1–7, 2019, doi: 10.33024/jpm.v1i2.1856.

[4] P. Dunstan, "Open Up and Discover Your Baby'S Language," 2006. [Online]. Available: https://www.babytaal.nl/media/PDF/ComprehensiveBooklet(2).pdf.

[5] A. Osmani, M. Hamidi, and A. Chibani, "Machine Learning Approach for Infant Cry Interpretation," in *International Conference on Tools with Artificial Intelligence*, 2017, pp. 182–186, doi: 10.1109/ICTAI.2017.00038.

[6] M. D. Renanti, A. Buono, and W. A. Kusuma, "Infant Cries Identification by using Codebook as Feature Matching, and MFCC as Feature Extraction," *J. Theor. Appl. Inf. Technol.*, vol. 56, no. 3, pp. 437–442, 2013.

[7] M. D. Renanti, "Software Penerjemah Tangis Bayi Versi Dunstan Baby Language Berbasis Android," *Pros. Semin. Nas. Teknol. Terap.*, 2016.

[8] T. N. Maghfira, T. Basaruddin, and A. Krisnadhi, "Infant cry classification using CNN - RNN," *J. Phys. Conf. Ser.*, vol. 1528, no. 1, 2020, doi: 10.1088/1742-6596/1528/1/012019.

[9] A. Sherstinsky, "Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network," *Phys. D Nonlinear Phenom.*, vol. 404, p. 132306, 2020, doi: 10.1016/j.physd.2019.132306.

[10] N. M. Rezk, M. Purnaprajna, T. Nordstrom, and Z. Ul-Abdin, "Recurrent Neural Networks: An Embedded Computing Perspective," *IEEE Access*, vol. 8, pp. 57967–57996, 2020, doi: 10.1109/ACCESS.2020.2982416.

[11] J. Wang, "Analysis and Design of a Recurrent Neural Network for Linear Programming," *IEEE Trans. Circuits Syst.*, vol. 40, no. 9, pp. 613–618, 1993, doi: 10.1109/81.244913.

[12] S. J. Lim, S. J. Jang, J. Y. Lim, and J. H. Ko, "Classification of snoring sound based on a recurrent neural network," *Expert Syst. Appl.*, vol. 123, pp. 237–245, Jun. 2019, doi: 10.1016/J.ESWA.2019.01.020.

[13] J. Li *et al.*, "Developing RNN-T models surpassing high-performance hybrid models with customization capability," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 2020-Octob, pp. 3590–3594, 2020, doi: 10.21437/Interspeech.2020-3016.

[14] G. Song, Z. Wang, F. Han, S. Ding, and M. A. Iqbal, "Music auto-tagging using deep Recurrent Neural Networks," *Neurocomputing*, vol. 292, pp. 104–110, May

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

2018, doi: 10.1016/J.NEUCOM.2018.02.076.

[15] R. Fu, Z. Zhang, and L. Li, "Using LTSM and GRU Neural Network Methods for Traffic Flow Prediction," in *31st Youth Academic Annual Conference of Chinese Association of Automation*, 2016, pp. 324–328, doi: 10.1109/YAC.2016.7804912.

[16] C. K. On, P. M. Pandiyan, S. Yaacob, and A. Saudi, "Mel-Frequency Cepstral Coefficient Analysis in Speech Recognition," *2006 Int. Conf. Comput. Informatics, ICOCI '06*, no. 2, pp. 2–6, 2006, doi: 10.1109/ICOCI.2006.5276486.

[17] X. C. Yuan, C. M. Pun, and C. L. Philip Chen, "Robust Mel-Frequency Cepstral Coefficients Feature Detection and Dual-tree Complex Wavelet Transform for Digital Audio Watermarking," *Inf. Sci. (Ny).*, vol. 298, pp. 159–179, 2015, doi: 10.1016/j.ins.2014.11.040.

[18] E. Yücesoy and V. V Nabiyev, "Comparison of MFCC , LPCC and PLP Features for The Determination of a Speaker's Gender," 2014, pp. 321–324.

[19] H. Yang, Y. Deng, and H. Zhao, "A Comparison of MFCC and LPCC with Deep Learning for Speaker Recognition," *ACM*, pp. 160–164, 2019.

[20] M. Sidiq, T. A. B. W, and S. Sa'adah, "Desain dan Implementasi Voice Command Menggunakan Metode MFCC dan HMMs," in *e-Proceeding of Engineering*, 2015, vol. 2, no. 1, pp. 1362–1373.

[21] E. Rejaibi, A. Komaty, F. Meriaudeau, S. Agrebi, and A. Othmani, "MFCC-based Recurrent Neural Network for automatic clinical depression recognition and assessment from speech," *Biomed. Signal Process. Control*, vol. 71, p. 103107, Jan. 2022, doi: 10.1016/J.BSPC.2021.103107.

[22] M. A. Suliman, A. M. Alrashdi, T. Ballal, and T. Y. Al-Naffouri, "SNR Estimation in Linear Systems with Gaussian Matrices," *IEEE Signal Process. Lett.*, vol. 24, no. 12, pp. 1867–1871, 2017, doi: 10.1109/LSP.2017.2757398.

[23] R. G. Barr, M. S. Kramer, C. Boisjoly, L. McVey-White, and I. B. Plesst, "Parental Diary of Infant Cry and Fuss Behaviour," *Arch. Dis. Child.*, vol. 63, pp. 380–387, 1988.

[24] M. G. Rahim, *Artificial Neural Network for Speech Analysis/Synthesis*. London: Chapman&Hall, 1994.

[25] J. G. Ackenhusen, *Real-time Signal Processing: Design and Implementation of Signal Processing Systems*. New Jersey: Prentice-Hall, Upper Saddle River, 2001.

[26] T. E. Quatueri, *Discrete-time Speech Signal Processing: Priciples and Practice*. Prentice Hall Signal Processing Series, 2002.

[27] B. Gold and N. Morgan, *Speech and Audio Signal Processing: Processing and Perception of Speech and Music*. New York: John Wiley & Sons, Inc, 2000.

[28] W. Kurniawan, "Identifikasi Speech Recognition Manusia dengan Menggunakan Average Energy dan Silent Ratio Sebagai Feature Extraction Suara pada Komputer," *Biospecies*, vol. 9, no. 1, pp. 1–6, 2016.

[29] D. Gupta, M. R. C, N. Manjunath, and M. PB, "Isolated Word Speech Recognition Using Vector Quantization (VQ)," *Int. J. Adv. Res. Comput. Sci. Softw. Eng.*, vol. 2, no. 5, 2012.

[30] J. G. Proakis and D. G. Manolakis, *Digital Signal Processing: Principles, Algorithms, and Applications*, Ke-3. New Jersey: Prentice Hall, Inc., 1996.

[31] A. Buono, "Representasi Nilai HOS dan Model MFCC sebagai Ekstraksi Ciri pada Sistem Indentifikasi Pembicara di Lingkungan Ber-noise Menggunakan HMM," Universitas Indonesia, 2009.

[32] U. Shrawankar and V. Thakare, "Noise Estimation and Noise Removal Techniques for Speech Recognition in Adverse Environment," in *IFIP International Federation for Information Processing*, 2010, pp. 336–342.

[33] B. Edde, *Radar: Principles, Technology, Applications*. Prentice-Hall, 1993.

[34] R. Hayati and R. Kurnia, "Simulasi Unjuk Kerja Discrete Wavelet Transform (DWT) dan Discrete Cosine Transform (DCT) untuk Pengolahan Sinyal Radar di Daerah Yang Ber-Noise Tinggi," *J. Nas. Tek. Elektro*, vol. 3, no. 1, pp. 32–43, 2014, doi: 10.25077/jnte.v3n1.53.2014.

[35] M. D. Pawar and R. D. Kokate, "Convolution Neural Network based Automatic Speech Emotion Recognition using Mel-Frequency Cepstrum Coefficients," *Multimed. Tools Appl.*, vol. 80, pp. 15563–15587, 2021, doi: 10.1007/s11042-020-10329-2.

[36] T. Hakanpaa, T. Waaramaa, and A.-M. Laukkanen, "Emotion Recognition from Singing Voices using Contemporary Commercial Music and Classical Styles," *J*

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Voice, vol. 33, no. 4, pp. 501–509, 2019.

[37] Q. Li *et al.*, "MSP-MFCC: Energy-efficient MFCC Feature Extraction Method with Mixed-Signal Processing Architecture for Wearable Speech Recognition Applications," *IEEE Access*, vol. 8, pp. 48720–48730, 2020.

[38] M. Zahit, *Robot Control With Voice Command*. Istanbul: Yıldız Technical University, Department of Computer Engineering, 2008.

[39] O. Viikki, D. Bye, and K. Laurila, "Acoustics, Speech, and Signal Processing," in *Proceedings of the 1998 IEEE International Coriference 2*, 1998, pp. 733–736.

[40] M. Greenwood and A. Kinghorn, ""SUVing: Automatic Silence/Unvoiced/Voiced Classification of Speech," The University of Sheffield, UK, 1999.

[41] Z. Xuan, C. Yining, L. Jia, and L. Runsheng, "Feature Selection in Mandarin Large Vocabulary Continuous Speech Recognition," in *Signal Processing, 2002 6th International Coriference ICSP*, 2002, pp. 508–511.

[42] P. P. Shinde and S. Shah, "A Review of Machine Learning and Deep Learning Applications," in *Proceedings - 2018 4th International Conference on Computing, Communication Control and Automation, ICCUBEA*, 2018, pp. 1–6, doi: 10.1109/ICCUBEA.2018.8697857.

[43] Z. C. Lipton, J. Berkowitz, and C. Elkan, "A Critical Review of Recurrent Neural Networks for Sequence Learning," no. June, 2015.

[44] S. Yang, X. Yu, and Y. Zhou, "LSTM and GRU Neural Network Performance Comparison Study: Taking Yelp Review Dataset as an Example," *Proc. - 2020 Int. Work. Electron. Commun. Artif. Intell. IWECAI 2020*, pp. 98–101, 2020, doi: 10.1109/IWECAI50956.2020.00027.

[45] S. Nosouhian, F. Nosouhian, and A. K. Khoshouei, "A review of recurrent neural network architecture for sequence learning: Comparison between LSTM and GRU," no. July, pp. 1–7, 2021, doi: 10.20944/preprints202107.0252.v1.