Phonemic Resonance Networks: A Bio-Acoustic Framework for Emotion-Aware Speech Synthesis

Dr. Elara Vance* Prof. Kaito Tanaka† Dr. Sofia Petrova‡

Introduction

Contemporary speech synthesis systems have achieved remarkable progress in naturalness and intelligibility, yet they largely fail to capture the nuanced emotional dimensions that characterize authentic human communication. The predominant approach treats emotional expression as a post-processing layer applied to neutral speech, fundamentally misrepresenting how emotions emerge from the complex interplay of physiological changes in the human vocal apparatus. This paper introduces Phonemic Resonance Networks (PRNs), a paradigm-shifting framework that reconceptualizes emotional speech synthesis through the lens of vocal biomechanics and acoustic resonance theory.

Human emotional expression in speech arises from coordinated changes in respiratory patterns, laryngeal tension, vocal fold vibration, and articulatory precision. These physiological modifications produce distinctive acoustic signatures that current synthesis methods struggle to replicate authentically. Our work addresses this limitation by developing a computational model that explicitly incorporates principles from vocal fold dynamics, formant theory, and articulatory phonetics.

Research Questions:

1. How can we mathematically model the relationship between emotional states and the resulting changes in vocal tract resonance patterns?
2. Can a neural architecture constrained by bio-acoustic principles generate more emotionally authentic speech than conventional approaches?
3. What insights do learned resonance embeddings provide about the acoustic correlates of emotional expression?

Our contributions include: (1) the PRN architecture with its novel resonance embedding layer, (2) a methodology for incorporating bio-physical constraints

*Department of Computational Linguistics, University of Cambridge
†Institute for Bio-Acoustic Research, Kyoto University
‡Neural Speech Laboratory, Moscow State University

into neural speech synthesis, (3) experimental validation demonstrating superior emotional naturalness, and (4) interpretable representations of emotional speech production.

Methodology

Bio-Acoustic Foundations

The PRN framework is grounded in three key bio-acoustic principles:

Vocal Fold Dynamics: Emotional states alter the mechanical properties of vocal folds, affecting fundamental frequency (F0), jitter, and shimmer. We model these changes using a modified mass-spring-damper system:

$$m\frac{d^2x}{dt^2} + b(\epsilon)\frac{dx}{dt} + k(\epsilon)x = F_{subglottal} \tag{1}$$

where ϵ represents emotional state parameters that modulate damping coefficient b and stiffness k.

Formant Emotion Mapping: Different emotions produce characteristic shifts in formant frequencies due to changes in vocal tract configuration. We establish emotion-specific formant transformation matrices:

$$F_{emo} = T(\epsilon) \cdot F_{neutral} \tag{2}$$

Articulatory Precision: Emotional intensity affects the clarity and precision of articulation, which we model through emotion-dependent smoothing kernels applied to articulatory trajectories.

PRN Architecture

The PRN architecture consists of four main components:

Resonance Embedding Layer: This core innovation transforms phonemic input into a multi-dimensional resonance space. Each phoneme is represented as a 128-dimensional vector encoding its typical spectral envelope, formant structure, and articulatory features.

Bio-Physical Constraint Module: This component ensures generated speech parameters remain within physiologically plausible ranges. It implements soft constraints on F0 contours, formant trajectories, and spectral tilt based on human vocal production limits.

Emotion-Conditioned Generator: A convolutional-recurrent hybrid network that generates mel-spectrograms conditioned on both text input and emotional state vectors. The architecture includes:

- Emotion-aware attention mechanisms - Multi-scale convolutional filters for spectral pattern learning - Gated recurrent units for temporal modeling - Residual connections for gradient flow

Vocoder with Emotional Modulation: A neural vocoder that converts mel-spectrograms to waveform, incorporating emotion-specific pulse shapes and noise characteristics.

Training Methodology

We trained PRNs using a multi-stage approach:

1. **Pre-training:** Initial training on large neutral speech corpora to learn fundamental speech patterns 2. **Emotion Fine-tuning:** Specialized training on emotional speech data with electroglottographic measurements 3. **Bio-Physical Regularization:** Incorporation of physiological constraints through penalty terms in the loss function

$$\mathcal{L}_{total} = \mathcal{L}_{recon} + \lambda_1 \mathcal{L}_{bio} + \lambda_2 \mathcal{L}_{emo} \tag{3}$$

where $\mathcal{L}_{bio}$ enforces vocal production constraints and $\mathcal{L}_{emo}$ ensures emotional discriminability.

Results

Experimental Setup

We evaluated PRNs against three state-of-the-art emotional TTS systems: Tacotron2-Emo, FastSpeech2-ES, and GST-Based TTS. Our novel dataset comprised 1,200 speech samples across eight emotional states (joy, sadness, anger, fear, surprise, disgust, neutral, tender) with simultaneous electroglottographic recordings.

Evaluation metrics included: - Mean Opinion Score (MOS) for naturalness and emotional appropriateness - Emotion Recognition Accuracy (ERA) from human listeners - Physiological Plausibility Score (PPS) based on electroglottographic correlation - Intelligibility measured through word error rate

Quantitative Results

Table 1: Comparison of Emotional Speech Synthesis Systems

System	MOS-Naturalness	MOS-Emotion	ERA	PPS
Tacotron2-Emo	3.42	3.18	68.3%	0.52
FastSpeech2-ES	3.67	3.45	72.1%	0.61

System	MOS-Naturalness	MOS-Emotion	ERA	PPS
GST-Based TTS	3.58	3.72	75.6%	0.58
PRN (Ours)	**4.23**	**4.65**	**89.2%**	**0.87**

PRNs achieved significantly higher scores across all metrics, with particularly strong performance in emotional appropriateness and physiological plausibility. The 47% improvement in emotional naturalness MOS demonstrates the effectiveness of our bio-acoustic approach.

Qualitative Analysis

Analysis of the learned resonance embeddings revealed several insightful patterns:

Emotion Clustering: The embeddings formed interpretable clusters corresponding to physiological emotion families. High-arousal emotions (anger, joy) clustered separately from low-arousal emotions (sadness, tenderness), with fear and surprise forming an intermediate group.

Acoustic Correlates: We identified specific resonance patterns associated with emotional states: - Anger: Increased higher formant energy, irregular F0 contours - Sadness: Formant compression, reduced high-frequency energy - Joy: Expanded formant bandwidth, regular F0 modulation

Cross-Lingual Transfer: Preliminary experiments showed that PRN embeddings learned from English data transferred effectively to Japanese and Russian, suggesting universal aspects of emotional vocal production.

Conclusion

This paper has presented Phonemic Resonance Networks, a novel framework for emotion-aware speech synthesis grounded in bio-acoustic principles. Our approach fundamentally differs from conventional methods by modeling how emotions physically manifest in the human vocal apparatus, rather than treating emotional expression as a superficial modulation.

The key innovations of our work include:

1. The resonance embedding layer that captures the acoustic signatures of emotional speech production 2. The integration of bio-physical constraints into neural network training 3. The hybrid architecture that combines spectral and temporal modeling with physiological plausibility 4. The interpretable representations that provide insight into emotional speech acoustics

Our experimental results demonstrate that PRNs generate significantly more emotionally authentic speech while maintaining high intelligibility. The learned embeddings form meaningful clusters that correspond to physiological emotion

families, offering new understanding of the acoustic correlates of emotional expression.

Future work will explore applications in clinical settings for speech therapy, cross-cultural emotional expression modeling, and real-time emotional adaptation in human-computer interaction. The PRN framework establishes a new direction for speech synthesis research that bridges computational methods with the biological reality of human vocal production.

Acknowledgments

This research was supported by the International Bio-Acoustics Research Consortium and the Neural Speech Foundation. We thank our participants and the speech pathology clinics that contributed data to this study.