Journal of information and communication convergence engineering
http://jicce.org

J. Inf. Commun. Converg. Eng. 22(2): 145-152, Jun. 2024 Regular paper

# Prediction of Closed Quotient During Vocal Phonation using GRU-type Neural Network with Audio Signals

**Hyeonbin Han[1], Keun Young Lee[2], Seong-Yoon Shin[3], Yoseup Kim[4], Gwanghyun Jo[1], Jihoon Park[5]*, and Young-Min Kim[6]***

[1]Department of Mathematical Data Science, Hanyang University ERICA, Ansan, Republic of Korea
[2]Independent scholar, Republic of Korea
[3]School of Computer Science and Engineering, Kunsan National University, Gunsan 54150, Republic of Korea
[4]Digital Healthcare Research Center, Deltoid Inc., 186 Jagok-ro, Seoul, Republic of Korea
[5]Division of Vocal Music, Nicedream Music Academy, 48 Eoun-ro, Daejeon, Republic of Korea
[6]Digital Health Research Divisions, Korea Institute of Oriental Medicine, Daejeon, Republic of Korea

## Abstract

Closed quotient (CQ) represents the time ratio for which the vocal folds remain in contact during voice production. Because analyzing CQ values serves as an important reference point in vocal training for professional singers, these values have been measured mechanically or electrically by either inverse filtering of airflows captured by a circumferentially vented mask or post-processing of electroglottography waveforms. In this study, we introduced a novel algorithm to predict the CQ values only from audio signals. This has eliminated the need for mechanical or electrical measurement techniques. Our algorithm is based on a gated recurrent unit (GRU)-type neural network. To enhance the efficiency, we pre-processed an audio signal using the pitch feature extraction algorithm. Then, GRU-type neural networks were employed to extract the features. This was followed by a dense layer for the final prediction. The Results section reports the mean square error between the predicted and real CQ. It shows the capability of the proposed algorithm to predict CQ values.

**Index Terms**: Vocal phonation, GRU, Artificial neural network, Electroglottography

## I. INTRODUCTION

Recently, attempts have been undertaken in the phonetics community to quantitatively analyze the vibratory behavior of human vocal folds during vocal phonation. A common method involves employing a windowed Fourier transform (spectrogram) to examine the audio waveform produced by the vocal activity. This technique facilitates the direct visualization of voice quality attributes such as harmonicity, kurtosis, and spectral centroid [1, 2]. Additionally, a few researchers have utilized mechanical or electrical devices to study vocal fold dynamics. Here, the focus was on metrics such as the subglottal pressure and the contact area of the vocal folds. For example, circumferentially vented Pneumotach split-flow air masks can measure pressure waveforms [3]. This enables the analysis of the nasal/oral aerodynamics. Although an air mask system is a highly effective and direct tool for voice quality evaluation, it is difficult to use. In contrast, electroglottography (EGG) provides a convenient and noninvasive technique for visualizing vocal fold vibrations during voice production [4,5]. By placing two electrodes around the vocal folds and passing a low-amperage current

**Received** 30 March 2024, **Revised** June 4 2024, **Accepted** 7 June 2024
***Corresponding Author** Jihoon Park (e-mail: nicedreamusic@gmail.com) Division of Vocal Music, Nicedream Music Academy
Young-Min Kim (e-mail: irobo77@kiom.re.kr), Digital Health Research Division, Korea Institute of Oriental Medicine

Open Access https://doi.org/10.56977/jicce.2024.22.2.145 print ISSN: 2234-8255 online ISSN: 2234-8883

near the thyroid cartilage, the variations in the vocal fold contact area can be captured during the glottal cycle. This is based on the principle that closed vocal folds allow for higher electrical admittance across the larynx, which results in a higher current between the electrodes. This method simplifies the evaluation of the vocal quality by visualizing the variations in the contact area of the vocal folds during vocal production.

One of the crucial metrics extracted from the EGG signal is the closed quotient (CQ). It represents the time ratio during which the vocal folds remain in contact throughout voice production [6,7]. Various theories have emphasized the significance of CQ in voice analysis. Typically, a higher CQ is associated with voices considered stronger or more pressed and is, attributed to the increased duration of vocal fold contact during phonation; it yields richer and more vibrant sounds. Consequently, a higher CQ was more prevalent in vocal production when the *chest* register was used than when the *head* register was used. Additionally, CQ is effective in clinical diagnostics. This is because alterations

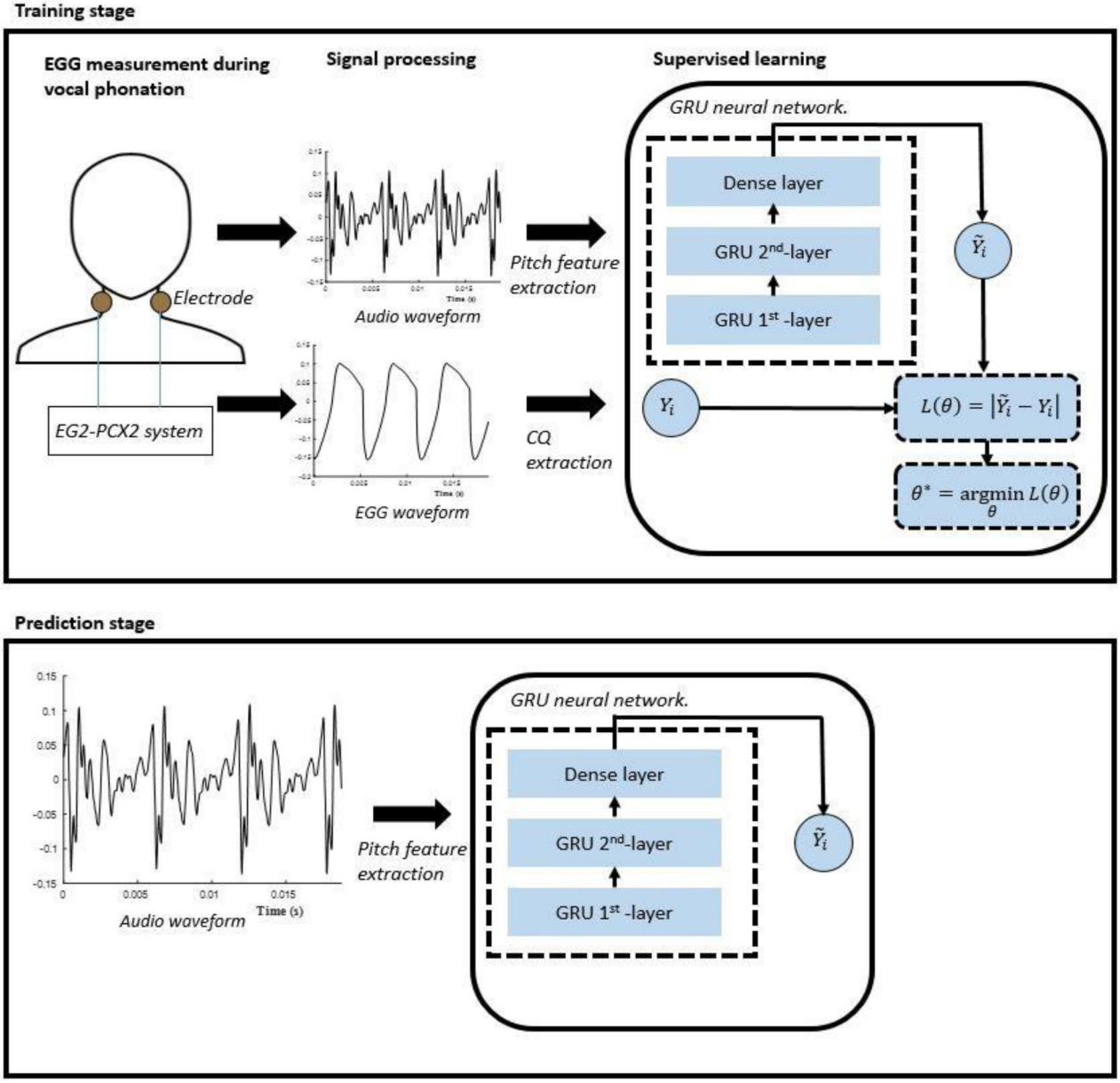



**Fig. 1.** Overall process of methods. The first step of the *training stage* is the measurement of a vocal phonation process in terms of the audio signal and EGG signal. An audio waveform is pre-processed by the pitch algorithm before being input to GRU neural network while the CQ value is extracted from the EGG waveform. By supervised learning of the GRU neural network, we obtain the prediction model for CQ value. After the model is trained via an auto-grad algorithm, the CQ value can be predicted from only the audio waveform in the *prediction stage*.

in vocal fold closure patterns owing to lesions or paralysis affect typical CQ values [8]. It is generally acknowledged that CQ decreases as the fundamental frequency increases. To summarize, the insights gained from analyzing CQ levels serve as important reference points in vocal training for professional singers.

Notwithstanding the critical role of the closed quotient (CQ) values in vocal analysis, obtaining these values conventionally requires mechanical and electrical measurements. These methods involve either the inverse filtering of airflows captured by a circumferentially vented mask or the post-processing of EGG waveforms. In this study, we introduced a novel algorithm to predict the CQ values from only audio signals. This eliminates the need for mechanical and electrical measurement techniques. Our approach began by constructing a dataset that pairs the vocal audio waveforms with their corresponding CQ values. We then developed a machine-learning algorithm that leverages supervised learning for training. Recently, significant developments have been achieved in the ANN community. For example, see [9,10] for computer vision, [11,12] for reinforcement learning, and [13,14] for natural language processing. In particular, recurrent neural networks (e.g., LSTM [15-17] and GRU [18-20]) have been demonstrated to be effective in handling time-series data. Therefore, we employed neural network architectures that incorporate gated recurrent unit (GRU) layers in the CQ prediction algorithm. To optimize the performance of the algorithm, we preprocessed the audio input using a pitch feature extraction algorithm [21,22] before substituting it into a GRU-type neural network. In the Results section, we report the performance of the proposed algorithm. For all the tests, the MSE between the predicted and real CQ values were below 8E-03. This indicated the capability of the proposed algorithm to analyze the vibratory behavior of the vocal fold contact area.

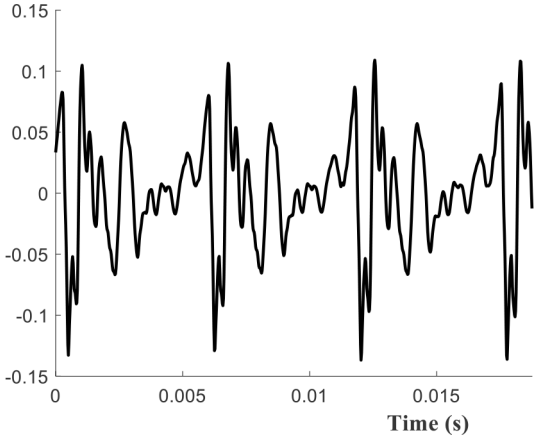
The remainder of this paper is organized as follows: In Section 2, we describe the GRU-based neural network algorithm for predicting the CQ values. The results are presented in Section 3. Finally, the conclusions are presented in Section 4.

## II. METHODS

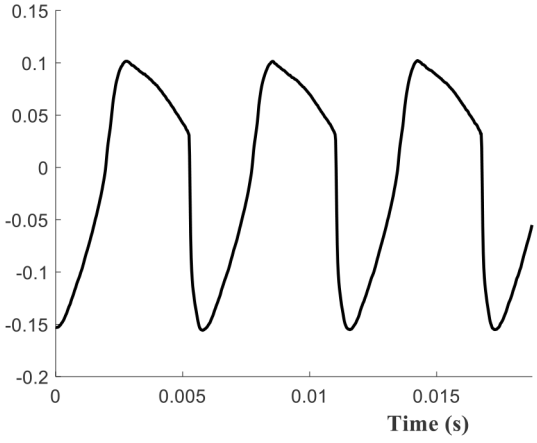
In this section, we describe the development of a novel algorithm for predicting CQ values. Utilizing a neural network based on GRU, the proposed algorithm is trained by audio and EGG signals. After the training stage is complete, our model can predict the CQ values only from audio signals. Subsection A outlines the methodology for data collection and the process of extracting the CQ value from an EGG waveform. Subsection B elaborates on the GRU-type neural networks and pitch feature extraction techniques for audio signals. The pitch algorithm reduces the length of the audio signal. GRU-type neural networks are employed to extract features from the pitch-reduced signal. This is followed by a dense layer for prediction. A comprehensive schematic of this process is shown in Fig. 1.

### *A. Data collection and CQ feature extraction*

In this subsection, we describe the data collection process and CQ extraction process from EGG signal. Data were collected from the vocal productions of the vowel /'a'/ by 22 individuals. Each of them spoke for approximately 10 s in a relaxed state. During this process, the audio signals were captured using a condenser microphone. Meanwhile, the EGG signals were recorded with two electrodes attached to the neck near the vocal folds, utilizing the *EG2-PCX2* system from Glottal Enterprises. For each individual's measured audio and EGG data, we divided the regions by 0.1 s. This yielded 2,076 samples. ($A_i$, $EGG_i$) denotes the audio waveforms and EGG waveforms for samples $i = 1, ..., 2076,$

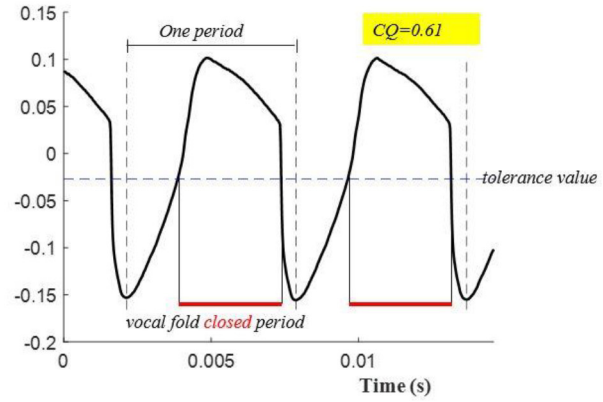



(a) Audio signal obtained by condenser microphone

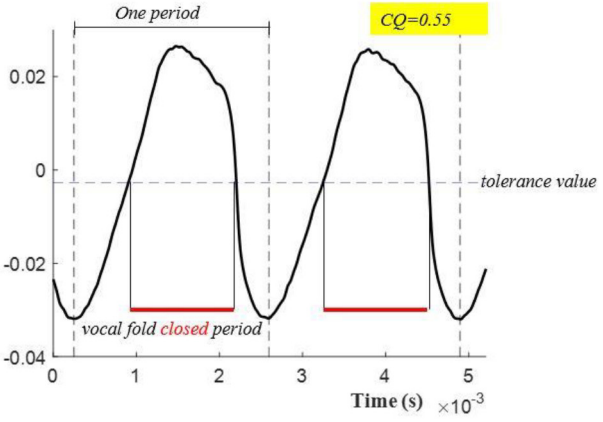



(b) EGG waveform obtained by *EG2-PCX2* system

**Fig. 2.** Typical examples of time-synchronized audio signal and EGG waveform obtained during vocal production.




(a) EGG waveform with CQ = 0.61.




(b) EGG waveform with CQ = 0.55.

**Fig. 3.** Two EGG waveforms with CQ = 0.61 (top) and 0.55 (bottom). The red line indicates the closed vocal fold sub-region where the EGG signal is above the tolerance level. The ratio of a period to the vocal-fold-closed period determines the CQ value.

respectively. Fig. 2 shows graphs of typical examples of $A_i$ and $EGG_i$. Here, we emphasize that EG2-PCX2 systems have synchronized audio and EGG waveforms during vocal phonation measurements. Therefore, the waveforms of $A_i$ and $EGG_i$ in Fig. 2 have identical fundamental frequencies with almost completely matched temporal appearances of the (local) minimum amplitudes in each period.

Now, we describe the process of extracting the CQ value from an EGG waveform (see Fig. 3). The CQ value is defined as follows:

$$CQ = \frac{\text{Length of closed vacal fold region}}{\text{Length of One Period}}. \quad (1)$$

Here, the closed vocal fold region is defined by the region where the EGG waveform is above the tolerance value (defined as 50% percent of the maximum amplitude of EGG). Based on the definitions in Eq. (1), the CQ value is between zero and one. For convenience, we denote $CQ_i$ using the CQ value extracted from $EGG_i$. Finally, we define the dataset in the form of samples $(X_i, Y_i)$. The input variable $X_i$ is an audio signal $A_i$. The target variable $Y_i$ is $CQ_i$ obtained by the extraction process for EGG waveforms.

### B. GRU-based neural network algorithm

In this subsection, we describe the development of a GRU-based neural network architecture and its training process. To obtain a high efficiency, the audio signal $X_i$ was preprocessed using the pitch feature extraction algorithm [17,18]. We briefly describe the pitch algorithm for the completeness of the work.

In general, extracting the pitch from an audio signal is identified as determining the peak of the frequency spectrum from a short-time Fourier transform (STFT). It remains to estimate accurate peaks in STFT. Paraboloid interpolation is generally applied using a quadratic polynomial near the peak. The maximum value of the quadratic polynomial is presumed to be the peak. Finally, values higher than 10% of the magnitude of the frequency spectrum were determined as the pitch. We use the notation

$$P_i = Pitch(X_i) \quad (2)$$

for the features extracted using the pitch algorithms. Here, $P_i$ represents vector-type data with a size smaller than that of $X_i$. Therefore, considering $P_i$ as a reduced version of an audio signal, we can efficiently employ a neural network algorithm.

We now describe GRU-based neural networks [16]. Given an input time series $x$, the hidden state $h_t$ of the GRU layer is obtained by sequentially computing Eqs. (3)-(6) for ($t$ = 1, 2, ..., $N$). Here, $N$ is the length of $x$.

$$r_t = \sigma(W_{ir}x_t + b_{ir} + W_{hr}h_{t-1} + b_{hr}) \quad (3)$$

$$z_t = \sigma(W_{iz}x_t + b_{iz} + W_{hz}h_{t-1} + b_{hz}) \quad (4)$$

$$n_t = \tanh(W_{in}x_t + b_{in} + r_t \oplus (W_{hn}h_{t-1} + b_{hn})) \quad (5)$$

$$h_t = (1 - z_t) \oplus n_t + z_t \oplus h_{t-1}. \quad (6)$$

Here $r_t$, $z_t$, and $n_t$ are the so-called reset, update, and new gates. $\sigma$ is the sigmoid function. An important parameter in the GRU layer is the hidden size. It indicates the number of features in the hidden state $h$. One of the advantages of the GRU layers is the convenience of stacking layers by considering the hidden state of the previous layer as an input to the subsequent layer. We introduced four types of GRU-based neural network algorithms accompanied by a pitch algorithm to predict the CQ value:

**Definition 1. GRU-based neural networks.**

*1) GRU 1L utilizes a single-layer GRU followed by a dense layer for regression.*

*2) GRU 2L employs two stacked GRU layers followed by a dense layer for regression.*

*3) BiGRU 1L uses a single-layer bidirectional GRU followed by a dense layer for regression.*

4) *CONV1D-GRU applies one-dimensional convolutional layer, followed by a single-layer GRU and a dense layer for the regression.* □

The four types of GRU-based neural networks in Definition 1 have one or two GRU layers, possibly with an accompanying one-dimensional convolutional layer. Here, all the neural networks listed in Definition 1 have a dense layer at the final stage of the CQ regression. Typically, GRU-type neural networks are trained using a backpropagation algorithm.

We are in a position to state the CQ prediction algorithm. Suppose $X_i$ is a typical audio signal. First, we apply a pitch algorithm to reduce the length of the audio signal. Here, $P_i$ is the feature extracted from the pitch algorithm. Next, by selecting a suitable type of GRU-based neural network, we obtain regression results by substituting the pitch features into the GRU-type neural network introduced in Definition 1. We use the notation $N(\cdot, \theta)$ for the neural network architecture. $\theta$ denotes the collection of parameters appearing in the GRU-type neural network. Here, we summarize the CQ prediction algorithm. We call it Pitch-GRU.

**Algorithm Pitch-GRU** (*Input:* $X_i$ *, output:* $\tilde{Y}_i$).

1. *Extract pitch features from the input audio sample*:

$$P_i = Pitch(X_i).$$

2. *Select one of the following GRU configurations to define neural network* $N(P_i, \theta)$:

*1) GRU 1L*

*2) GRU 2L*

*3) BiGRU 1L*

*4) CONV1D-GRU 1L*

3. *Predict CQ by*

$$\tilde{Y}_i = N(P_i, \theta). \quad \square$$

Note that we can alternatively substitute raw audio-signal $X_i$ directly to $N(\cdot, \theta)$:

**Algorithm 2** (*Input:* $X_i$*, output:* $\tilde{Y}_i$).

1. *Select one of the following GRU configurations to define neural network* $N(P_i, \theta)$:

*1) GRU 1L*

*2) GRU 2L*

*3) BiGRU 1L*

*4) CONV1D-GRU 1L*

2. *Predict CQ by*

$$\tilde{Y}_i = N(X_i, \theta). \quad \square$$

Pitch-GRU is the primary proposed method. Algorithm 2 is used for a comparison to emphasize the enhanced performance of the pitch algorithm.

## III. RESULTS

In this section, we report the performance of the Pitch-GRU algorithm. Note that Algorithm 2 (which does not use pitch extraction) was used for the comparison.

Data were collected from the vocal productions of the vowel /'a'/ by 22 individuals. This yielded 2,076 audio and EGG pairs of samples. These samples were divided into training, valid, and test sets consisting of 1,230, 412, and 434 samples, respectively.

For all the tests, the hidden sizes in the GRU layers were set to 10. The losses were defined as the mean squared error (MSE) between $Y_i$ and $\tilde{Y}_i$. The GRU-based neural networks were trained using Adam optimization [23] with a learning rate of 0.1. All the tests were conducted using an NDIVIA RTX A5000 instrument.

**Table 1.** Number of parameters and training/validation/test losses, and CPU time of Pitch-GRU with different neural networks

| Model | Parameter | Training Loss | Validation loss | Test Loss | CPU time |
|---|---|---|---|---|---|
| GRU 1L | 401 | 7.6E-3 | 7.1E-3 | 8.0E-3 | 99.3 s |
| GRU 2L | 1061 | 7.7E-3 | 6.8E-3 | 7.6E-3 | 107.4 s |
| BiGRU 1L | 801 | 7.7E-3 | 5.1E-3 | 8.0E-3 | 104.8 s |
| CONV1D-GRU | 8627 | 7.5E-3 | 7.3E-3 | 7.6E-3 | 146.6 s |

**Table 2.** Number of parameters and training/validation/test losses, and CPU time of Algorithm 2 with different neural networks (GRU 1L, GRU 2L, BiGRU 1L, Conv1D GRU)

| Model | Parameter | Training loss | Validation Loss | Test Loss | CPU time |
|---|---|---|---|---|---|
| GRU 1L | 401 | 7.9E-3 | 5.5E-3 | 8.8E-3 | 526.9 s |
| GRU 2L | 1061 | 7.8E-3 | 6.6E-3 | 9.0 E-3 | 649.4 s |
| BiGRU 1L | 801 | 7.8E-3 | 6.9E-3 | 9.0 E-3 | 610.7 s |
| CONV1D-GRU | 8627 | 7.2E-3 | 9.9E-3 | 9.6 E-3 | 240.2 s |

We report the performance of Pitch-GRU in terms of the number of parameters, losses, and CPU time. It is noteworthy that for all the cases, the test errors were below 8.0E-3. This indicated a close match between the predicted and actual CQ values. The test loss was the smallest when GRU 2L was employed. However, the CPU time was shortest when CONV1D-GRU was used for the neural network.

Now, we compare Pitch-GRU and Algorithm 2. Because we did not use pitch feature extraction in Algorithm 2, the CPU time was longer for this algorithm. This occurred because in Pitch-GRU, pitch extraction reduces the length of the audio signal, which yields a small sequence for the input for neural networks. In terms of accuracy, Pitch-GRU has fewer errors than Algorithm 2. This shows that the pitch algorithm is capable of capturing the important features of audio signals. Therefore, we conclude that the proposed pitch-GRU algorithm is computationally efficient while obtain-

ing a high CQ prediction accuracy.

It is reasonable to consider whether other types of feature extraction algorithms enhance GRU-type neural networks for predicting CQ values. Therefore, we changed the first step of Pitch-GRU by replacing the pitch algorithm with MFCC [24], Chroma [25], ZCR [26], and RMS_E [27]. The test losses of the resulting algorithm are listed in Table 3. We observed that the pitch algorithm is most accurately employed with GRU-type neural networks. This validates our selection of the pitch algorithm for audio-feature extraction.

**Table 3.** Loss table of GRU-type neural network with feature extracted by MFCC, Chroma, ZCR, Pitch, and RMS_E algorithms

| Model | MFCC | Chroma | ZCR | Pitch | RMS_E |
|---|---|---|---|---|---|
| GRU 1L | 1.5E-02 | 9.4E-03 | 8.7E-03 | 8.0E-03 | 8.5E-03 |
| GRU 2L | 1.3E-01 | 8.8E-03 | 9.1E-03 | 7.6E-03 | 8.7E-03 |
| BiGRU 1L | 1.7E-02 | 9.1E-03 | 9.3E-03 | 8.0E-03 | 8.7E-03 |
| CONV1D-GRU | 8.9E-03 | 8.7E-03 | 8.3E-03 | 7.6E-03 | 9.2E-03 |

**Table 4.** Comparison of Pitch-GRU with random forest and XGBoost algorithm

| Model | Training loss | Validation loss | Test Loss | CPU time |
|---|---|---|---|---|
| Pitch-GRU | 7.7E-3 | 7.7E-3 | 7.6E-3 | 107.4 s |
| Random Forest | 3.5E-4 | 7.8E-2 | 7.4E-2 | 6.4 s |
| XGBoost | 2.1E-3 | 7.6E-2 | 7.5E-2 | 2.5 s |

Finally, we compared the performance of Pitch-GRU employed with GRU 2L with that of other tree ensemble-type algorithms. In Table 4, we report the losses and CPU-time for Pitch-GRU, random forest [28,29], and XGBoost [30]. It is observed that the test loss of the Pitch-GRU algorithm was lower than those of random forest and XGBoost.

## IV. CONCLUSION

In this study, we developed a new method called Pitch-GRU to predict CQ using audio signals during vocal phonation. Data were collected from vocal productions of the vowel /‘a’/ by 22 individuals. The CQ values were extracted from EGG waveforms. By matching the audio signals and CQ values, we trained the GRU-based neural networks using supervised learning. To enhance the efficiency, the audio signal was preprocessed using the pitch feature extraction algorithm. The results revealed that, the MSE errors between the predicted CQ and real CQ was below 9E-03 for all the cases. This demonstrates the capability of the proposed algorithm in analyzing the vocal fold behavior during vocal phonation. Next, we discussed the likely utilization of the proposed Pitch-GRU algorithm. Because our algorithm can predict the CQ in real time, it can be used efficiently as a reference tool for educating professional singers or as vocal cord exercises for patients with vocal fold disorders. In future works, we can consider different vowel such as /‘i’/ or /‘u’/.

## ACKNOWLEDGEMENTS

This study was supported by a grant (NRF KSN1824130) from the Korea Institute of Oriental Medicine.

## REFERENCES

[ 1 ] E. B. Lacerda and C. A. B. Mello, “Automatic classification of laryngeal mechanisms in singing based on the audio signal,” *Procedia Computer Science,* vol. 112, pp. 2204-2212, Feb. 2017. DOI: 10.1016/j.procs.2017.08.115.

[ 2 ] A. Zysk and P. Badura, “An Approach for Vocal Register Recognition Based on Spectral Analysis of Singing,” *International Journal of Cognitive and Language Sciences*, vol. 11, no. 2, pp. 207-212, Jan. 2017. DOI: 10.5281/zenodo.1128825.

[ 3 ] R. K. Shosted, “Vocalic context as a condition for nasal coda emergence: aerodynamic evidence,” *Journal of the International Phonetic Association*, vol. 36, no. 1, pp. 39-58, May 2006. DOI: 10.1017/S0025100306002350.

[ 4 ] P. Fabre, “Percutaneous electric process registering glottic union during phonation: glottography at high frequency; first results,” *Bulletin de L'academie Nationale de Medecine*, vol. 141, no. 3-4, pp. 66-69, Jan. 1957. DOI: 10.1007/BF02991550.

[ 5 ] V. Hampala, M. Garcia, J. G. Švec, R. C. Scherer, and C. T. Herbst, “Relationship between the electroglottographic signal and vocal fold contact area,” *Journal of Voice*, vol. 30, no. 2, pp. 161-171, Mar. 2016. DOI: 10.1016/j.jvoice.2015.03.018.

[ 6 ] F. M. B. Lã and J. Sundberg, “Contact quotient versus closed quotient: a comparative study on professional male singers,” *Journal of Voice*, vol. 29, no. 2, pp. 148-154, Mar. 2015. DOI: 10.1016/j.jvoice.2014.07.005.

[ 7 ] K. Verdolini, R. Chan, I. R. Titze, M. Hess, and W. Bierhals, “Correspondence of electroglottographic closed quotient to vocal fold impact stress in excised canine larynges,” *Journal of Voice*, vol. 12, no. 4, pp. 415-423, Apr. 1997. DOI: 10.1016/S0892-1997(98)80050-7.

[ 8 ] J. Y. Lim, S. E. Lim, S. H. Choi, J. H. Kim, K. M. Kim, and H. S. Choi, “Clinical characteristics and voice analysis of patients with mutational dysphonia: clinical significance of diplophonia and closed quotients,” *Journal of voice*, vol. 21, no. 1, pp. 12-19, Jan. 2007. DOI: 10.1016/j.jvoice.2005.10.002.

[ 9 ] K. He, X. Zhang, S. Ren, and, J. Sun, “Deep residual learning for image recognition,” in *Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas: US, pp. 770-778, 2016. DOI: 10.1109/CVPR.2016.90.

[10] J. Ho, A. Jain, and P. Abbeel, “Denoising diffusion probabilistic models,” in *Proceeding of the 33th Advances in Neural Information Processing Systems*, Vancouver: CA, pp. 6840-6851, 2020. DOI: 10.48550/arXiv.2006.11239.

[11] D. Silver, A. Huang, CJ. Maddison, and A. Guez, “Mastering the game of Go with deep neural networks and tree search”. *Nature*, vol. 529, pp. 484-489, Jan. 2016. DOI: 10.1038/nature16961.

[12] T. Haarnoja, A, Zhou, P. Abbeel, and S. Levine, “Soft actor-critic:

Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *Proceeding of International Conference on Machine Learning*, Stockholm: SE, pp. 1861-1870, 2018. DOI: 10.48550/arXiv.1801.01290.

[13] T. Haarnoja, A, Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *Proceeding of International Conference on Machine Learning*, Stockholm: SE, pp. 1861-1870, 2018. DOI: 10.48550/arXiv.1801.01290.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, AN. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceeding of the 30th Advances in Neural Information Processing Systems*, Long beach: US, pp. 6000-6010, 2017. DOI: 10.48550/ arXiv.1706.03762.

[15] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," in *Proceeding of the 33 th Advances in Neural Information Processing Systems*, Vancouver: CA, pp. 1877-1901, 2020. DOI: 10.48550/arXiv.2005.14165.

[16] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735-1780, Nov. 1997. DOI: 10.1162/neco.1997.9.8.1735.

[17] J. Chen X. Xue. "A transfer learning-based long short-term memory model for the prediction of river water temperature," *Engineering Applications of Artificial Intelligence*, vol. 133, pp. 108605, 2024. DOI: 10.1016/j.engappai.2024.108605.

[18] C. Qin, D. Qin, Q. Jiang, and B. Zhu, "Forecasting carbon price with attention mechanism and bidirectional long short-term memory network," *Energy*, vol. 299, pp. 131410, 2024. DOI: 10.1016/ j.energy.2024.131410.

[19] R. Dey and F. M. Salem, "Gate-variants of gated recurrent unit (GRU) neural networks," in *Proceeding of 2017 IEEE 60th International Midwest Symposium on Circuits and Systems*, Boston: US, pp. 1597-1600, Aug. 2017. DOI: 10.1109/MWSCAS.2017. 8053243.

[20] Y. Yevnin, S. Chorev, I. Dukan, and Y. Toledo, "Short-term wave forecasts using gated recurrent unit," *Ocean Engineering*, vol. 268, no. 15, pp. 113389, 2023. DOI: 10.1016/j.oceaneng.2022.113389.

[21] L. Zhang, J. Zhang, W. Gao, F. Bai, N. Li, and N. Ghadimi, "A deep learning outline aimed at prompt skin cancer detection utilizing gated recurrent unit networks and improved orca predation algorithm," *Biomedical Signal Processing and Control*, vol. 90, pp. 105858, 2024. DOI: 10.1016/j.bspc.2023.105858.

[22] E. Terhardt, G. Stoll, and M. Seewann, "Algorithm for extraction of pitch and pitch salience from complex tonal signals," *The Journal of the Acoustical Society of America*, vol. 71, no. 3, pp. 679-688, Mar. 1982. DOI: 10.1121/1.387544.

[23] D. Talkin and W. B. Klejin, "A robust algorithm for pitch tracking (RAPT)," *Speech Coding and Synthesis*, pp. 497-518, 1995.

[24] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980,* Dec. 2014. DOI: 10.48550/arXiv.1412.6980.

[25] C. Ittichaichareon, S. Suksri, and T. Yingthawornsuk, "Speech recognition using MFCC," in *Proceeding of International Conference on Computer Graphics, Simulation and Modeling*, Pattaya: TH, vol. 9. pp. 135-138, Jul. 2012.

[26] M. Muller and S. Ewert, "Chroma Toolbox: MATLAB implementations for extracting variants of chroma-based audio features," *International Society for Music Information Retrieval*, 2011.

[27] B. P. Das and R. Parekh, "Recognition of isolated words using features based on LPC, MFCC, ZCR and STE, with neural network classifiers," *International Journal of Modern Engineering Research*, vol. 2. no. 3. pp. 854-858, May-Jun. 2012.

[28] H. Panti, A. Jagtap, V. Bhoyar, and A. Gupta, "Speech emotion recognition using MFCC, GFCC, chromagram and RMSE features," *in Proceeding of 2021 8th International Conference on Signal Processing and Integrated Networks (SPIN)*, Nodia: IN, Aug. 2021. DOI: 10.1109/SPIN52536.2021.9566046.

[29] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5-32, 2001. DOI: 10.1023/A:1010933404324.

[30] J. Hu and S. Szymczak, "A review on longitudinal data analysis with random forests," *Briefings in Bioinformatics*, vol. 24, no. 2, pp. 1-11, 2023. DOI: 10.1093/bib/bbad002.

[31] T. Chen, C. Guestrin, and Carlos Guestrin. "Xgboost: A scalable tree boosting system." in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, pp. 785-794. DOI: 10.1145/2939672.2939785.

**Hyeonbin Han**
has been pursuing his B.S. from Department of Mathematical Data Science, Hanyang University ERICA, since 2023. His research interests include computer vision, deep reinforcement learning, and chat-bot.

**Keun Young Lee**
received his Ph.D. from Department of Mathematical Sciences, KAIST, in 2009. From 2017 to 2020, he was a faculty member of Department of Mathematics, Sejong University, Republic of Korea. From 2020, he has been an independent scholar in Republic of Korea. His research interests include Banach space theory, machine learning, and fuzzy theory.

**Seong-Yoon Shin**
received his M.S. and Ph.D. from Department of Computer Information Engineering, Kunsan National University, Gunsan, Republic of Korea, in 1997 and 2003, respectively. From 2006, he has been a professor in School of Computer Science and Engineering. His research interests include image processing, computer vision, and virtual reality. He can be contacted at the following email ID: s3397220@kunsan.ac.kr

**Yoseup Kim**
received his B.S. in Material Science and Engineering, Business and Technology Management (double major), and Chemical and Biological Engineering (minor) from KAIST, Daejeon, Republic of Korea, in 2015. He received his M.D. from Yonsei University College of Medicine in 2024, and has been leading several government R&D projects as a CEO and principal investigator at Deltoid Inc. since 2020. His research interests include motion and visual analysis in the field of digital healthcare

**Gwanghyun Jo**
received his B.S., M.S., and Ph.D. from Department of Mathematical Sciences, KAIST, Daejeon, Republic of Korea, in 2018. From 2018 to August 2019, he was a Postdoctoral Researcher with KAIST. He was a faculty member of Department of Mathematics, Kunsan National University during 2019–2023. He has been a faculty member of Department of Mathematical Data Sciences, Hanyang University ERICA. His research interests include numerical analysis and simulation of various fluids problems originating from hemodynamics, and petroleum engineering.

**Jihoon Park**
graduated with a Bachelor of Music in Classic Vocal from Department of Music, Chungnam University. He also completed the Intermediate Course for Voice Correction Specialist certified by the Korean Vocology Association (KOVA). Since 2014, he has been performing as a soloist tenor or actor at prestigious venues such as the Korea National Theater, Daejeon Art Center, Daejeon Observatory, and Daegu International Musical Festival (DIMF). He has also organized numerous concerts including classic and jazz concerts such as "Playing Love" and "The moment". He has held directorial positions at Nicedream Music Academy and Baekyang Studio. Additionally, he is an active member of the Korean Vocology Association (KOVA). His research interests include classical vocal phonation, vocal education, and scientific analysis of vocal behavior.

**Young-Min Kim**
received his B.S. in Mechanical Engineering from Yonsei University, Seoul, Republic of Korea, in 1999; M.S. in Mechanical Engineering from POSTECH, Republic of Korea, in 2001: and Ph.D. in Mechanical Engineering from KAIST, Republic of Korea, in 2011. From 2002 to 2006, he was a Research Scientist with Human-Welfare Robotic System Research Center, KAIST, Republic of Korea. Since 2011, he has been a Principal Researcher with Digital Health Research Division, Korea Institute of Oriental Medicine, Republic of Korea. His research interests include the medical devices for personalized healthcare, wearable sensors for daily health monitoring, sophisticated human–robot interface (HRI) technology, and innovative HRI applications.