EGGCODEC: A ROBUST NEURAL ENCODEC FRAMEWORK FOR EGG RECONSTRUCTION AND F0 EXTRACTION

Rui Feng[†,1], *Yuang Chen*[†,1,2], *Yu Hu*[1], *Jiahong Yuan*[*,1,2]

[1] National Engineering Research Center of Speech and Language Information Processing, University of Science and Technology of China, Hefei, P. R. China
[2] Interdisciplinary Research Center for Linguistic Sciences, University of Science and Technology of China, Hefei, P. R. China

ABSTRACT

This paper introduces EGGCodec, a robust neural Encodec framework engineered for electroglottography (EGG) signal reconstruction and F0 extraction. We propose a multi-scale frequency-domain loss function to capture the nuanced relationship between original and reconstructed EGG signals, complemented by a time-domain correlation loss to improve generalization and accuracy. Unlike conventional Encodec models that extract F0 directly from features, EGGCodec leverages reconstructed EGG signals, which more closely correspond to F0. By removing the conventional GAN discriminator, we streamline EGGCodec's training process without compromising efficiency, incurring only negligible performance degradation. Trained on a widely used EGG-inclusive dataset, extensive evaluations demonstrate that EGGCodec outperforms state-of-the-art F0 extraction schemes, reducing mean absolute error (MAE) from 14.14 Hz to 13.69 Hz, and improving voicing decision error (VDE) by 38.2%. Moreover, extensive ablation experiments validate the contribution of each component of EGGCodec. For reproducibility, our code is available at **`https://github.com/RuiFeng-USTC/eggcodec_cya_and_fr`**.

Index Terms— F0 extraction, EGG reconstruction, speech inverse filtering, speech signal processing.

1. INTRODUCTION

Fundamental frequency (F0) extraction is a foundational task in speech signal processing since it reflects the rate of vocal fold vibration and carries essential information pertinent to prosody and speaker characteristics [1]. The accurate extraction of F0 is thus essential for various practical applications, including speech recognition, speech synthesis, speaker identification, prosody analysis, and music research [2, 3, 4, 5, 6].

Currently, numerous studies have been conducted on F0 extraction from speech signals. To improve F0 extraction accuracy, the authors in [5] proposed an auditory gain harmonic detection method exploiting selective Gammachirp filters to highlight harmonics and reduce the noise masking. In [7], the authors focused on melody extraction, tackling harmonic resolvability, which is a key challenge in F0 extraction. A deep learning approach for synthesizing F0 trajectories for speaker anonymization was presented in [8], addressing noise sensitivity and computational complexity. In [1], we introduced Wav2F0 that combines the Wav2vec 2.0 model with fully connected layers and LSTM. The Crepe [9] employs deep convolutional neural networks (CNN) directly on time-domain waveforms for data-driven F0 extraction. Despite these advances, the intricate vibration mechanisms of the vocal folds, coupled with the variability of recording conditions, render the task of accurately extracting F0 from waveforms a formidable challenge. Unlike microphone-captured speech signals, EGG signals provide higher accuracy and stability, reflecting vocal fold vibrations more precisely, and are thus well-suited for F0 extraction [10]. EGG remains the most common non-invasive method for observing vocal cord vibration, measuring resistance changes between electrodes placed on the thyroid cartilage [11, 12].

In this context, META Corporation in [13] developed **Encodec**, a neural network-based speech codec that effectively compresses and reconstructs speech signals using a convolutional encoder, symmetric decoder, and a generative adversarial network (GAN)-based framework to enhance perceptual audio quality [14]. Due to the strong capability of Encodec in maintaining high audio quality and fidelity, Encodec is widely used in studies related to F0 and EGG signal processing. For example, the author in [15] proposed a high-fidelity speech codec leveraging both speech and EGG signals for speech compression, while the authors in [16] introduced FunCodec, offering comparable speech quality with lower computational cost using the SEANet architecture. Other works include EnCLAP [17] proposed for automatic audio captioning and SpeechX [18] for zero-shot TTS and various speech transformations, both utilizing Encodec for robust signal processing. Preliminary experiments show Encodec provides a stable EGG signal fitting. However, due to the complexity of Encodec's structure, limitations in the loss function, and the

[†] These authors are co-first author. [*] Corresponding author.
{yuangchen21, fengruimse}@mail.ustc.edu.cn, jiahongyuan@ustc.edu.cn.

inherent instability of the GAN discriminator during practical applications, directly using the Encodec framework to reconstruct EGG signals is challenging.

This paper designs a robust neural Encodec framework called EGGCodec for precise EGG reconstruction and F0 extraction. EGGCodec leverages the superior ability of EGG signals to characterize vocal fold vibrations and locate them as the reconstruction target for F0 extraction. The structure and training process of EGGCodec have been greatly simplified by removing the GAN discriminators. To accurately measure the similarity between the reconstructed and target signals, a multi-scale frequency-domain loss function is proposed to capture the nuanced relationship between original and reconstructed EGG signals, complemented by a time-domain correlation loss to improve generalization and accuracy. Extensive evaluations demonstrate that EGGCodec outperforms state-of-the-art schemes across multiple performance metrics.

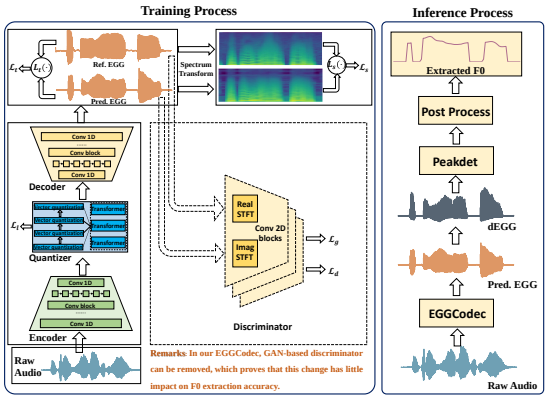

Fig. 1. The framework of the proposed EGGCodec.

2. THE DEVELOPED EGGCODEC FRAMEWORK

As illustrated in Fig. 1, we present the detailed framework of EGGCodec, which shows the input signals during the training process, including the raw audio signal, the label signal (i.e., Ref. EGG), and our output signal (i.e., Pred. EGG).

2.1. Transformation of the Reconstruction Target

EGGCodec processes speech signals while synchronously collecting EGG signals. During training, 3-7 dB of white noise is added to the speech signals to augment the data. The primary estimation target of EGGCodec is to reconstruct EGG waveforms from original speech for accurate F0 extraction. Speech signals are first encoded into compact representations, which are subsequently quantized by the quantizer and then reconstructed into waveforms by the decoder. By using the EGG signal as the target of the loss function, the model learns to generate outputs that closely match EGG signals rather than speech signals. It helps our proposed EGGCodec focus on reconstructing the vocal cord vibration signal from the speech input, to more accurately capture F0's fine details. This approach not only improves the accuracy of F0 extraction but also enhances the ability of EGGCodec to characterize the dynamics of vocal fold opening and closing [1].

2.2. The Design of Loss Functions

To ensure high-fidelity between reconstructed waveforms and target EGG signals across time-frequency domains, we propose a multi-scale frequency-domain loss combining L1 and L2 norms to measure Mel-spectrogram differences across varied window lengths. This is complemented by a time-domain correlation loss to improve generalization. Together, these losses enable EGGCodec to effectively capture multi-scale information features, formalized as follows:

$$\mathcal{L}_s = \frac{1}{6}\sum_{i=5}^{10}\left(\left\|\mathrm{S}(y_{\mathrm{pred}},2^i)-\mathrm{S}(y_{\mathrm{ref}},2^i)\right\|_1 + \left\|\mathrm{S}(y_{\mathrm{pred}},2^i)-\mathrm{S}(y_{\mathrm{ref}},2^i)\right\|_2\right), \quad (1)$$

where y_{pred} and y_{ref} represent the reconstructed waveform and the target EGG waveform, respectively. $\mathrm{S}(y,w)$ is the log-Mel spectrogram of signal y derived from the STFT with window length w. In Eq. (1), we employ a linear combination of L1 and L2 losses applied to the Mel-spectrogram, computed across multiple frequency windows to ensure spectral consistency. In the time domain, EGGCodec further improves the preservation of the signal's phase information by introducing a cosine distance loss that is

$$\mathcal{L}_{\cos} = 1 - \frac{\langle \mathbf{y}_1, \mathbf{y}_2\rangle}{\|\mathbf{y}_1\|\|\mathbf{y}_2\|}, \quad (2)$$

where $\mathbf{y}_1$ and $\mathbf{y}_2$ are two arbitrary time series. The cosine distance assesses their similarity via the normalized inner product, rendering it scale invariant. A lower value indicates a higher degree of correlation between the two sequences, whereas values closer to 1 indicate weaker similarity. Unlike Euclidean-based metrics that emphasize absolute amplitude disparities, cosine distance prioritizes relative patterns. To further enhance reconstruction accuracy, a hybrid time-domain loss function that combines L1, L2, and cosine distance is formulated as follows:

$$\mathcal{L}_t = \frac{\mathcal{L}_{\mathrm{L1}}(y_{\mathrm{pred}},y_{\mathrm{ref}}) + \mathcal{L}_{\mathrm{L2}}(y_{\mathrm{pred}},y_{\mathrm{ref}})}{\lambda} + \mathcal{L}_{\cos}(y_{\mathrm{pred}},y_{\mathrm{ref}}). \quad (3)$$

where λ denotes the weighting factor, empirically set to 100, to balance the contributions of the different loss terms, ensuring stable training and optimal performance. The terms $\mathcal{L}_{\mathrm{L1}}(y_{\mathrm{pred}},y_{\mathrm{ref}})$ and $\mathcal{L}_{\mathrm{L2}}(y_{\mathrm{pred}},y_{\mathrm{ref}})$ quantify absolute and squared differences, respectively, while $\mathcal{L}_{\cos}(y_{\mathrm{pred}},y_{\mathrm{ref}})$ enhances them with a scale-invariant similarity metric. To balance their contributions, we introduce a scaling factor of 100 to the L1 and L2 losses, empirically derived from extensive

[1]The proposed EGGCodec follows the EnCodec framework that consists mainly of an initial convolution, multiple residual and downsampling modules, and a temporal modeling component [13].

trials. This factor establishes a practical normalization magnitude, preventing dominance by L1 and L2 losses while preserving both absolute and pattern-based similarity fidelity.

Furthermore, we assign different weights to the aforementioned loss functions and combine them to form the reconstruction loss function, as follows:

$$\mathcal{L}_{\text{reco}} = \mathcal{L}_s + \lambda \times \mathcal{L}_t + \mathcal{L}_g + \mathcal{L}_d + \mathcal{L}_l, \tag{4}$$

where $\mathcal{L}_g$ embodies the generator's adversarial loss, ensuring perceptual fidelity in the reconstructed signal, while $\mathcal{L}_d$ reflects the discriminator's adversarial loss, distinguishing real from synthesized signals. Additionally, $\mathcal{L}_l$ denotes the entropy coding loss, optionally employed with a Transformer-based language model to enhance compression efficiency [13]. The coefficient of $\lambda = 100$, empirically tuned for $\mathcal{L}_t$, balances the loss terms, mitigating magnitude disparities between $\mathcal{L}_s$ and $\mathcal{L}_t$ that could destabilize training. Rigorous optimization confirms this factor fosters both stability and peak performance.

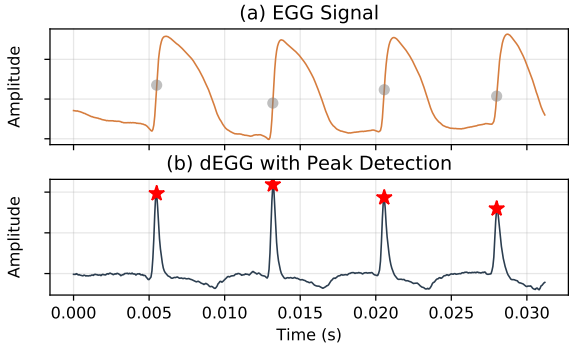

Fig. 2. The F0 extraction from dEGG signals using the peakdet algorithm to highlight peaks.

2.3. The Design of F0 Extraction Scheme

In EGGCodec, we propose an encoder-decoder architecture for accurate EGG reconstruction, combined with differential EGG (dEGG) processing to enhance F0 extraction. The reconstructed EGG is differentiated to generate the dEGG signal, as shown in Fig. 2, where peaks correspond to vocal fold closure instants. Using the peakdet algorithm [12, 19, 20, 21], these peaks are detected as periodic markers to calculate vibration periods and derive F0, followed by frame-level frequency prediction through post-processing. By aligning dEGG signal maxima with vocal fold dynamics, EGGCodec ensures robust and accurate F0 estimation. The importance of preproposing EGG signals from the PTDB-TUG dataset [22] is demonstrated in Fig. 3. As revealed in Fig. 3 (b), unfiltered reference EGG signals exhibit substantial low-frequency components originating from throat muscle artifacts during speech production rather than vocal fold vibrations. These extraneous components, which bear minimal relevance to fundamental frequency extraction, risk interfering with model training. Fig. 3 (c) illustrates that omitting the 50 Hz high-pass filter allows these low-frequency artifacts to disrupt learning. Our evaluation systematically compares filtered versus unfiltered EGG references, validating the necessity of this preprocessing step.

3. PERFORMANCE EVALUATIONS

3.1. Parameter Settings and Dataset

EGGCodec is trained using the Adam optimizer [23] with a learning rate of $\eta = 10^{-3}$, momentum coefficients $\beta_1 = 0.9$ and $\beta_2 = 0.999$, and $\epsilon = 1 \times 10^{-8}$, for 20 epochs with a batch size of 14 audio clips, accelerated on an NVIDIA RTX 4090 GPU. Training uses the PTDB_TUG corpus [22], containing 576 minutes of English speech from 20 speakers (10 male/10 female) with synchronized EGG recordings. To ensure multi-condition robustness, additive white noise is applied at SNR levels of 3 dB, 5 dB, 7 dB, and clean conditions. Evaluation is performed on the CSTR-FDA dataset [24], a gold-standard pitch determination corpus containing 5.53 minutes of speech (1 male/1 female) with ground-truth frequency contours. Performance is quantified using the PPMCC [25, 26, 27], which measures the linear correlation between reconstructed and reference EGG signals. Higher PPMCC indicates closer alignment with the original EGG waveform, preserving the periodic characteristics of vocal fold motion critical for accurate F0 estimation. EGGCodec is benchmarked against pYIN [28], Crepe [9], and Wav2F0 [1], evaluating mean absolute error (MAE), 50-cent raw pitch accuracy (RPA) [29], 20% gross pitch error (GPE) [30], voicing decision error (VDE) [29], and PPMCC.

3.2. Experimental Results

As depicted in Fig. 4, we compare the reconstructed EGG signals obtained by EGGCodec with the original EGG signals. The reconstructed EGG signals (Fig. 4 (c)) exhibit a high degree of consistency with the original signals (Fig. 4 (b)) in the vibrating regions of the vocal cords. In the voiced regions of the audio signal, the reconstructed EGG signal successfully avoids generating erroneous waveforms when the original EGG signal shows no significant vibration, suggesting the reconstruction process effectively reduces the interference from non-vocal vibrations. Moreover, the reconstructed EGG waveforms maintain nearly the same clarity and simplicity as the original signals, preserving their essential characteristics while reducing potential distortions. This faithful reconstruction preserves essential details for F0 extraction while reducing noise and irrelevant information, providing a cleaner, more reliable foundation for F0 extraction. As shown in Fig. 5, the effect with or without (w/o) noise-augmented training on EGG reconstruction is investigated. From Fig. 5 (b) and (c), it can be seen that training without noise augmentation increases the reconstructed signal's vulnerability to noise during unvoiced segments, increasing the risk of misclassifying noise as F0. In contrast, noise-enhanced training yields perfect reconstructed EGG signals, enabling accurate vocal fold cycle detection. Experimental results reveal that the stability of the reconstructed signal is essential for accurate peak detection. Incorrect EGG reconstructions yield abnormal dEGG

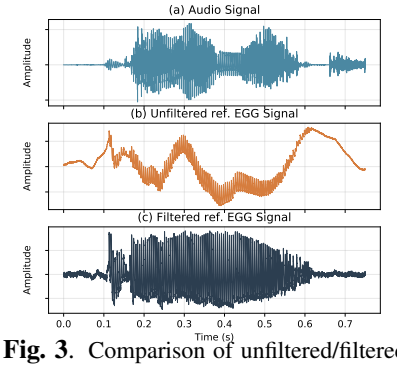

Fig. 3. Comparison of unfiltered/filtered EGG and the processing for PTDB-TUG.

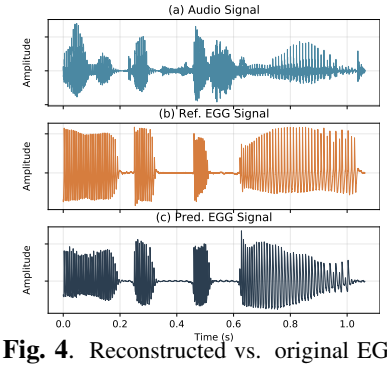

Fig. 4. Reconstructed vs. original EGG and audio signals.

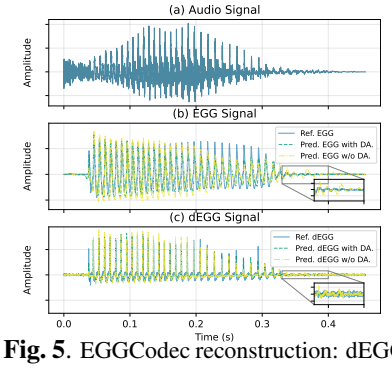

Fig. 5. EGGCodec reconstruction: dEGG peaks without noise augmentation.

peaks, which potentially cause segmentation errors in vibration cycles and ultimately affect the accuracy of F0 extraction.

To systematically evaluate each EGGCodec component's contribution, we designed multiple control groups for ablation experiments. The **(EGGCodec, Optimal)** configuration integrates cosine distance and L1/L2 norm losses in the time domain for robust error characterization, employs L1/L2 losses in the frequency domain to capture spectral features, and uses 3 dB SNR white noise augmentation with GAN-based adversarial training to enhance noise resilience and fidelity. A 50 Hz high-pass filter is applied to the reference EGG signal to remove low-frequency interference while preserving critical features, achieving a better balance between accuracy and robustness across evaluation metrics. Other control groups include: **(EGGCodec, Cos)** and **(EGGCodec, L1/L2)** solely consider cosine distance and L1/L2 losses in time domain, respectively; **(EGGCodec, w/o Time)**, sans time-domain losses; **(EGGCodec, w/o Freq)**, sans frequency-domain losses; **(EGGCodec, w/o NDA)**, **(EGGCodec, 5dB NDA)**, and **(EGGCodec, 7dB NDA)**, varying noise augmentation (none, 5 dB SNR, 7 dB SNR); **(EGGCodec, w/o GAN)**, excluding GAN training; and **(EGGCodec, Unfiltered)**, assessing the effectiveness of high-pass filtering.

Table 1. Performance comparison of the EGGCodec with the state-of-the-art baselines and its ablation studies.

Model Name	PPMCC ↑	MAE (Hz) ↓	RPA (%) ↑	GPE (%) ↓	VDE (%) ↓
pYIN [28]	-	36.9	62.9	24.3	26.1
crepe [9]	-	15.9	87.8	**7.9**	9.5
Wav2F0 [1]	-	15.2	81.8	9.7	8.2
(EGGCodec, Optimal)	0.834	**13.7**	86.0	9.1	**5.5**
(EGGCodec, Cos)	0.818	**15.8**	85.4	10.1	**5.9**
(EGGCodec, L1/L2)	0.468	54.7	72.9	24.8	23.2
(EGGCodec, w/o Time)	0.002	40.3	56.6	35.8	9.2
(EGGCodec, w/o Freq)	0.82	**15.3**	86.5	10.2	**6.0**
(EGGCodec, 5dB NDA)	0.828	16.7	84.7	10.9	**6.5**
(EGGCodec, w/o NDA)	0.819	17.3	84.1	12.2	**7.4**
(EGGCodec, 7dB NDA)	0.839	17.4	84.7	11.1	**6.5**
(EGGCodec, w/o GAN)	0.812	14.2	86.1	9.5	**5.5**
(EGGCodec, Unfiltered)	0.278	26.7	73.9	19.0	10.3

Table 1 presents a comparative performance analysis of EGGCodec against baselines. **(EGGCodec, Optimal)** achieves outstanding performance in F0 extraction, with an MAE of just 13.69 Hz, outperforming all baselines. **(EGGCodec, Optimal)** significantly outperforms pYIN, which records a much higher MAE of 36.85 Hz, achieving the lowest VDE at 5.5%, indicating superior robustness in voiced/unvoiced classification. Among control groups, **(EGGCodec, w/o GAN)** achieves an MAE of 14.17 Hz and an RPA of 86.1%. **(EGGCodec, Cos)** and **(EGGCodec, L1/L2)** demonstrate competitive performance, with the former maintaining a relatively low MAE of 15.76 Hz. Moreover, **(EGGCodec, Optimal)** and **(EGGCodec, 7dB NDA)** achieve the highest PPMCC values, i.e., 0.834 and 0.839, respectively, indicating superior EGG reconstruction and providing a more reliable foundation for F0 extraction. In contrast, **(EGGCodec, Unfiltered)** has a significantly lower PPMCC of 0.278, suggesting a poor correlation with the reference EGG, which may negatively impact subsequent F0 extraction. Thus, EGGCodec not only enhances the accuracy of EGG reconstruction but also contributes to the stability and reliability of F0 extraction. It can also be observed that **(EEGCodec, 7dB NDA)** presented PMCC slightly higher than **(EEGCodec, Optimal)**. This is mainly because PMCC measures the linear correlation of the waveform as a whole and is not affected by absolute amplitude offset and scale variation. Moreover, MAE and GPE are based on a point-by-point comparison of the F0 values of the reconstructed signal with the reference signal. The difference between the metrics is more likely to reflect the different responses of the F0 extraction to the microstructure of the signal rather than a degradation of the overall reconstruction quality.

4. CONCLUSION

This paper proposes an innovative neural Encodec framework named EGGCodec, which advances EGG signal reconstruction and F0 extraction. EGGCodec delivers exceptional accuracy and generalization on an EGG-rich dataset. Comparisons with state-of-the-art F0 extraction schemes highlight EGGCodec's significant gains, and ablation studies demonstrate the contribution of each component to its performance. In our future work, we intend to explore the F0 estimation results of the proposed EGGCodec under urban acoustic noise and different SNRs, in order to further demonstrate the robustness of EGGCodec to noise.

5. REFERENCES

[1] Rui Feng, Yin-Long Liu, Zhen-Hua Ling, and Jia-Hong Yuan, "Wav2F0: Exploring the potential of Wav2Vec 2.0 for speech fundamental frequency extraction," in *2024 IEEE 14th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, 2024, pp. 169–173.

[2] Qisheng Yang, Weiqiu Jin, Qihang Zhang, Yuhong Wei, Zhanfeng Guo, Xiaoshi Li, Yi Yang, Qingquan Luo, He Tian, and Tian-Ling Ren, "Mixed-modality speech recognition and interaction using a wearable artificial throat," *Nature Machine Intelligence*, vol. 5, no. 2, pp. 169–180, 2023.

[3] Jixun Yao, Qing Wang, Yi Lei, Pengcheng Guo, Lei Xie, Namin Wang, and Jie Liu, "Distinguishable speaker anonymization based on formant and fundamental frequency scaling," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[4] Sudarsana Reddy Kadiri, Paavo Alku, and B. Yegnanarayana, "Extraction and utilization of excitation information of speech: A review," *Proceedings of the IEEE*, vol. 109, no. 12, pp. 1920–1941, 2021.

[5] Anderson Queiroz and Rosângela Coelho, "Harmonic detection from noisy speech with auditory frame gain for intelligibility enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 2522–2531, 2024.

[6] Larissa Cristina Berti, Evelyn Alves Spazzapan, Marcelo Queiroz, et al., "Fundamental frequency related parameters in brazilians with COVID-19," *The Journal of the Acoustical Society of America*, vol. 153, no. 1, pp. 576–585, 2023.

[7] Keren Shao, Ke Chen, Taylor Berg-Kirkpatrick, and Shlomo Dubnov, "Towards improving harmonic sensitivity and prediction stability for singing melody extraction," in *ISMIR*, 2023, pp. 657–663.

[8] Ünal Ege Gaznepoglu and Nils Peters, "Deep learning-based F0 synthesis for speaker anonymization," in *2023 31st European Signal Processing Conference (EUSIPCO)*. IEEE, 2023, pp. 291–295.

[9] Jong Wook Kim, Justin Salamon, Peter Li, and Juan Pablo Bello, "Crepe: A convolutional representation for pitch estimation," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 161–165.

[10] Christian T Herbst, "Electroglottography–an update," *Journal of Voice*, vol. 34, no. 4, pp. 503–526, 2020.

[11] Minh-Châu Nguyen, *Exploring Machine Learning perspectives for electroglottographic signals*, Ph.D. thesis, LIG (Laboratoire informatique de Grenoble), 2023.

[12] Nathalie Henrich, Christophe d'Alessandro, Boris Doval, and Michèle Castellengo, "On the use of the derivative of electroglottographic signals for characterization of nonpathological phonation," *The Journal of the Acoustical Society of America*, vol. 115, no. 3, pp. 1321–1332, 2004.

[13] Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi, "High fidelity neural audio compression," *Transactions on Machine Learning Research*, 2023, Featured Certification, Reproducibility Certification.

[14] Kundan Kumar, Rithesh Kumar, Thibault De Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre De Brebisson, Yoshua Bengio, and Aaron C Courville, "Melgan: Generative adversarial networks for conditional waveform synthesis," *Advances in neural information processing systems*, vol. 32, 2019.

[15] Lingfeng Zhang, Lijiang Chen, Yuye Su, Chunfeng Cui, and Qi Zhao, "AECodec: High fidelity neural audio codec based on speech and electroglottograph," in *2024 3rd International Conference on Artificial Intelligence, Human-Computer Interaction and Robotics (AIHCIR)*, 2024, pp. 111–115.

[16] Zhihao Du, Shiliang Zhang, Kai Hu, and Siqi Zheng, "FunCodec: A fundamental, reproducible and integrable open-source toolkit for neural speech codec," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 591–595.

[17] Jaeyeon Kim, Jaeyoon Jung, Jinjoo Lee, and Sang Hoon Woo, "En-CLAP: Combining neural audio codec and audio-text joint embedding for automated audio captioning," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 6735–6739.

[18] Xiaofei Wang, Manthan Thakker, Zhuo Chen, Naoyuki Kanda, Sefik Emre Eskimez, Sanyuan Chen, Min Tang, Shujie Liu, Jinyu Li, and Takuya Yoshioka, "SpeechX: Neural codec language model as a versatile speech transformer," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 3355–3364, 2024.

[19] Alexis Michaud, "Final consonants and glottalization: new perspectives from hanoi vietnamese," *Phonetica*, vol. 61, no. 2-3, pp. 119–146, 2005.

[20] Martine Mazaudon and Alexis Michaud, "Tonal contrasts and initial consonants: a case study of tamang, a 'missing link'in tonogenesis," *Phonetica*, vol. 65, no. 4, pp. 231–256, 2009.

[21] Alexis Michaud, "A measurement from electroglottography: DECPA, and its application in prosody," in *Speech Prosody 2004*. International Speech Communication Association (ISCA), 2004, pp. pp–633.

[22] Gregor Pirker, Michael Wohlmayr, Stefan Petrik, and Franz Pernkopf, "A pitch tracking corpus with evaluation on multipitch tracking scenario.," in *Interspeech*, 2011, pp. 1509–1512.

[23] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[24] Paul C Bagshaw, Steven M Hiller, and Mervyn A Jack, "Enhanced pitch tracking and the processing of F0 contours for computer aided intonation teaching.," 1993.

[25] Gerhard Nahler and Gerhard Nahler, "Pearson correlation coefficient," *Dictionary of Pharmaceutical Medicine*, pp. 132–132, 2009.

[26] Mittapalle Kiran Reddy and Paavo Alku, "Exemplar-based sparse representations for detection of parkinson's disease from speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 1386–1396, 2023.

[27] Rui Feng, Yu-Ang Chen, Yin-Long Liu, Jia-Hong Yuan, and Zhen-Hua Ling, "Wav2Nas: An exploratory approach to nasalance estimation in speech," in *2024 IEEE 14th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, 2024, pp. 1–5.

[28] Matthias Mauch and Simon Dixon, "pYIN: A fundamental frequency estimator using probabilistic threshold distributions," in *2014 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 2014, pp. 659–663.

[29] Yixuan Zhang, Heming Wang, and DeLiang Wang, "F0 estimation and voicing detection with cascade architecture in noisy speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 3760–3770, 2023.

[30] Lyudmila Sukhostat and Yadigar Imamverdiyev, "A comparative analysis of pitch detection methods under the influence of different noise conditions," *Journal of voice*, vol. 29, no. 4, pp. 410–417, 2015.