Integrating Vocal Feedback to Optimize Neuromuscular Electrical Stimulation

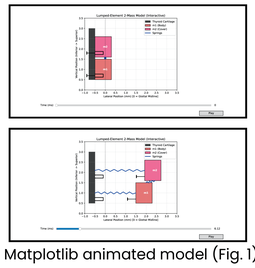
INTRODUCTION

Vocal fold paralysis occurs when one or both vocal cords lose their ability to move due to nerve damage, often resulting in breathy speech, increased risk of aspiration, and reduced vocal strength (National Institute on Deafness and Other Communication Disorders [NIDCD], 2011). Traditional therapies, such as voice exercises or surgical medialization, can partially restore voice but often fail to replicate natural laryngeal control. Recent research suggests that **neuromuscular electrical stimulation (NMES)** may improve vocal fold mobility and strength by reactivating paralyzed muscles (de Almeida et al., 2022). However, current NMES methods use fixed stimulation parameters and **lack personalized feedback mechanisms**, limiting their long-term efficacy. This project proposes a feasibility study for an **adaptive NMES system that uses machine learning to personalize NMES stimulation parameters based on acoustic feedback analysis.**

METHODOLOGY

Vocal Fold Biomechanics

The vocal folds are modeled as a **lumped-element two-mass system**, separating the structure into body and cover layers (Fig. 1). These masses are governed by a **mass-spring-damper framework** where springs represent tissue tension and mechanical coupling, while dampers simulate energy dissipation due to tissue viscosity. Net force dynamics balance the restorative structural forces against the aerodynamic pressures acting on each mass (eq. 1 & eq. 2). These driving pressures are derived from fluid dynamics. The Continuity Equation maintains a constant volumetric flow rate of incompressible air (eq. 3), while Bernoulli's Principle calculates the static pressure drop caused by the accelerated airflow through the variable glottal area (eq. 4)



Matplotlib animated model (Fig. 1)

$$m_1 a_1 = P_{g1} - b_1 v_1 - k_1(x_1 - x_{01}) - k_c(x_1 - x_2) \quad \text{(eq. 1)}$$

$$m_2 a_2 = P_{g2} - b_2 v_2 - k_2(x_2 - x_{02}) - k_c(x_2 - x_1) \quad \text{(eq. 2)}$$

$$v(t) = \frac{U}{A(t)} \propto \frac{U}{L \cdot x(t)} \quad \text{(eq. 3)}$$

$$P(x) = P_{lung} - \frac{1}{2}\rho\left(\frac{U}{A(x)}\right)^2 \quad \text{(eq. 4)}$$

Transitioning to Sound

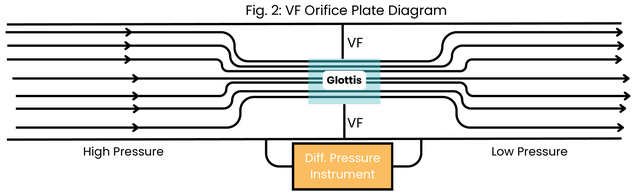


Fig. 2: VF Orifice Plate Diagram

The mechanical displacement of the vocal folds is translated into acoustic sound pressure using aerodynamic calculations and digital signal processing. First, the Orifice Flow Law (eq. 5) **converts the simulated glottal gap into a clean volumetric flow** (eq. 6) using a pre-calculated discharge constant (eq. 7). To introduce biological realism, subglottal pressure is modulated by a 5 Hz neuromuscular tremor (eq. 8).

$$Q = C_d A\sqrt{\frac{2\Delta P}{\rho}} \quad \text{(eq. 5)}$$

$$U_{g\_clean} = c \cdot x(t)\sqrt{P_{sub}} \quad \text{(eq. 6)}$$

$$c = C_d L\sqrt{\frac{2}{\rho}} \quad \text{(eq. 7)}$$

$$P_{sub\_dynamic} = P_{sub}(1 + \text{tremor\_depth} \cdot \sin(2\pi \cdot 5.0 \cdot t)) \quad \text{(eq. 8)}$$

The human vocal tract is then simulated as a **series of four digital resonators acting as an Infinite Impulse Response (IIR) filter cascade for the "Ah" vowel.** The discrete-time difference equation for each filter (eq. 9) relies on tuning coefficients derived from formant frequencies and bandwidths (eq. 10-13):

$$y_i[n] = U_g[n] + a_1 y_i[n-1] - a_2 y_i[n-2] \quad \text{(eq. 9)}$$

$$R = e^{-\pi \cdot B \cdot \Delta t} \quad \text{(eq. 10)} \qquad \theta = 2\pi \cdot f \cdot \Delta t \quad \text{(eq. 11)} \qquad a_1 = 2R\cos(\theta) \quad \text{(eq. 12)} \qquad a_2 = R^2 \quad \text{(eq. 13)}$$

Finally, the microphone detects changes in air density. Therefore, the radiated sound pressure is calculated as the first discrete derivative of the filtered flow, representing the acoustic radiation at the mouth (eq. 14).

$$P_{sound}[n] = y_4[n] - y_4[n-1] \quad \text{(eq. 14)}$$

Creating the Dataset

The theoretical models are executed programmatically to generate a synthetic dataset mapping physical parameters to acoustic outputs. The physics model solves the differential equations of motion using a 4th-Order Runge-Kutta (RK4) integration step. This method calculates the rates of change over a standard 0.5-second duration at a 44100 Hz sample rate. To create diverse training data, the underlying physical parameters (k_1, k_2, kc, b_1, b_2) are randomized using independent variance multipliers. The average structural integrity of these parameters dictates a binary diagnosis of "Healthy" or "Weak". The glottal gap array is passed through the acoustic cascade to produce audio waves. These arrays are saved as .wav files alongside corresponding .json files containing the ground-truth physical parameters and diagnostic labels.

ML Approach

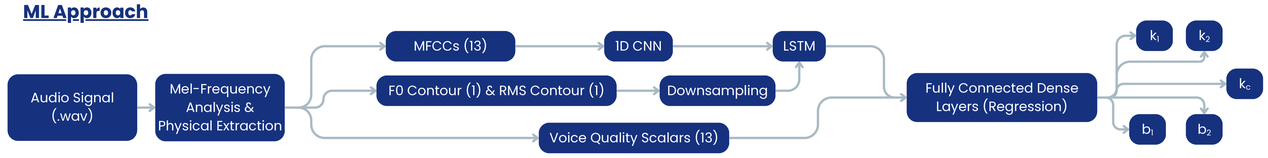


Fig. 3: Hybrid CNN-LSTM Architecture

A deep learning model solves the inverse problem by predicting the underlying physical and mechanical parameters from the resulting acoustic audio waves. The pipeline extracts two distinct sets of features from the audio data. The temporal set consists of **13 Mel-Frequency Cepstral Coefficients (MFCCs), fundamental frequency (F0) contours, and Root Mean Square (RMS) energy**. The scalar bypass set contains **13 source-level features (e.g., Jitter, Shimmer, Harmonics-to-Noise Ratio)** that bypass the vocal tract filter and directly reflect glottal oscillation regularity. The model utilizes a parallel pathway design. A **1D Convolutional Neural Network (CNN) scans the MFCCs for localized acoustic anomalies**, such as aspiration bursts. The output is concatenated with the down sampled F0 and RMS contours and fed into a **Long Short-Term Memory (LSTM) network to track continuous temporal variations like neuromuscular tremor.** The temporal summary generated by the LSTM is concatenated with the scalar features. This combined representation is passed through fully connected layers to map the acoustic profile back to the five physical targets (k_1, k_2, b_1, b_2, k_c). The network is trained using Mean Squared Error (MSE) loss with an Adam optimizer. The pipeline implements early stopping and a learning rate scheduler that halves the learning rate if validation loss plateaus.

RESULTS

The Hybrid CNN-LSTM architecture was successfully trained and evaluated on a synthetic dataset of 3000 simulated patient profiles (2400 training, 600 testing). The model **converged at epoch 34, with training officially halted at epoch 59** via an early stopping mechanism (25-epoch patience) to prevent overfitting. Regression analysis demonstrated varying degrees of predictive accuracy depending on the targeted mechanical parameter. The model achieved its strongest Coefficient of Determination (R^2) scores in predicting lower tissue damping (b_1) at **70.38% (MSE: ~6e^{-08})** and inter-mass coupling (k_c) at **60.50% (MSE: 3.25)**. Predictions for structural spring tension showed moderate correlation, yielding R^2 scores of **38.03%** for upper cover tension (k_2) and **37.46%** for lower body tension (k_1). Upper tissue damping (b_2) proved the most difficult parameter to isolate acoustically, returning an R^2 of 14.34%. As shown in Fig. 4, parameters with higher R^2 scores, such as b_1 and k_c, exhibit tight clustering along the $y = x$ ideal prediction line. Despite the model struggling to isolate secondary cover-layer dynamics, these results confirm that the algorithm can successfully infer primary biomechanical deficits, particularly body-layer viscosity and tissue coupling strength, using solely non-invasive acoustic features.

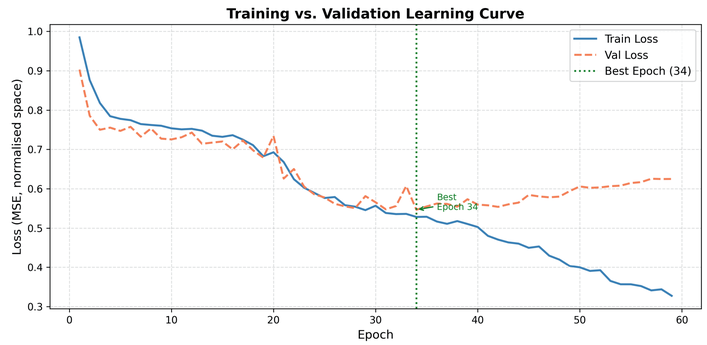


Fig. 4: Learning Curve

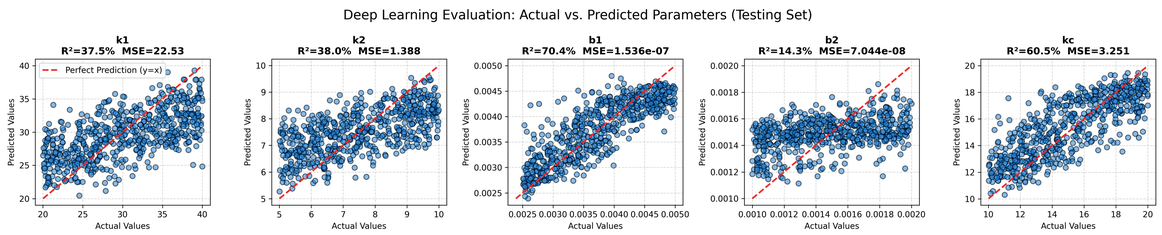


Fig. 5: Regression Results

DISCUSSION

This study **successfully demonstrated the theoretical feasibility of using machine learning to map non-invasive acoustic feedback to underlying laryngeal mechanics**. This serves as a crucial foundational step toward creating an adaptive, closed-loop Neuromuscular Electrical Stimulation (NMES) system that can personalize therapy based on real-time vocal fold health.

The Hybrid CNN-LSTM architecture **excelled at isolating specific mechanical deficits, particularly lower tissue viscosity (b_1) and inter-mass coupling (k_c).** The high predictive accuracy for these parameters correlates directly to their distinct, deterministic acoustic footprints. In the simulation, b_1 solely governs neuromuscular tremor (100% weight), producing a distinct 5 Hz low-frequency amplitude modulation that the LSTM can easily track over time. Similarly, k_c solely governs vocal tract drift (100% weight), creating distinct spectral shifts that are highly detectable by the CNN's MFCC spatial analysis.

Conversely, the model **struggled to accurately predict upper cover damping (b_2)**. This failure highlights a limitation in acoustic isolation: the biomechanical influence of b_2 is encoded through only a 20% share of the aspiration channel, competing against k_1 (30%), k_2 (30%), k_c (15%), and b_1 (5%). Lacking a primary, concentrated acoustic signature, its variance was masked by the dominating variables. Furthermore, structural spring tensions (k_1, k_2) yielded only moderate accuracy because they primarily drove aspiration noise. Aspiration is a stochastic, broadband Gaussian turbulence, making it inherently harder to regress back to a deterministic mechanical stiffness value compared to the distinct sine-wave modulations of tremor and drift.

While the lumped-element two-mass model effectively proves the algorithmic concept, relying on a purely computational dataset introduces significant clinical limitations. The synthetic dataset does not account for dynamic human vocal tract geometries, varying lung pressures, or compensatory muscular behaviors often seen in vocal paralysis patients (such as ventricular fold recruitment). Furthermore, environmental noise and natural recording artifacts present in real-world audio were not simulated, representing a hurdle for practical application.

CONCLUSION

This computational study **established a novel, proof-of-concept deep learning pipeline capable of isolating hidden vocal fold biomechanics using purely non-invasive acoustic data.** By proving that specific mechanical degradations, specifically body-layer viscosity and tissue coupling, translate to quantifiable acoustic markers that neural networks can decode, this research validates the theoretical foundation for acoustically driven, adaptive NMES therapy.

The Hybrid CNN-LSTM was a well-suited choice: the CNN detects localized spectral anomalies in the MFCCs while the LSTM tracks the time-dependent tremor and F0 drift patterns that encode the mechanical parameters. However, **the persistent difficulty predicting b_2 (R^2=14.3%) suggests future work should explore Transformer-based encoders**, whose multi-head self-attention can capture non-local temporal dependencies that LSTM hidden states compress into a fixed-size bottleneck, as well as parameter-specific decoder heads to reduce cross-parameter interference in the shared output layer.

Future work must transition from synthetic modeling to clinical validation. The model will be retrained and tested using real human audio samples from established clinical voice pathology databases. Ultimately, the refined diagnostic algorithm will be integrated with a physical NMES hardware prototype to facilitate in vivo testing, evaluating the system's ability to adjust electrical stimulation parameters in real-time based on live patient phonation.

REFERENCES

Amarante Andrade, P., Pereira Padilha, J., Leite Vieira, V., da Costa Marques, M., Fonseca e Silva, K., & Lemos, S. M. A. (2023). Electrical stimulation in voice therapy: A systematic review. Journal of Voice, 37(5), 812.e1–812.e14. https://doi.org/10.1016/j.jvoice.2021.04.023

de Almeida, A. N. S., da Cunha, D. A., Silva, H. J., Siqueira Júnior, L. T., Santos, T. S., & Silva, K. C. (2022). Effect of electrical stimulation on the treatment of dysphonia: A systematic review. Journal of Voice, 36(5), 650–660. https://doi.org/10.1016/j.jvoice.2020.09.014

Ishizaka, K., & Flanagan, J. L. (1972). Synthesis of voiced sounds from a two-mass model of the vocal cords. Bell System Technical Journal, 51(6), 1233–1268. https://doi.org/10.1002/j.1538-7305.1972.tb02651.x

National Institute on Deafness and Other Communication Disorders [NIDCD]. (2011, October). Vocal fold paralysis (Pub. No. 11-4306) [Fact sheet]. U.S. Department of Health and Human Services, National Institutes of Health. https://www.nidcd.nih.gov/sites/default/files/Documents/health/voice/NIDCD-Vocal-Fold-Paralysis.pdf