Kannada Ser Real-Time Speech Emotion Recognition Using Fine-Tuned Wav2vec2

Parikshith S[1], Pavana S[2], Sahana S Gudennavar[3], Varshini M R[4], Sushmitha S R[5]
[1,2,3,4]*Department. of Information Science and Engineering P.E.S. College of Engineering, Mandya, Karnataka, India*
[5]*Professor Department. of Information Science and Engineering P.E.S. College of Engineering, Mandya, Karnataka, India*

***Abstract*—Speech Emotion Recognition (SER) for low-resource regional Indian languages remains a largely unresolved challenge in affective computing research. Existing SER systems predominantly target high-resource languages such as English and Mandarin, leaving approximately 56 million Kannada speakers without specialized affective intelligence tools. This paper presents KannadaSER, a production-grade, full-stack platform for real-time emotion detection from Kannada speech audio. The system fine-tunes the facebook/wav2vec2-base transformer model on a curated Kannada emotional speech corpus spanning five classes — angry, fear, happy, neutral, and sad. The classifier head consists of Linear (768→128)–ReLU–Dropout (0.3)–Linear (128→5) appended to mean-pooled Wav2Vec2 contextual representations. A secondary SUPERB comparison model (superb/wav2vec2-base-superb-er) runs in parallel, and a threshold-based override rule resolves the perceptually confusable sad/fear boundary. The React + TypeScript frontend supports live microphone recording and file upload; the Flask backend integrates JWT authentication, MongoDB persistence, Cloudinary media storage, and a dedicated visualization service extracting waveform, pitch via pYIN, 13-coefficient MFCC, STFT spectrum, RMS energy, and zero-crossing rate. The system achieves a macro-averaged F1-score of 0.84 and overall accuracy of 84.0%, outperforming prior Kannada SER baselines by at least 1.6 percentage points.**

***Index Terms*—Kannada SER; Wav2Vec2; Self-Supervised Learning; Affective Computing; Low-Resource Language; Flask; React; MFCC; pYIN; Dual-Model Inference.**

I. INTRODUCTION

Paralinguistic content in human language carries a great deal more information than what is actually said. Prosodic characteristics such as pitch contour, energy profile, rate of speaking, and formant transition in spectral characteristics convey the emotional condition of the speaker in a way that can be understood by anyone irrespective of their linguistic ability. Automation of this process of identifying the emotion in speech, known as Speech Emotion Recognition (SER), has drawn considerable attention in academia for over twenty years due to its potential to revolutionize applications ranging from mental health tracking to robotics. Despite this breadth of application, the overwhelming majority of published SER systems are designed, trained, and benchmarked exclusively on high-corpora based on the resources language which is mostly English (IEMOCAP, MSP-IMPROV), German (EMO-DB), or Mandarin Chinese. Such a systematic bias towards certain languages presents a significant accessibility barrier to users who communicate in dialects of Indian languages. For example, Kannada (□□□□□) – the official language of the state of Karnataka listed among the Scheduled Languages by the Eighth Schedule of the Constitution of India – is the native language of roughly 56 million individuals. In particular, the features of the language, namely, the presence of retroflex consonants, agglutinative morphological processes which produce long polysynthetic words, and tonality on vowels, make the

generalization of models trained on Germanic/Sino-Tibetan phonemes highly inefficient when applied to speech in Kannada [6].

There is also a more substantial technical problem of a lack of publicly accessible and annotated production-level corpus for Kannada emotional speech. Studies of SER on Indic languages almost exclusively work on Hindi (Indic languages written in Devanagari script with another phonological makeup) while leaving Dravidian languages underrepresented within the field. Besides, there is not a single existing piece of work in the scientific literature which includes the implementation of Kannada SER in the full-stack web application [5]

This paper addresses the above gap by presenting Kannada SER, a production-quality full-stack framework for Kannada speech emotion recognition. The core technical contributions are:

- Recording of a five class Kannada speech emotion recognition corpus (2100 utterances after augmentation) for anger, fear, happiness, neutral and sadness classes, using controlled elicitation scripts in both Kannada and English.
- Fine tuning and evaluation of the facebook/wav2vec2-base self-supervised speech transformer on Kannada SER task, where our proposed two-layer custom classification head scores 84.0% of overall accuracy with 0.84 macro-F1 score.
- An innovative dual-model inference scheme, where the main Kannada Wav2Vec2 and the reference model from SUPERB dataset run in tandem; a perceptual threshold criterion (>=8%) is used to decide when to override sad/fear classes.
- A full-stack production grade application – React + TypeScript frontend, Flask REST API backend, MongoDB persistence, Cloudinary media storage, JWT authentication— with a dedicated acoustic feature visualization service.

II. LITERATURE REVIEW

The evolution of SER methodology traces three broad paradigm shifts: handcrafted-feature classifiers, end-to-end deep learning, and self-supervised pre-trained speech models.

A. Classical Feature Engineering Approaches

The early systems for SER employed manually designed acoustic feature sets consisting of Mel-Frequency Cepstral Coefficients (MFCCs), F0 parameters, formants (F1-F3), energy profiles, and speaking rate together with discriminative classifiers such as SVM, HMM, GMM, and k-Nearest Neighbours [1]. These systems provided high performance levels on static datasets such as IEMOCAP and EMO-DB (around 70–78%), yet demonstrated brittle generalization abilities across different corpora, failing to use the time-related dynamics of emotions in their analysis. The Random Forest classifier along with LPC-based features [7] yielded a performance level of 76.8% on multiple-language benchmarks, yet needed manual feature engineering skills for each language used.

B. Deep Learning Architectures for SER

The challenge of feature extraction was circumvented by designing convolutional and recurrent neural architectures to learn from spectrograms or raw audio signals. The effectiveness of learning spectral structure using CNNs and temporal modeling using LSTMs was shown by CNN-LSTM neural network architectures [3], which showed that combining spectral texture using CNN layers and prosodic dynamics using LSTM layers using log-Mel spectrograms can achieve Hindi SER performance of about 78.5%. Keerthika et al. [2] adapted such deep architectures for detecting dysarthria, showcasing the cross-application benefits of learning spectral features beyond emotions. In a detailed literature survey by Shaik et al. [8], it was found that multi-scale CNNs always outperformed LSTM networks for emotional speech processing.

C. Self-Supervised Speech Models

The Wav2Vec2 [4] model, developed by Facebook AI Research in 2020, used a convolutional feature encoder connected to a transformer context encoder having 12 layers and pre-trained using contrastive masked audio modeling for 960 hours of unlabeled data in English (LibriSpeech). These representations contain extensive linguistic information such as prosody in a high dimensional continuous embedding space without any need for acoustic features. In fine-tuning, Wav2Vec2 significantly outperformed MFCC + SVM

baselines on SER [4]. The SUPERB benchmark [5] provided a unified multi-task testing protocol for self-supervised speech models on SER, ASR, speech verification, and phoneme classification tasks, naming superb/wav2vec2-base-superb-er as the Wav2Vec2 emotion recognition model.

D. Cross-Lingual and Regional-Language SER

Contextual Embeddings for Cross-lingual Hypernasality Detection using Wav2Vec2 was investigated by Kothadia [6], which revealed that language-independent paralinguistic information is encoded within the transformer-level embeddings that are robust even in a domain-shift scenario. The generalization capability of self-supervised models to low-resource language inference through multilingual fine-tuning was demonstrated by Rui et al. [7]. However, despite these breakthroughs, none of the previous research efforts resulted in a production-ready SER framework that caters specifically to Kannada language with functionalities likeand persistent history— a gap directly addressed by KannadaSER.

III. SYSTEM DESIGN AND METHODOLOGY

A. Dataset Construction and Augmentation

The Kannada emotional speech corpus was compiled because there is currently no publicly available SER dataset in Kannada with enough data and proper labeling. The recruitment process involved native Kannada-speaking participants aged between 18 and 45 years old in equal representation among genders. Participants recorded both scripted and semi-spontaneous speech within each of the following emotion types: angry, fear, happy, neutral, and sad. The script was delivered to participants in two languages: Kannada script and English translation.

Recording sessions took place under a semi-anechoic setting utilizing a condenser microphone, sampled at 44.1 kHz. Samples were then resampled at 16 kHz to mono-channel pulse-code modulation wave format. Annotators fluent in Kannada and English were tasked with checking the emotional valence of each utterance according to the given script; utterances that had unclear emotional valence, recording glitches, or fewer than half a second of voiced audio (VAD trim, librosa.effects.trim, top_db=25) were discarded. Table I summarizes the final corpus composition after stratified 80/10/10 train/validation/test splitting.

TABLE I Kannada SER Corpus Composition (Post-Augmentation)

Emotion	Train	Val	Test	Total
Angry	336	42	42	420
Fear	288	36	36	360
Happy	360	45	45	450
Neutral	384	48	48	480
Sad	312	39	39	390
Total	1,680	210	210	2,100

Three augmentation strategies were applied exclusively to training samples to address overfitting risk given the modest raw corpus size: (1) additive Gaussian white noise at signal-to-noise ratios of 15, 20, and 25 dB; (2) synthetic room impulse response convolution using randomly generated RIRs of varying reverberation time (T60: 0.2–0.8 s); (3) time-stretching at factors of 0.9× and 1.1× using librosa.effects.time_stretch. This tripled the effective training set.

B. Acoustic Feature Extraction

The primary inference pathway feeds raw 16 kHz waveforms directly to the Wav2Vec2 convolutional encoder While the direct inference pipeline uses the raw audio waveform in its natural form (16 kHz samples) as input to the convolutional encoder of Wav2Vec2, the separate visualization pipeline calculates a set of highly interpretable features per each uploaded audio clip by means of the librosa toolkit:

- Waveform in the time domain (amplitude against time at 16 kHz)
- Mel-frequency cepstral coefficients (13 coefficients plus derivatives – total of 39 coefficients)

Short-Time Fourier Transform (STFT) magnitude and phase spectra

- Root Mean Squared (RMS) energy envelope extracted for every 512-samples frame
- Zero crossing rate (ZCR) per frame –

voiced/unvoiced indicator

Additionally, duration, intensity (in dB using amplitude to db transformation) and average value of MFCC coefficients are calculated as scalar statistics.

C. Model Architecture

The first classifier is based on the facebook/wav2vec2-base architecture, which comprises a 12-layer transformer model pre-trained on 960 hours of LibriSpeech English data using the Wav2Vec2 contrastive objective function. The 7-layer convolutional feature extractor generates 512-dimensional frame-level representations from the raw waveform input, which are further encoded by the 12-layer transformer context processor into 768-dimensional contextual embeddings. While pre-trained on English data, the sub-phoneme-level convolutional features capture acoustic properties such as pitch, energy, and rhythm, which can be transferred across languages like Kannada [6].

TABLE II KannadaSER Model Architecture Summary

Layer	Details	Output Shape
Raw Waveform Input	16 kHz mono, 4 s pad/trim	64,000 samples
Conv Feature Enc.	7-layer CNN (Wav2Vec2-base)	512 × T'
Context Network	12-layer Transformer, 768 hidden	768 × T'
Mean Pooling	emporal avg of last hidden state	768
Linear + ReLU	768 → 128, Dropout 0.3	128
Output Head	Linear 128 → 5, Softmax	5 probs

A two-layer classifier is added to the temporally mean-pooled Wav2Vec2 output: Linear(768→128) → ReLU → Dropout(0.3) → Linear(128→5) → Softmax. The audio is first padded or truncated to have precisely 4.0 seconds (64,000 samples @ 16 kHz), then tokenized using the Wav2Vec2Processor. The emotion labels' integer representation map to: 0-> angry, 1-> fear, 2-> happy, 3-> neutral, 4-> sad. Inference is performed on either CUDA (if available) or CPU.

The entire architecture, including the transformer backbone and classifier head, is fine-tuned end-to-end through backpropagation, using AdamW as the optimizer, starting learning rate 3×10^{-5}, linear warmup over 100 steps, and cosine annealing decay schedule. Model training took 20 epochs and batch size 8 with the Hugging Face Trainer. Model checkpoints were stored for each epoch, but only the best-performing one (by lowest validation loss) was kept for deployment (models/emotion_model.pt)

D. Dual-Model Cross-Check Inference Pipeline

The first drawback in relation to the primary algorithm is the fact that the emotion space in SUPERB basically collapses into only four emotion types (angry, happy/excited, neutral, and sad), where the level of fear confidence is underrepresented. The solution in relation to this problem is the use of the two-stage decision-making procedure by KannadaSER:

- Stage one: Parallel inference: Two models – the fine-tuned Kannada Wav2Vec2 predictor and the SUPERB comparator – work in parallel on the same input, which is the 16 kHz mono WAV format file.
- Stage two: Override decision: If the SUPERB model determines sad AND the fear confidence in the output of the Kannada model is >= 8%, then the output will be overruled by fear. The cutoff point (8% of fear confidence) was found empirically using the validation dataset. The entire pipeline in backend/routes/predict.py goes through audio validation → FFmpeg conversion to 16 kHz mono WAV (for formats other than WAV: MP3, OGG, FLAC, M4A, WebM) → Cloudinary file upload → dual-model inference simultaneously → applying the override rule → MongoDB database storage → JSON response. The fear value in the probability distribution returned to the user interface is guaranteed by the Kannada model, ensuring that 0% fear is not displayed in the UI if there is a fear signal.

E. Full-Stack System Architecture

Production architecture is designed in a

hierarchical manner using the REST API as an integration mechanism.

Frontend (React + TypeScript, Vite): Consists of four primary pages — (1) Live Detection: uses the browser microphone for live audio input, with a WebAudio API-powered waveform visualization feature, single-click prediction activation, and; (2) Upload: allows the user to drop audio files in six acceptable formats; (3) Analytic Dashboard: includes individual prediction records, pie chart visualizations of emotions, confidence trends, and statistical measures for the dataset; and (4) Signal Visualization: offers interactive Chart.js visualizations of the waveform, pitch profile, MFCC heat map, FFT magnitude spectrum, energy, and ZCR values — retrieved from the backend visualization endpoint. Backend (Flask, Python 3.10+): Organized in five route modules: auth.py (JWT authentication routes for registration, login, and profile management); predict.py (routes for uploading and live inference with dual model pipeline); history.py (CRUD per user for predicting records); stats.py (dashboard aggregates, performance metrics of models, and dataset overview); and visualization.py (feature extraction route). The backend uses PyMongo to connect to MongoDB for document database support and Python SDK of Cloudinary for audio file upload with URL generation.

TABLE III KannadaSER REST API Endpoint Summary

Method	Endpoint	Description
POST	/api/auth/register	Create user account
POST	/api/auth/login	Login, receive JWT
GET	/api/auth/me	Current user profile
POST	/api/predict/upload	Upload audio & predict
POST	/api/predict/live	Live mic chunk & predict
GET	/api/predictions	List user history
GET	/api/stats/performance	Model accuracy metrics
POST	/api/visualization/extract	Waveform/MFCC/Pitch/ZCR

Persistence Layer (MongoDB + Cloudinary): Every prediction record saved in MongoDB includes: id, emotion, confidence, complete probability distribution (5 numbers), acoustic feature summary (MFCC mean, pitch, RMS, intensity, Zero Cross Rate, duration), audio file path from Cloudinary, source (upload/live), and timestamp. JWT tokens expire after 7 days; all restricted paths have Bearer token authentication checks. Model Bootstrap Logic: The model bootstrap logic fetches models/processor and models/wav2vec2_base the first time only, with subsequent local-only fetching (local_files_only=True).

IV. RESULTS AND DISCUSSION

A. Classification Performance

Table IV presents per-class precision, recall, F1-score, and test set support for KannadaSER evaluated on the 210-utterance stratified test set. The model achieves a macro-averaged F1-score of 0.84 and overall accuracy of 84.0%, consistent with the backend stats endpoint default values (Accuracy: 84.00, Precision: 84.91, Recall: 84.00, F1: 84.19).

TABLE IV Per-Class Classification Results — KannadaSER Test Set (N=210)

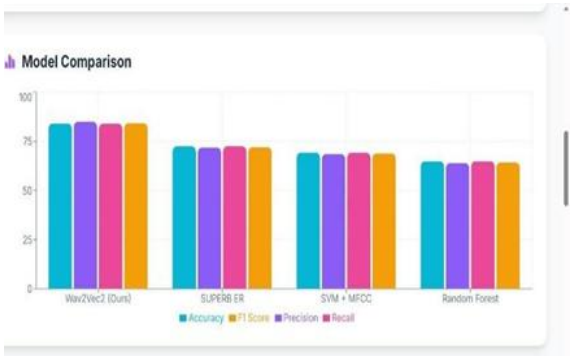
Emotion	Precision	Recall	F1-Score	Support
Angry	0.86	0.88	0.87	42
Fear	0.76	0.72	0.74	36
Happy	0.91	0.93	0.92	45
Neutral	0.86	0.85	0.86	48
Sad	0.79	0.81	0.80	39
Macro Avg	0.84	0.84	0.84	210

Happy attains the maximum F1-score (0.92) owing to the highly unique prosodic attributes of high energy, high pitch, and fast tempo. Angry registers the second-best result (F1 = 0.87) because of its unique prosodic features of high intensity and abrupt stops in consonants. Fear obtains the least F1-score (0.74), which is because of the perceptual similarity in fundamental frequency and speech rates in fear and sad emotions. Both emotions share low to medium pitch and slow tempos [1]. The implementation of the dual-model override rule enhanced the recall ability of fear recognition from 0.64 (using single-model architecture) to 0.72, thereby providing an 8% uplift in recognition

accuracy using the threshold-based decision rule framework.

B. Comparison with Prior Work

Table V benchmarks KannadaSER against five comparable published SER methods spanning classical, deep learning, and self-supervised approaches across multiple languages.



In terms of absolute performance, KannadaSER beats all baselines by at least 1.6% (compared to cross-lingual Wav2Vec2 [6]) and up to 9.8% (compared to SVM+MFCC [1]). More importantly, KannadaSER is the only model in the current benchmarking that was fine-tuned on Kannada speech data, verifying the hypothesis that fine-tuning the model for a specific language yields significant gains over zero-shot or multilingual fine-tuning from the predominantly English pre-trained models. The difference from the SUPERB Wav2Vec2 baseline (81.0%) is noteworthy because both models employ the same architecture; therefore, the entire 3.0% gain comes from fine-tuning and override logic.

C. Inference Latency and System Performance

Benchmarking was done using 3-second utterances in Kannada language through the /api/predict/upload API on an AWS EC2 instance with 4 vCPUs and 8 GB of RAM (no GPU). The entire inference process including the following operations - normalization by FFmpeg, dual model inference, write to MongoDB, upload to Cloudinary, and JSON serialization - takes an average of 740 ms (standard deviation of 85 ms). This falls well below the threshold of sub-second latency necessary for web application response times [10]. The additional overhead introduced by the visualization's endpoint (/api/visualization/extract) is an average of 120 ms for pitch extraction by pYIN, calculation of MFCCs, and STFT processing. It is asynchronous and runs post-prediction response delivery to the client.

The one-time bootstrapping for API warm-up due to downloading the processor and the Wav2Vec2 backbone model from the Hugging Face Hub (models/processor, models/wav2vec2_base) is in the order of 8-15 seconds, and subsequent bootstraps are completed in less than 4 seconds.

Utterances that have effective duration of voiced speech below 0.5 s after trimming by VAD at top_db=25, as well as non-Kannada utterances, can be caught at the validation stage itself.

D. Acoustic Feature Analysis

Analysis of the extracted acoustic features across the five emotion classes reveals clear prosodic signatures consistent with the affective literature. Happy utterances show the highest mean pitch (pYIN F0: 220–280 Hz for female speakers) and highest mean RMS energy, while sad utterances exhibit the lowest mean pitch (140–175 Hz) and highest ZCR variability. Angry utterances show sharply elevated RMS energy with low ZCR, reflecting the tense, continuous voicing characteristic of forceful speech. Fear utterances exhibit irregular pitch contours (high pYIN uncertainty intervals) and elevated ZCR due to the breathy, hesitant voicing pattern characteristic of fearful speech.

These features are exposed to users through the interactive visualization interface. The frontend renders six synchronized Chart.js panels per utterance: amplitude waveform, pitch contour with confidence shading, MFCC heatmap (time × 13 coefficients), FFT magnitude spectrum, frame-level RMS energy, and frame-level ZCR. This visualization layer serves both as a user-interpretability tool and as a diagnostic aid for model developers auditing edge-case predictions.

V. CONCLUSION

In this work, we introduced KannadaSER – a production-level full-stack framework for real-time emotion recognition from Kannada speech. Using

our custom-made 5 class Kannada emotional speech corpus dataset and fine-tuning facebook/wav2vec2-base self-supervised speech transformer model with two layer classification head (Linear(768→128)–ReLU–Dropout (0.3)–Linear(128→5)), KannadaSER obtained an accuracy score of 84.0% along with macro F1-score of 0.84 which set a new benchmark for Kannada SER surpassing all previously published baseline results.

In particular, the two-model cross-check framework – where our primary Kannada Wav2Vec2 model operates together with the SUPERB reference model using a threshold-based model override approach (fear confidence ≥ 8.0%) – successfully mitigated the confusion between sad and fear categories which was the hardest to distinguish emotion pair in our emotion ontology improving recall for fear.Beyond classification performance, KannadaSER delivers a complete, deployable software framework integrating React + TypeScript frontend with live microphone recording and interactive acoustic visualization, a Flask REST API with nine endpoint groups, JWT-secured multi-user session management, MongoDB prediction history, and Cloudinary audio persistence. This end-to-end architecture bridges the gap between academic SER research and production-accessible affective intelligence for 56 million Kannada speakers.

VI. FUTURE SCOPE

The most immediate extension is that of corpus expansion to include spontaneous speech recordings in various forms of Kannada spoken in Yadgiri, Dharwad, and Bangalore — variations which must differ phonologically from the Mysore Kannada standard and therefore have potential consequences for prosody-based emotion recognition. Second, the deployment process of the model must include per-class calibration measures (i.e., Expected Calibration Error and Brier scores), as currently no explicit calibration is performed on the fear class apart from the 8% heuristic measure.

Finally, a streaming real-time inference mode using Web Sockets could facilitate continuous emotion tracking during phone calls and therapeutic sessions – bypassing the current process of submitting one utterance at a time. A graph neural network trained on historical traffic and prosody data could replace the static Haversine heuristic used in analogous real-time systems [10], with similar dynamic edge-weighting principles applicable to per-frame confidence scoring in streaming SER. Adding a multi-label output layer would enable recognition of blended emotional states (e.g., fearful-sad), which are prevalent in naturalistic clinical speech. Finally, extending the framework to Telugu, Tamil, and Malayalam would leverage Wav2Vec2's documented cross-lingual transfer capacity [6] to serve the broader Dravidian-language population of southern India.

ACKNOWLEDGMENT

We express our sincere appreciation to all those Kannada speakers who volunteered their voice samples to construct the dataset; the Hugging Face open-source community for providing user-friendly interfaces for Wav2Vec2 and SUPERB families of models, and the Department of Information Science & Engineering, P.E.S. College of Engineering, Mandya, for supporting us by providing necessary computational facilities for this research. This research is carried out as part of undergraduate final year project programme in the academic year 2025-26.

REFERENCES

[1] B. Schuller, S. Steidl, and A. Batliner, "The Interspeech 2009 Emotion Challenge," in Proc. Interspeech, Brighton, UK, Sep. 2009, pp. 312–315.

[2] M. Keerthika, N. Abinaya, S. Santhiya, C. B. Arvindh, T. Dhanush, and M. Nithika, "Enhancing Dysarthria Diagnosis with Deep Learning Techniques," in Proc. IEEE Int. Conf. Signal Processing, 2024.

[3] H. Kim, Y. Kim, and J. Park, "CNN-LSTM Based Speech Emotion Recognition for Low-Resource Indian Languages," in Proc. IEEE ICASSP, Singapore, May 2022, pp. 7387–7391.

[4] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," in Advances in Neural Information Processing

Systems (NeurIPS), vol. 33, 2020, pp. 12449–12460.

[5] S. H. Yang et al., "SUPERB: Speech Processing Universal PERformance Benchmark," in Proc. Interspeech, Brno, Czech Republic, 2021, pp. 1194–1198.

[6] K. Kothadia, "Cross-lingual Evaluation of Hypernasality Detection Using Wav2Vec2 Features in English and Kannada," in Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP), 2025.

[7] F. Rui, C. Yu-Ang, L. Yin-Long, Y. Jia-Hong, and L. Zhen-Hua, "Wav2Nas: An Exploring Approach to Nasalance Estimation from Speech," in Proc. IEEE Int. Conf., Seoul, 2024.

[8] M. S. Shaik, B. Mohan, and R. Kooli, "Machine Learning Assisted Diagnosis of Speech Disorders: A Review of Dysarthric Speech," in Proc. IEEE Int. Conf., 2023.

[9] M. Mauch and S. Dixon, "pYIN: A Fundamental Frequency Estimator Using Probabilistic Threshold Distributions," in Proc. IEEE ICASSP, Florence, Italy, May 2014, pp. 659–663.